Available online at www.sciencedirect.com
ScienceDirect
Computers, Environment and Urban Systems
31 (2007) 4–18
Computers, Environment and Urban Systems
www.elsevier.com/locate/compenvurbsys

A dual approach to cluster discovery in point event data sets

Allan J. Brimicombe *

Centre for Geo-Information Studies, University of East London, University Way, London E16 2RD, UK

Received 30 April 2004; received in revised form 4 July 2005; accepted 4 July 2005

Abstract

Spatial data mining seeks to discover meaningful patterns in data where a prime dimension of interest is geographical location. Consideration of a spatial dimension becomes important where data either refer to specific locations and/or have significant spatial dependence which needs to be considered if meaningful patterns are to emerge. For point event data there are two main groups of approaches to identifying clusters. One stems from the statistical tradition of classification which assigns point events to a spatial segmentation. A popular method is the *k*-means algorithm. The other broad approach is one which searches for 'hot spots' which can be loosely defined as a localised excess of some incidence rate. Examples of this approach are GAM and kernel density estimation. This paper presents a novel variable resolution approach to 'hot spot' cluster discovery which acts to define spatial concentrations within the point event data. 'Hot spot' centroids are then used to establish additional distance variables and initial cluster centroids for a *k*-means classification that produces a segmentation, both spatially and by attribute. This dual approach is effective in quickly focusing on rational candidate solutions to the values of *k* and choice of initial candidate centroids in the *k*-means clustering. This is demonstrated through the analysis of a business transactions database. The overall dual approach can be used effectively to explore clusters in very large point event data sets.

Keywords: Spatial data mining; Point events; Clustering; Hot spots; Geocomputation; Robust normalisation

* Tel.: +44 20 8223 2352; fax: +44 20 8223 2918.
E-mail address: a.j.brimicombe@uel.ac.uk

0198-9715/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2005.07.004

1. Introduction

Prior to the 1990s the spatial sciences, and the application of geographical information systems (GIS) in particular, suffered from a paucity of digital data sets. The 1990s were a period of transition into data-richness, a trend which accelerates today. Digital spatial data sets have grown rapidly in coverage, volume of records and numbers of attributes per record (Gahegan, 2003; Miller & Han, 2001). This state change has come about as a result of:

- improved technology and wider use of GPS, remote sensing and digital photogrammetry for collecting data on topographic and other physical objects;
- the introduction of new approaches to obtaining lifestyle and preference data such as through loyalty cards;
- dramatic increases in computing power to process raw data coupled with falling unit costs of data storage and data processing;
- the advent of data warehousing technologies;
- more efficient means of accessing and delivering data on-line.

The technical advances in hardware, software and data have been so profound that they have fundamentally affected the range of problems studied and the methodologies used to do so (Macmillan, 1998). An exponential rise in the size of databases, their increasing complexity and the rate at which they can accumulate on a daily basis have therefore lead to an urgent need for techniques that can mine very large databases for the knowledge they contain. Consequently, an active area of research has focused on *spatial data mining* which can be defined as *techniques for the discovery of meaningful patterns from large data sets where a prime dimension of interest is geographical location*. This paper focuses on clustering as a central aspect of spatial data mining and seeks to demonstrate the benefits of using 'hot spot' approaches to clustering in tandem with segmentation approaches to clustering. This is demonstrated using a case study analysis of a business transactions database. The following section discusses the theoretical perspectives and the dichotomy between the two different approaches to clustering. A form of 'hot spot' type clustering is then introduced and is subsequently used in the case study to guide a *k*-means classification of spatial and non-spatial attributes for a customer database. This forms the basis of a dual approach to cluster discovery as alluded to in the title of the paper.

2. Cluster detection in point event data

Transactions databases, be they for business, crime or health, can be regarded as point event data sets if each record has a specific geographical identifier such that geocoding can be achieved at the resolution of an address or postcode. From a location perspective the point event is a binary occurrence – either it happened there or it did not. From a data perspective, the binary occurrence may have added dimensions of attributes that describe the nature or content of the transaction which may relate to the location, the individual or the event that has been recorded. The traditional approach to non-spatial analyses of attributes may reveal apparently meaningful knowledge but may well be lacking in perspicacity or may even be misleading if underlying spatial distributions and dependencies are ignored. The exploratory analysis of point event data seeks to identify patterns using all

the dimensions of the data from which causal processes can be hypothesised or inferred (Fotheringham, 1992; Unwin, 1996). The analysis of point event patterns in geography, ecology and epidemiology has a long tradition (e.g. Clark & Evans, 1954; Cliff & Ord, 1981; Harvey, 1966; Knox, 1964; Mantel, 1967; Snow, 1855). Over the past decade, however, two broad thrusts have lead to a renewed interest in analysing point event patterns. These are the rise of geocomputation and a re-focusing away from global towards local analyses.

The adoption of geocomputational approaches to spatial data analysis represents a paradigm shift in which computers and hence computational tools play a pivotal role in the form of analysis as an essential defining ingredient of the science alongside observation, experimentation and theory (Armstrong, 2000; Couclelis, 1998; Fotheringham, 1998; Longley, Brooks, McDonnell, & MacMillan, 1998; Openshaw & Abrahart, 2000). The rationale for a geocomputational approach is driven by the advent of data-richness (discussed above), by the growing necessity in some areas for large data sets as a prerequisite for non-trivial analyses and by the growth of computationally intensive simulation models. The tools for geocomputation naturally include GIS but they are increasingly viewed as just one class of tool to be used alongside neural networks, artificial intelligence, heuristics, spatial statistics, fuzzy computation, fractals, genetic algorithms, cellular automata, simulated annealing and parallel computing (Brimicombe, 2003). The other broad thrust has been a re-focusing within quantitative geography towards spatial variation at the local level rather than in the search for global patterns (Fotheringham, 1997; Fotheringham & Brunsdon, 1999; Fotheringham, Brunsdon, & Charlton, 2000). The new emphasis is on exploring and understanding the spatial differences between localities rather than on quantifying their more general, global similarities. Such approaches are often data-rich and geocomputational.

Broad patterns detected in point event data are usually classified as random, uniform or clustered. Spatially random data are assumed to have no underlying spatial process of interest that can be modelled. Phillips (1999) has nevertheless pointed out that such apparent randomness may be attributable to chaotic, complex deterministic patterns. For spatial uniformity a space-filling, mutual exclusion process can be hypothesised. It is, however, clustered patterns that raise the strongest hypotheses for and interest in identifying underlying processes. Thus, where data either refer to specific locations and/or have significant spatial dependence which has to be considered if meaningful patterns are to emerge, spatial cluster detection methods lie at the heart of spatial data mining (Estivill-Castro & Lee, 2002; Halls, Bulling, White, Garland, & Harris, 2001; Kiang, 2001; Miller & Han, 2001; Murray, 2000; Murray & Estivill-Castro, 1998; Openshaw, 1998). Within the context of geocomputation and local analyses, there are two broad approaches to cluster detection and it is here that a significant dichotomy in the meaning of 'cluster' arises as discussed in the remainder of this section.

One set of approaches is allied to mainstream statistics of cluster analysis arising from the work of Sokal and Sneath (1963). Clustering in this context is a means of classification or grouping where clusters are "groups of highly similar entities" (Aldenderfer & Blashfield, 1984, p. 7). Spatially, cluster analysis will seek to form a segmentation into regions which minimise within-cluster variation but maximise between-cluster variation. There is a general expectation that the clustering mutually exclusively includes all point events and is therefore space-filling within the geographical extent of the data under consideration. Examples of this approach are to be found in Murray and Estivill-Castro (1998), Murray

(2000), Han, Kamber, and Tung (2001) and Kiang (2001). A widely-used clustering algorithm is the *k*-means classification (MacQueen, 1967) due to its relative efficiency in processing large numbers of cases having many attribute variables. Its weakness, however, is sensitivity to outliers (Han et al., 2001) and the need to specify *ab initio* the number (k) of desired clusters and optionally the location of N initial candidate centroids. Such prior specification is counter to the spirit of spatial data mining in which the data themselves should indicate the number and location of clusters rather than as speculated by the analyst. This has lead Halls et al. (2001) and Estivill-Castro and Lee (2002) to use Dirichlet and Delaunay diagrams respectively to define spatial clusters. These algorithms, however, will fail where points occupy the same location (as will often happen with geocoding, say, at postcode level) and to spatially de-duplicate the data set will lead to important data loss. The case study given below uses *k*-means classification because of its general popularity and its accessibility through many statistical packages. Though not inherently a spatial tool, it can achieve spatial segmentations using X and Y co-ordinate values or, as will be illustrated in the case study below, when converted into distance variables.

The other broad set of approaches treat a cluster as a 'hot spot' which can be loosely defined as a localised excess of some incidence rate, though there is no generally accepted definition of a 'hot spot'. This approach is typified by Openshaw's Geographical Analysis Machine (GAM) and its descendants (Openshaw, 1994, 1998; Openshaw, Charlton, Wymer, & Craft, 1987). Similar approaches are based around kernel density functions in which the highest densities are accepted as 'hot spots' (e.g. Gatrell & Rowlingson, 1994; Gatrell, Bailey, Diggle, & Rowlingson, 1996; Rowlingson & Diggle, 1993). The 'hot spot' approach is the mainstay of spatial epidemiology (Lawson, 2001) which seeks to identify any significantly elevated risk above that which might be expected from an at-risk background population. Defining a population at risk is clearly critical to this approach and in some, if not many, data mining applications this may not be possible at the outset. Mis-specification of an at-risk background population is likely to lead to erroneous results. Furthermore, intrinsic to this approach is that some of the points form 'hot spots' and the rest are no longer the focus of analysis. This is a fundamental difference from the set of techniques discussed in the previous paragraph where every point is assigned mutually exclusively to a group. Section 4 of this paper demonstrates how these two approaches can be usefully brought together. The choice of a 'hot spot' technique to use in such a demonstration remains problematic. Where an at-risk population is not initially specified (in other words, 'hot spot' detection is based solely on the distribution of point events), kernel density mapping is popularly used and is accessible through, for example, the Spatial Analyst extension to ArcView® and public domain MapBasic® software for MapInfo® (see Atkinson & Unwin, 2002). The kernel density algorithm requires the setting of two parameters: the underlying grid size and bandwidth of the kernel. Reasonable values for these parameters can be difficult to estimate. There are rules of thumb suggested, for example, in Fotheringham et al. (2000) and Atkinson and Unwin (2002), and some software provide default values. Nevertheless, best practice would suggest a form of sensitivity analysis by varying the parameters as illustrated in Fig. 1. The point event data in Fig. 1(a) has a very obvious main cluster as well as a possible number of other smaller clusters. Fig. 1(b)–(d) show the effect of increases in bandwidth on the kernel density mapping. All of these have been clipped from a broader geographical set so as to avoid edge effects (Atkinson & Unwin, 2002; Koch & Denike, 2001). Whilst the obvious main cluster continues to dominate at all bandwidths, the smaller 'hot spots' denoted

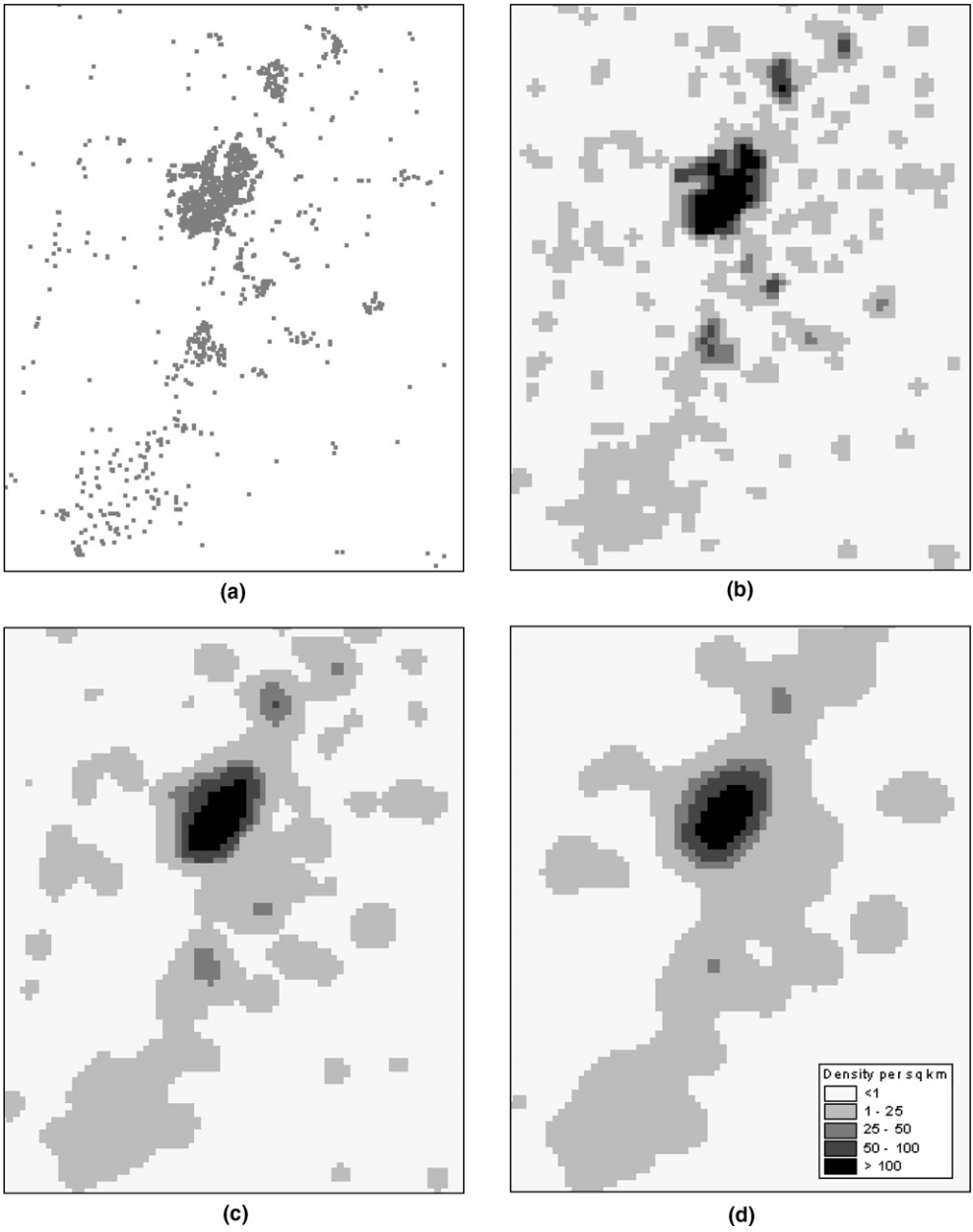

Fig. 1. Kernel density mapping: (a) distribution of point event data, (b) 250 m grid, 500 m bandwidth, (c) 250 m grid, 1000 m bandwidth, (d) 250 m grid, 1500 m bandwidth.

by the higher densities are less permanent and very much dependent on the bandwidth selected; the smaller the bandwidth the more 'hot spots' that can be made to appear. This problem is shared with many interpolation techniques and for this reason an alternative approach to 'hot spot' detection using a form of spatial segmentation is deployed in the case study. This alternative approach is described in the next section.

3. A variable resolution approach to the analysis of point event data

A recursive decomposition of space into gradually smaller spatial units that are nevertheless space-filling is generally referred to as a hierarchical tessellation. Any such decomposition requires some predefined criteria such as decomposition ratio and minimum size of spatial units in order to guide and finally terminate the algorithm. Possibly the most studied hierarchical tessellation is the quadtree (Samet, 1984) in which an initial square region covering the entire study area is repeatedly decomposed into quadrants using a fixed decomposition ratio of 1:4 until a predefined level of cell homogeneity or atomic (minimum) size is reached. A more generalised framework for hierarchical tessellations that includes variable decomposition ratios and rectangular cell shapes has been established by Tsui and Brimicombe (1997a). These adaptive recursive tessellations allow a variable resolution approach to the decomposition of space, in other words, no longer are scale and resolution treated as being uniform across an area but are allowed to vary spatially in response to patterns within the point event data. General applicability of adaptive recursive tessellations to spatial analysis are given in Tsui and Brimicombe (1997b) with more a specific application to point pattern analysis in Brimicombe and Tsui (2000).

The algorithm, as implemented here, uses a divide and conquer approach which treats each point as a binary occurrence of some phenomenon without reference to further descriptive attributes. Firstly, the most important parameter, the atomic or minimum cell size, is established by comparing the median nearest neighbour distance between point events with the average expected nearest neighbour distance. The larger of the two is accepted and squared to give the atomic cell size. Where the areal extent of the study area is ambiguous (i.e. not defined by some administrative or other boundary), the calculation of the average expected nearest neighbour distance is rendered intractable. In such cases a convex hull is established around the point event data set and buffered by 1% of the convex hull area so that no points lie on the boundary. The initial bounding rectangle is taken as a 2^n multiple of the atomic cell size to cover the study area. Variable resolution decomposition into different size cells is then carried out such that a quadrant is left undivided if it contains no points, if it has reached the atomic cell size or if the variance of points at the next level of decomposition is greater than or equal to one. On completion of the algorithm, cells containing zero point events are deleted and the remainder can be displayed as density classes or, if data on an at-risk population are available, incident rates can be calculated and tested for significance. In general a two stage process is adopted: an initial visualisation of density clusters and a subsequent visualisation of rates or risk (Brimicombe, 2003). Whilst the authors of GAM-type and epidemiological approaches are dismissive of identifying clusters without reference to an at-risk or control population, count data on their own do reflect workload, revenue stream or commitment of resources in meeting a spatially distributed demand (such as in response to crime). Once point event densities and any additional attribute dimensions are clearly understood, then a second stage analysis of risk can be carried out, if necessary, for the application at hand. Tests have shown the variable resolution decomposition algorithm to be consistently effective in comparison with other approaches of point event cluster detection (Brimicombe & Tsui, 2000). The resulting polygons are termed Geo-ProZones (GPZ) as they represent zones of geographical proximity in the point event pattern (see example in the next section) and may be interpreted as 'hot spots' should they have either a localised excess in density or incident rate. One important difference of this approach to, say, the kernel density

mapping, is that GPZ are spatial segmentations into polygons rather than an interpolation into a surface. As such there are no edge effects and because all parameters are set consistently within the algorithm, there is a unique solution.

The method being proposed in this paper is to use GPZ cluster centroids as a guide to setting up and running the *k*-means clustering. GPZ clusters are used to analyse the spatial distribution of binary events, typical of a 'hot spot' approach, in order to suggest an initial value for *k* and for identifying *N* candidate centroids located within the 'hot spots'. The *k*-means clustering can then include other descriptive attributes of the point events to derive a spatial segmentation inclusive of all point events and all data dimensions. This would bring together the two broad approaches to cluster detection in spatial data mining discussed above.

4. Geo-ProZones and *k*-means clustering in tandem: a case study

The case study focuses on an analysis of a business transactions database for one year. The database contained details of 2390 customers of whom 2361 (98.7%) could be geocoded to postcode level. This is a relatively small database by data mining standards but is sufficiently tractable to allow experimentation and tracking of special cases and checking results without being deluged by the data. The distribution of geocoded customers is given in Fig. 2(a). This particular business operates a service from a single hub or outlet situated to the north east of London, UK. The owner sees the business as serving a regional market rather than a national one with customers from outside the immediate region representing opportunist sales to customers passing through the area or temporarily visiting. Although the 'region' is spatially undefined (i.e. does not correspond to any particular administrative boundaries), from an inspection of the spatial distribution and the database, 327 customers were deemed as being from outside the target regional market and excluded from further analysis. The study data set of 2034 customers is given in Fig. 2(b) with an enlargement of the core area in Fig. 2(c) which is the same as Fig. 1(a). As well as the visually obvious clustering of customers in certain locations, there is a discernible linear trend from southwest to northeast following the alignment of a motorway through the area. The additional

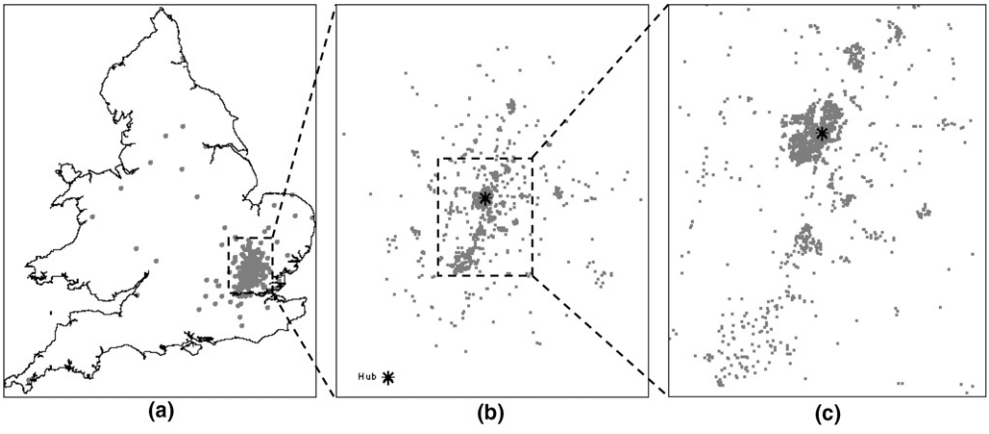

Fig. 2. Distribution of customers: (a) nationally, (b) regionally (~60 km × 70 km), (c) locally (~17 km × 21 km).

Table 1
GB-Profiles classification in 10 classes (based on Openshaw and Blake, 1996)

1	Struggling: Multi-ethnic areas – pensioners and single parents – high unemployment – local authority rented flats
2	Struggling: Council tenants – blue collar families and single parents – local Authority rented terraces
3	Struggling: Less prosperous pensioner areas – retired blue collar residents – local authority rented semi-detached houses
4	Struggling: Multi-ethnic areas; less prosperous private renters – young blue collar families with children – privately renting terraces and bedsits
5	Aspiring: Academic centres & student areas – young educated white collar singles and couples – privately rented bedsits and flats
6	Aspiring: Young married suburbia – young well-off blue collar couples and families – mixed tenure terraces
7	Climbing: Well-off suburban areas – young white collar couples and families – buying semi-detached and detached houses
8	Established: Rural farming communities – mature well-off self-employed couples and pensioners – owning or privately renting large detached houses
9	Prospering: Affluent achievers – mature educated professional families – owning and buying large detached houses
10	Established: Comfortable middle-agers – mature white collar couples and families – owning and buying semi-detached houses

attributes for each customer were SPEND (total amount of purchases by each customer for the year) and a geodemographic lifestyle classification based on customer postcode. In this instance the GB-Profiles classification into 10 classes (Openshaw & Blake, 1996) has been used and their characteristics are given in Table 1. These were assigned as 10 binary variables GB-PROF1 through GB-PROF10. It is recognised that in any geodemographic lifestyle classification there can be variability at the level of the neighbourhood, but that this is outweighed by the degree of similarity to be found amongst residents (Sleight, 1997). Also, the classes themselves may overlap in the characteristics they describe. This might militate against the use of binary variables. Nevertheless, GB-Profiles (using the 10 class option) returns a single class type for a given postcode based on majority characteristics and it is on this basis that binary variables have been deployed.

Turning now to the k-means classification, the number of clusters needs to be specified *ab initio* and furthermore there is an option to choose the first N observations as candidate centroids where $N = k$. The problem is in identifying suitable (or hypothesised) starting values of k and whether N candidate spatial centroids are to be specified. Clearly it would be useful to have some guideline and this is where, in this study, a 'hot spot' approach to clustering (in this case GPZ) has been used to inform the k-means classification.

The spatial distribution of customers in Fig. 2(b) requires a bounded study area for the GPZ algorithm and this is established using a slightly buffered convex hull (Fig. 3(a)). The GPZ algorithm is then applied and the results displayed as density classes. GPZ for density classes $n >= 2$ are given in Fig. 3(b) and (c) where the density value n in the legend represents a point event density of 2^n points per atomic cell size. The density pattern in Fig. 3(c) is visually comparable to that of Fig. 1(b) but represents a spatial segmentation into polygons that has a unique solution for the point event data set. Which of these GPZ can be classed as a 'hot spot' for guiding the k-means classification? This is a recurring problem of density mapping met, for example, on an almost daily basis in crime mapping. Generally, higher densities are interpreted as 'hot spots'. In this instance there are six GPZ of density

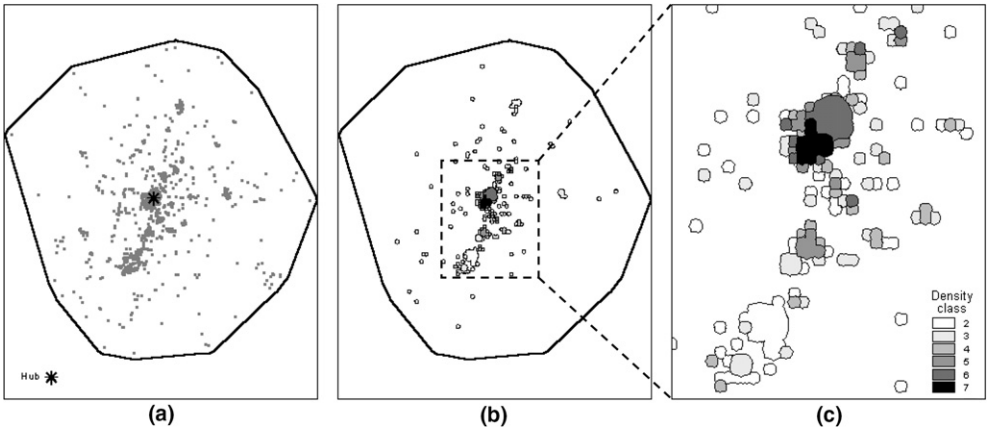

Fig. 3. Geo-ProZone clustering (a) buffered convex hull, (b) clusters, (c) detail of clusters at local scale.

class 6 or above that can be usefully interpreted as separate 'hot spots' (numbers 1–6 in Fig. 4) together with a seventh at density class 5 that represents a comparatively large, separated and distinct unit (number 7 in Fig. 4). An eighth has been inserted to represent a group of lower density GPZ (number 8 in Fig. 4) that whilst not as 'hot' as the others, nevertheless represents a sizeable number of more dispersed (rural) customers. Thus the suggestion is that $k = 8$ and that the N initial centroids are the centroids of the eight chosen GPZ polygons. The choice of 'hot spots' and therefore initial values for k and N do admittedly have a subjective element. However, whilst sensitivity analysis can be carried

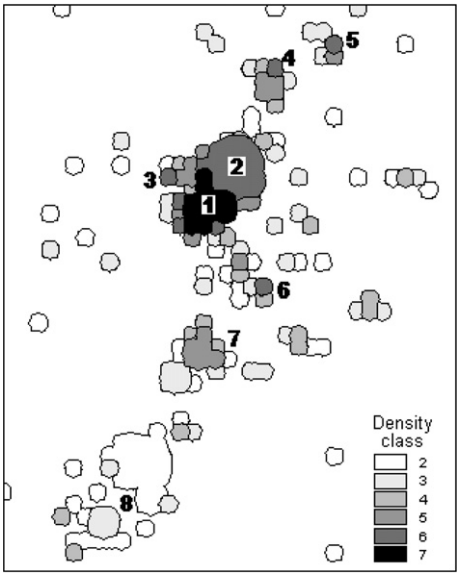

Fig. 4. $k = 8$ centroids suggested by Geo-ProZone clustering (note: for clarity, numbers 3–8 have been placed adjacent to the density polygon used to establish centroid co-ordinates).

out as part of the analysis (see below), this juncture of sentient human intervention allows an important pause for reflection and interpretation of the outcome of the GPZ (or any other chosen method of 'hot spot' clustering) and what the implications are likely to be for the next stage of analysis.

The decision to go forward with $k = N = 8$ initiates further structuring of the data prior to being submitted to the k-means algorithm. The $N = 8$ customers nearest to the candidate centroids from the GPZ clustering were ordered as in Fig. 4 and placed at the top of the data set to represent the candidate centroids from within the data set. In order to make the spatial relationship of each data record to the candidate centroids more explicit, a set of distance measures were used instead of the two X and Y co-ordinate variables. Although the k-means algorithm would use the two X and Y co-ordinate variables (along with all the other dimensions) in calculating distance in the data space to the N candidate centroids, it is argued that a set of distance measures would give more weight to the spatial relationships. Thus eight new variables DIST1 to DIST8 were calculated to represent the Euclidean distance from each customer to each candidate centroid. These new distance variables and SPEND were then normalised using the technique of *robust normalisation* (Brimicombe, 1999, 2000). 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 normalisation using z-scores. The binary GB-Profiles variables were given a slight stretch so that binary absence [0] was re-scaled to $[-1]$ to fall in line with the interquartile range $[-1, +1]$ of the robust normalised variables. An analysis of the robust normalised values of SPEND showed that 10 cases could be deemed outliers and likely to bias the k-means clustering; they were omitted from further analysis. The k-means clustering could then be run. The first run used only the distance variables without other attributes (Fig. 5(a)) and then

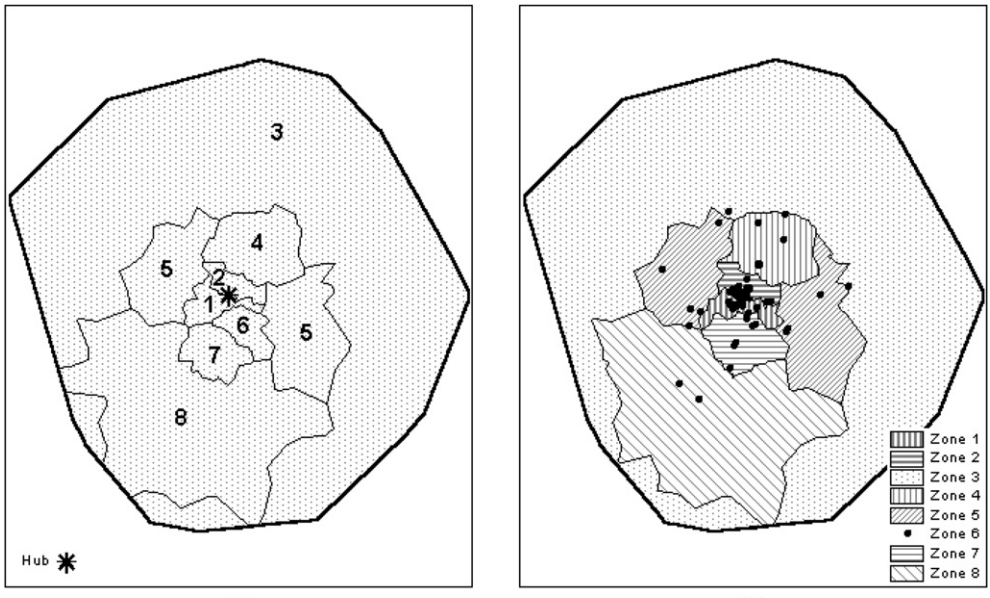

Fig. 5. k-means classification: (a) $k = 8$, distance only, (b) $k = 8$, all variables.

Table 2
Characteristics of $k = 8$ clusters (zones 1–8 in Fig. 5(b)); med. = median, tri. = trimean

$k = 8$	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%

using the SPEND and GB-Profile variables (Fig. 5(b)). The characteristics of each of the resulting clusters are summarised in Table 2.

The *k*-means clustering in Fig. 5(a) is based solely on the distance attributes DIST1 to DIST8 and has resulted in a mutually exclusive spatial segmentation (zones) with all variables significant at $p < .001$. Of note is that cluster 5 in Fig. 5(a) has resulted in two spatially separated zones to the east and west of the hub. This may be influenced by the fact that the *N* candidate centroids formed a broadly linear arrangement from southwest to northeast permitting the two parts of cluster 5 to emerge orthogonal and roughly equidistant to this axis. Fig. 5(b) has included all variables. From a spatial perspective clusters 1–5 and 8 in Fig. 5(a) remain as zones 1–5 in Fig. 5(b) with only slight modification; clusters 6 and 7 in Fig. 5(a) have merged into zone 7 in Fig. 5(b). With only seven spatially mutually exclusive zones in Fig. 5(b) one cluster, zone 6, has 'floated free' to be a cluster that is not spatially mutually exclusive with the others and is a surprise outcome of the *k*-means clustering. All variables are significant at $p < .001$ except for GB-PROF5 which is not significant with $p > 0.1$. If GB-PROF5 is removed from the *k*-means clustering, cluster memberships remain exactly the same and all variables are significant at $p < .001$. Table 2 summarises the characteristics of the zones (Z1–Z8) including median distance of customers within each zone to the business hub and trimean of SPEND where trimean (Tukey, 1977) is defined as:

$$(\text{lower quartile} + (\text{median} \times 2) + \text{upper quartile})/4$$

From inspection of Table 2, the eight zones have quite different characteristics, either spatially and/or in their attributes. That variable GB-PROF5 is not significant is perhaps not surprising given that it accounts for only a small percentage of customers and is spread across five zones. Although zones 1 and 2 could initially be viewed as being part of the same cluster in Figs. 1(b) and 3(c), their characteristics have emerged as being quite different. Z1 in Table 2 has the lowest trimean SPEND and is dominated by GB-PROF7 (*climbing*) in contrast to its spatial neighbour Z2 which has higher trimean SPEND and more than half of its membership characterised by GB-PROF8 (*established*) and GB-PROF9 (*prospering*). Another surprise is Z8, which is comparatively far from the hub, has 39% of its membership in three *struggling* geodemographic classes and yet has the third highest trimean of SPEND. But

probably of most interest from a business perspective is Z6, 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 predominantly from the *aspiring*, *established* and *prospering* geodemographic classes.

Whilst this result can be deemed 'useful' from a business perspective in as much as:

- the five important spenders (Z6) have been separated out and profiled;
- a profiled segmentation of customers (spatially and by attribute) for the regional market area has been achieved;

it has to be recognised that an unknown number of other 'useful' and statistically significant segmentations may be achievable. Whilst it is not feasible to exhaustively test for all other possibilities, a level of sensitivity can be quickly and easily established by systematically reducing the initial number of 'hot spots' and thus reducing k and N. This is illustrated firstly in Fig. 6(a) and Table 3 for initial centroids 1 to 7 in Fig. 4 and secondly in Fig. 6(b) and Table 4 for initial centroids 1 to 6 in Fig. 4, in other words, the successive omission of 'hot spots' having lower densities of point events. As the number of 'hot spots' is reduced, so the clusters become increasing concentric around the hub. By $k = 6$, both GB-PROF3 and GB-PROF5 are not significant with $p > 0.05$ (their removal nevertheless results in all variables used having $p < 0.05$) and Z1 and Z2 are neither spatially distinct nor compact as segmentations (differentiated primarily by SPEND) and for mapping purposes have had to be combined (Fig. 6(b)). Z6, the higher spenders, remain distinct. Both $k = 8$ and $k = 7$ have little to choose from as significant market segmentations (spatially and by attribute), the difference resting on the inclusion of initial centroid 8 as a subjective

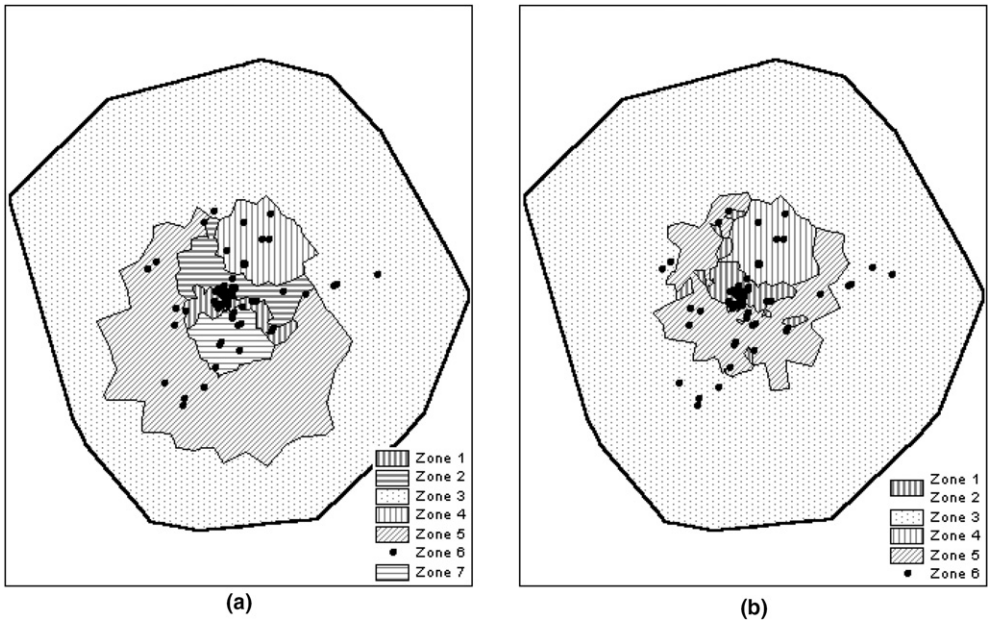

Fig. 6. Testing sensitivity: (a) $k = 7$, all variables, (b) $k = 6$, all variables.

Table 3
Characteristics of $k = 7$ clusters (zones 1–7 in Fig. 6(a)); med. = median, tri. = trimean

$k = 7$	$Z1$	$Z2$	$Z3$	$Z4$	$Z5$	$Z6$	$Z7$
n	643	480	199	186	208	78	230
Med. dist.	1127	1248	18057	5265	12354	2326	5862
Tri. spend	142	243	240	201	222	2762	223
GB-Prof1					2%		
GB-Prof2			3%		27%	1%	0.4%
GB-Prof3	7%	5%	4%	5%	7%	4%	13%
GB-Prof5	2%	2%	3%		3%		3%
GB-Prof6	7%	3%	7%	6%	13%	4%	4%
GB-Prof7	47%	13%	23%	52%	11%	35%	8%
GB-Prof8	9%	33%	24%	12%	12%	21%	26%
GB-Prof9	14%	31%	27%	24%	16%	29%	39%
GB-Prof10	15%	13%	11%		9%	6%	6%

Table 4
Characteristics of $k = 6$ clusters (zones 1–6 in Fig. 6(b)); med. = median, tri. = trimean

$k = 6$	$Z1$	$Z2$	$Z3$	$Z4$	$Z5$	$Z6$
n	836	227	359	199	338	65
Med. dist.	1068	1349	14,673	5163	6188	3685
Tri. spend	114	930	246	196	169	3073
GB-Prof1			1%			
GB-Prof2			17%		0.3%	2%
GB-Prof3	7%	5%	6%	5%	8%	3%
GB-Prof5	2%	1%	3%		2%	
GB-Prof6	6%	2%	11%	6%	4%	5%
GB-Prof7	36%	30%	19%	49%	6%	32%
GB-Prof8	16%	22%	13%	15%	32%	25%
GB-Prof9	18%	25%	18%	26%	43%	31%
GB-Prof10	14%	15%	11%		5%	3%

interpretation of the 'hot spot' clustering. However, any reduction on $k = 7$ in this instance breaks down as a segmentation that provides clusters that can be distinctively characterised both spatially and by attribute. The dual approach has thus, overall, allowed the analysis to quickly focus on $k = 8$ and $k = 7$ as rational candidate solutions. One might go on to speculate on increasing the N candidate centroids and hence $k = 9$ or higher as part of the sensitivity analysis. This is problematic as further 'hot spots' need to be recognised in the GPZ clustering and may not be justified. The outcome would be influenced by the spatial location of additional candidate centroid(s) and would result in further splits in the zones. Zones most likely to be split as k increases would be those with least homogeneity in their attributes, a likely candidate for $k = 9$ being Z8 in Table 2.

5. Conclusions

This paper has explored and demonstrated a dual approach in spatial data mining of point event data. The sequence has been:

- use a 'hot spot' style clustering of point events (in this instance, Geo-ProZones) treating each point as a binary event to suggest k number of classes centred on N initial candidate centroids (the 'hot spots') where $k = N$;
- create k new attributes for each point event being the Euclidean distance to each initial candidate centroid;
- bring the records spatially nearest to the N initial candidate centroids in order to the top of the data set;
- if necessary, normalise the data and check for outliers;
- run k-means clustering on all attributes using the first N records as candidate centroids;
- analyse the effectiveness of the approach.

The technique has been demonstrated on a business transaction database in order to achieve a significant customer segmentation both spatially and by attribute. Although the example used here has been a relatively small data set, it has allowed the workings of the technique to be explained and visualised. The variable resolution approach to producing GPZ clusters has shown itself to be effective for interpreting density 'hot spots'. Whilst the identification of 'hot spots' can be carried out by ranking the density of the GPZ clusters, it remains subjective. Hence the need for sensitivity analysis of rational candidate solutions. The overall dual approach can be used effectively to explore clusters in very large point event data sets.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. California: Sage.

Armstrong, M. P. (2000). Geography and computational science. *Annals of the Association of American Geographers, 90*, 146–156.

Atkinson, P. J., & Unwin, D. J. (2002). Density and local attribute estimation of an infectious disease using MapInfo. *Computers and Geosciences, 28*, 1095–1105.

Brimicombe, A. J. (1999). Small may be beautiful – but is simple sufficient? *Geographical and Environmental Modelling, 3*, 9–33.

Brimicombe, A. J. (2000). Constructing and evaluating contextual indices using GIS: A case of primary school performance. *Environment & Planning A, 32*, 1909–1933.

Brimicombe, A. J. (2003). *GIS, environmental modelling and engineering*. London: Taylor & Francis.

Brimicombe, A. J., & Tsui, H. Y. (2000). A variable resolution, geocomputational approach to the analysis of point patterns. *Hydrological Processes, 14*, 2143–2155.

Clark, P. J., & Evans, F. C. (1954). Distance to nearest neighbour as a measure of spatial relations in populations. *Ecology, 35*, 445–453.

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: Models and applications*. London: Pion.

Couclelis, H. (1998). Computation and space. 25th Anniversary Issue. *Environment & Planning B*, 41–47.

Estivill-Castro, V., & Lee, I. (2002). Argument free clustering for large spatial point-data sets via boundary extraction from Delaunay Diagram. *Computers, Environment and Urban Systems, 26*, 315–334.

Fotheringham, A. S. (1992). Exploratory spatial data analysis and GIS. *Environment and Planning A, 24*, 1675–1678.

Fotheringham, A. S. (1997). Trends in quantitative methods I: Stressing the local. *Progress in Human Geography, 21*, 88–96.

Fotheringham, A. S. (1998). Trends in quantitative methods II: Stressing the computational. *Progress in Human Geography, 22*, 283–292.

Fotheringham, A. S., & Brunsdon, C. (1999). Local forms of spatial analysis. *Geographical Analysis, 31*, 340–358.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography*. London: Sage.

Gahegan, M. (2003). Is inductive machine learning just another wild goose (or might it lay the golden egg)? *International Journal of Geographical Information Science, 17*, 69–92.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers, NS, 21*, 256–274.

Gatrell, A. C., & Rowlingson, B. S. (1994). Spatial point process modelling in a geographical information system environment. In A. S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 147–163). London: Taylor & Francis.

Halls, P. J., Bulling, M., White, P. C. L., Garland, L., & Harris, S. (2001). Dirichlet neighbours: Revisiting Dirichlet tessellation for neighbourhood analysis. *Computers, Environment and Urban Systems, 25*, 105–117.

Han, J., Kamber, M., & Tung, A. (2001). Spatial clustering methods in data mining. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 188–217). London: Taylor & Francis.

Harvey, D. W. (1966). Geographical processes and point patterns: Testing models of diffusion by quadrat sampling. *Transactions of the Institute of British Geographers, 40*, 81–95.

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis, 38*, 161–180.

Knox, E. G. (1964). The detection of space–time interactions. *Applied Statistics, 13*, 25–29.

Koch, T., & Denike, K. (2001). GIS approaches to the problem of disease clusters: A brief commentary. *Social Science & Medicine, 52*, 1751–1754.

Lawson, A. B. (2001). *Statistical methods in spatial epidemiology*. Chichester: Wiley.

Longley, P. A., Brooks, S. M., McDonnell, R., & MacMillan, B. (1998). *Geocomputation: A primer*. Chichester: Wiley.

Macmillan, W. (1998). Epilogue. In P. A. Longley et al. (Eds.), *Geocomputation: A primer* (pp. 257–264). Chichester: Wiley.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on maths and statistics problems*, Vol. 1. (pp. 281–297). Berkeley, CA.

Mantel, M. (1967). The detection of disease clustering and a generalised regression approach. *Cancer Research, 27*, 209–220.

Miller, H. J., & Han, J. (2001). *Geographic data mining and knowledge discovery*. London: Taylor & Francis.

Murray, A. T. (2000). Spatial characteristics and comparisons of interaction and median clustering models. *Geographical Analysis, 32*, 1–18.

Murray, A. T., & Estivill-Castro, V. (1998). Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science, 12*, 431–443.

Openshaw, S. (1994). Two exploratory space–time attribute pattern analysers relevant to GIS. In A. S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 83–104). London: Taylor & Francis.

Openshaw, S. (1998). Building automated geographical analysis and explanation machines. In P. A. Longley et al. (Eds.), *Geocomputation: A Primer* (pp. 95–115). Chichester: Wiley.

Openshaw, S., & Abrahart, R. J. (2000). *GeoComputation*. London: Taylor & Francis.

Openshaw, S., & Blake, M. (1996). *GB Profiler 91*. Department of Geography, University of Leeds.

Openshaw, S., Charlton, M. E., Wymer, C., & 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*, 359–377.

Phillips, J. D. (1999). Spatial analysis in physical geography and challenge of deterministic uncertainty. *Geographical Analysis, 31*, 359–372.

Rowlingson, B. S., & Diggle, P. J. (1993). Splancs: Spatial point pattern analysis code in S-Plus. *Computers and Geosciences, 19*, 627–655.

Samet, H. (1984). The quadtree and related hierarchical data structure. *Computing Surveys, 16*, 187–260.

Sleight, P. (1997). *Targeting customers: How to use geodemographic and lifestyle data in your business*. Henley-on-Thames: NTC Publications.

Snow, J. (1855). *On the mode of communication of cholera*. London: Churchill Livingstone.

Sokal, R., & Sneath, P. (1963). *Principles of Numerical Taxonomy*. San Francisco: Freeman.

Tsui, H. Y., & Brimicombe, A. J. (1997a). Adaptive recursive tessellations (ART) for Geographical Information Systems. *International Journal of Geographical Information Science, 11*, 247–263.

Tsui, H. Y., & Brimicombe, A. J. (1997b). Hierarchical tessellations model and its use in spatial analysis. *Transactions in GIS, 2*, 267–279.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Unwin, D. (1996). GIS, spatial analysis and spatial statistics. *Progress in Human Geography, 20*, 540–551.