Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data

Vladimir Estivill-Castro and Ickjai Lee
Department of Computer Science & Software Engineering
The University of Newcastle
Callaghan, NSW 2308, Australia
{vlad, ijlee}@cs.newcastle.edu.au

Abstract. We incorporate two knowledge discovery techniques, clustering and association-rule mining, into a fruitful exploratory tool for the discovery of spatio-temporal patterns. This tool is an autonomous pattern detector to reveal plausible cause-effect associations between layers of point and area data. We present two methods for this exploratory analysis and we detail algorithms to effectively explore geo-referenced data. We illustrate the algorithms with real crime data. We demonstrate our approach to a new type of analysis of the spatio-temporal dimensions of records of criminal events. We hope this will lead to new approaches in the exploration of large volumes of spatio-temporal data.

Keywords. Spatio-temporal association rules. Exploratory data analysis. Knowledge Discovery and Data Mining.

1. INTRODUCTION

Learning cause-effect associations and exercising control over a complex phenomenon like crime can only be accomplished by domain knowledge and thorough understanding. Raw data is too detailed and thus the next step is to apply statistical analysis. Nevertheless, despite being informed about a phenomenon, we may still lack an understanding about causes or plausible hypotheses to explain patterns in the phenomenon.

Identifying salient features contributing to frequency concentration of crime is essential to prevention. While geocomputation applications have penetrated the first two stages, namely facilitating data collection and statistical report generation, advanced autonomous techniques in exploratory analysis are not widely adopted. Knowledge comes from sophisticated and exploratory analysis of the relations, associations and statistics in the data. Data collection results in data-rich environments but the bottleneck is in analysis and hypothesis generation towards knowledge discovery.

Traditional spatial statistical analytical methods have been dominant approaches for identifying spatial patterns. However, these traditional approaches are computationally expensive and confirmatory. In addition, they necessitate prior information and domain knowledge (Miller and Han, 2001). That is, these methods confirm known or expected patterns, but do not easily detect unknown or unexpected patterns residing in large spatial databases. Thus, spatial statistical analysis becomes inappropriate and unsuitable for data-rich environments (Miller and Han, 2001; Openshaw, 1999; Openshaw and Alvanides, 1999).

Detecting clusters of crime incidents is important to crime activity analysis (Levine, 1999). When clustering for a particular layer (Besag and Newell, 1991; Marshall, 1991; Openshaw, 1987) in data-poor environments, it is not impossible to consider information in other layers that possibly impact on the target layer and thus likely function as a population-at-risk since the combination of background layers is limited to a small number. However, it is a daunting task to incorporate all possible background information in massive spatial databases. In such data-rich environments, one alternative is to find patterns of concentrations in the target layer and then investigate possible causal factors (crime generators) or associations (Estivill-Castro and Murray, 1998; Knorr *et al*., 1996; Knorr *et al*., 1997) based on spatial clusters. This approach minimizes human-generated bias and constraints on data, which is important in exploratory spatial analysis (Openshaw, 1987; Openshaw, 1999).

This paper reports on research for autonomous (that is, little or no intervention from a criminologist) computer diagnosis and pattern detection for generating highly likely and plausible hypotheses using data mining techniques: clustering and association rules. The autonomous pattern detector reveals cause-effect associations and aggregated groups of spatial concentrations leading to expanding the knowledge and understanding of crime. Our techniques incorporate Knowledge Discovery and Spatial Data Mining (KD-SDM) into the analysis of the spatio-temporal dimensions of the computer records of criminal events. Real crime data is used to demonstrate the potential of this approach. The results of exploratory data analysis by KD-SDM techniques do not reveal information about individuals, but exhibit general patterns.

The rest of paper is organized as follows. Section 2 revisits clustering and association-rule mining. In Section 3, we propose two association-rule mining approaches using spatial clusters. Section 4 provides experimental results with real crime data sets that confirm the virtue of our approach. Section 5 draws conclusions.

2. CLUSTERING AND ASSOCIATION RULES

Clustering and association-rule mining are two core techniques in spatial data mining (Miller and Han, 2001) and geographical data mining (Openshaw, 1999). Clustering is closely related to intensity measurement (first order effect) whilst association-rule mining is more related to dependency measurement (second order effect). Thus, the combination of these two techniques will reveal the structure of complex geographical phenomena, since the first order effect and the second order effect formulate geographical phenomena (Bailey and Gatrell, 1995).

Spatial clustering is a series of processes grouping a set of georeferenced point-data, $P = \{p_1, p_2, \ldots, p_n\}$ in some study region S, into smaller homogeneous subgroups due to contiguity (proximity in space). Several spatial clustering approaches have been proposed (Eldershaw and Hegland, 1997; Ester *et al*., 1996; Estivill-Castro and Houle, 1999; Estivill-Castro and Lee, 2000a; Estivill-Castro and Lee, 2000b; Kang *et al*., 1997; Karypis *et al*., 1999; Ng and Han, 1994; Wang *et al*., 1997; Wang *et al*., 1999; Zhang *et al*., 1996; Zhang *et al*., 2001) to detect patterns of spatial concentrations in large spatial databases. They have their strengths and weaknesses. Detected spatial aggregations are indicative of interesting areas (global hot spots and localized excesses) that require further analysis to find causal factors or possible correlations (Estivill-Castro and Lee, 2000a). Thus, clustering answers queries like "Where do clusters occur?'', "How many cluster reside in *S*?'' and provides further suggestions for investigation, like "Why they are there?''. Despite of the wealth of clustering methods, relatively little research (Estivill-Castro and Murray, 1998; Knorr *et al*., 1997) has been conducted on post-clustering analysis (cluster reasoning and correlation analysis using clusters). This is due to difficulties with efficiency, effectiveness and degree of autonomy in clustering methods, difficulty of cluster shape extraction and lack of adequate correlation measures. Recent clustering methods (Estivill-Castro and Lee, 2000b; Karypis *et al*., 2000) using dynamic thresholds overcome typical problems of traditional clustering methods that use global thresholds (Eldershaw and Hegland, 1997; Ester *et al*., 1996; Estivill-Castro and Houle, 1999; Kang *et al*., 1997; Ng and Han, 1994; Wang *et al*., 1997; Wang *et al*., 1999; Zhang *et al*., 1996). Thus, they are able to identify quality clusters including clusters in heterogeneous densities and clusters in different sizes and shapes. In addition, Lee (Lee, 2001) proposed a robust automatic cluster shape detection method for post-clustering analysis. It derives cluster boundaries and approximates shapes of clusters with the boundaries. These recent advances set the scene for post-clustering analysis.

Association-rule mining has been a powerful tool for discovering correlations among massive databases (Agrawal *et al*., 1993). An association rule is an expression in the form of $X \Rightarrow Y$ (c%), where X is the *antecedent* and Y is the *consequent*, X and Y are sets of items in transactional databases, and $X \cap Y = f$. It is interpreted as "c% of data that satisfy X also satisfy Y''. Here, a relational table summarizes a set of records. The relational table has a number of rows corresponding to transactions and a number of columns corresponding items. The value of an item for a given record is "1" if the item is in the transaction, "0" otherwise. From the relational table, we mine association rules that correlate the presence of a set of items with another set of items. Each rule has an associated *support* and *confidence*. Defined as follows

$$\textit{Support} \text{ is an estimate for } Pr[X \cap Y],$$

$$\textit{confidence} \text{ is an estimate for } Pr[X \cap Y] / Pr[X].$$

The *support* is the ratio of transactions that satisfy both X and Y to the number of transactions in databases. The *confidence* is the conditional probability of Y given X. Since users are interested in large *support* and high *confidence* (*strong rules* (Koperski and Han, 1995)), two thresholds (*minimum support* and *minimum confidence*) are used for pruning rules to find strong association rules. The association-rule mining is to compute all association rules satisfying user-specified *minimum support* and *minimum confidence* constraints. Although association-rule mining is popular in data mining community (Agrawal *et al*., 1993; Agrawal and Srikant, 1994; Fayyad *et al*., 1996, Fu and Han, 1995), few research (Estivill-Castro and Murray, 1998; Koperski and Han, 1995) has been conducted on spatial association rules. The rule

$$\textit{is\_a}(x, \text{house}) \wedge \textit{near\_by}(x, \text{beach}) \Rightarrow$$

$$\textit{is\_expensive}(x) \ (95\%).$$

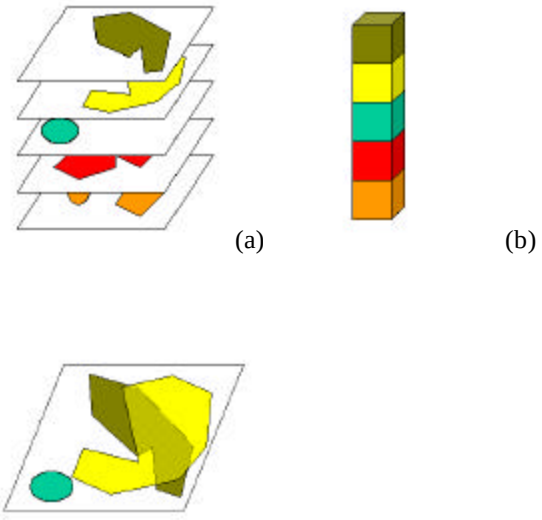
is discovered by spatial association-rule mining (Koperski and Han, 1995). The rule suggests that 95% of houses attached to beaches are expensive. This traditional association-rule mining uses predicates that contain either spatial-spatial relationship (*near_by*) or spatial-aspatial relationship (*is_a* and *is_expensive*). Here, predicates are pre-defined before association-rule mining and only rules that are somehow related to the predicates are extracted. Defining predicates requires domain knowledge and necessitates concept hierarchies that are difficult to define. For instance, it is hard to classify a house (US$200,000) as expensive or not. The decision purely depends on domain knowledge or personal decision. Thus, concept hierarchies are domain-specific and thus applicability is limited to case at hand. Also, this traditional association-rule mining is specialized for extended-relational databases and SAND architectures. Thus, this is not well-suited for layer-based multivariate association mining.

3. MINING MULTIVARIATE ASSOCIATIONS USING CLUSTERING

Crime hot spot analysis is one of popular techniques to understand complex crime activities (Levine, 1999). We identify these crime hot spots with cluster analysis and find possible cause-effect relations with association-rule mining. In this section, we propose two approaches for cluster association-rule mining. Figure 1: explains a vertical-view approach and a horizontal-view approach.

In this example, we consider five geographical layers as depicted in Figure 1:(a). If we pinpoint a location within *S*, the location will have five associated attributes corresponding to the five layers. Figure 1:(b) shows these 5 attributes vertically (referred as an attribute cube in this paper). Values of attributes become true (1) if the location lies within regions (clusters) of corresponding layers, false (0) otherwise. The vertical-view approach tries to discover interesting associations from the whole set of attribute cubes. For instance, an association rule "$layer(1) \wedge layer(2) \Rightarrow layer(4)$ (70%)" is derived if 70% of attribute cubes satisfying attribute values of layer 1 and layer 2, also have the value true in layer 4.

Figure 1: Multivariate associations mining.



(a) (b)

(c)

The horizontal-view approach overlays all the layers into a target layer, and then attempts to find associations from the target layer using intersection (overlapping) areas. Figure 1:(c) overlays the first, second and third layers depicted in Figure 1:(a). The first and second layers intersect while the third layer does not intersect with the other two. Thus, the association between the first layer and second layer is higher than that of the first and third and that of the second and third.

In the two approaches, we only consider spatial clusters of point-data as candidates for mining associations. These aggregated spatial groups of concentrations represent and summarize the distribution of *P*. We believe that there are some reasons (possibly attracters) for spatial concentrations. That is, particular environments (crime generators) attract crimes and thus result in concentrations. Spatial association mining is to find possible attracters or some positive contributors to spatial clusters. Noise points are the points that are not affected by the attractors. Thus, they are ignored for mining associations. For this task, we need a threshold to differentiate clusters from noise points. The threshold is not an absolute number, but a ratio in this paper since each point-data layer has different number of points within it. This will be further discussed later in this section. Following subsections will discuss the two proposed approaches in detail.

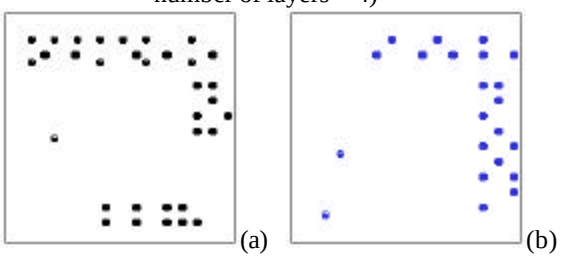
3.1 Vertical-view Approach

The vertical-view approach is similar to the popular raster model in the sense that they model space with regular cells. The algorithmic procedure of the vertical-view approach is as follows:

1. Find spatial clusters for point-data layers.
2. Segment all the layers with a finite number of regular cells (rectangles).
3. Construct a $m \times n$ relational table with the binary {0,1} values.
4. Apply association-rule mining to the table.

We first compute spatial clusters of point-data by cluster analysis. After that, we segment *S* and construct attribute cubes. With attribute cubes, we build a relational $m \times n$ table with the binary domain {0,1}, where *m* denotes the number of attribute cubes and *n* denotes the number of layers. In the table, $m_i[n_j]$= 1 if an attribute cube m_i satisfies event in a layer n_j. Finally, we mine associations that correlate the presence of a set of layers with another set of layers. Since the relational table of layer-based GIS is exactly the same as that of transactional databases except layers replace items and locations replace transactions, it is now straightforward to discover associations among layers using traditional association-rule mining. We illustrate this with an example.

Figure 2: Vertical-view approach with 4 cells (the number of layers = 4)



(a) (b)

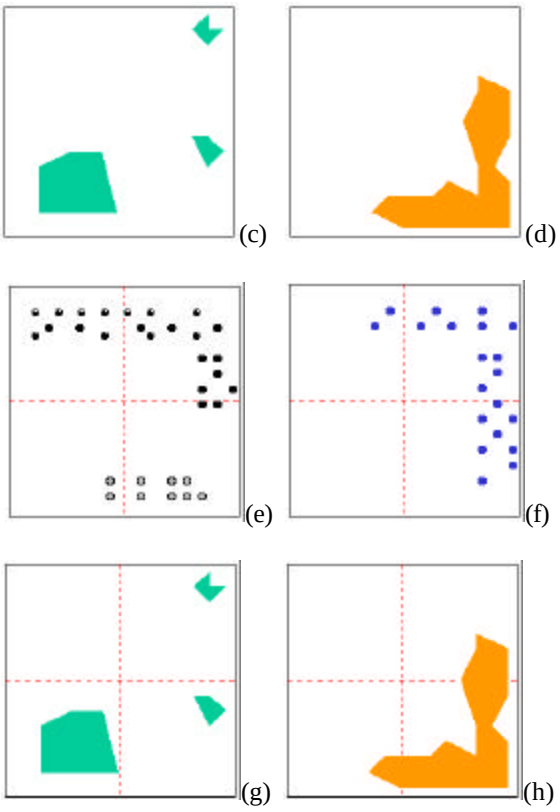


Figure 2: and Table 1 illustrate the vertical-view approach with 4 geographical layers. Let us assume that Figure 2:(a)-(d) show railway stations (point), crime incidents (point), parks (area) and urban areas (area), respectively. The first process is to find homogeneous groups of spatial concentrations of point-data layers. Two clusters of railway stations and one cluster of crime incidents are shown in Figure 2:(e) and Figure 2:(f), respectively. Noise points are ignored. Then, we frame S with collectively exhaustive and mutually exclusive cells. A 2×2 grid is used in this case. After that, we compute a 4×4 relational table before we apply association-rule mining. Table 1 describes the relational table.

Table 1: 4×4 table.

	layer(1)	*layer*(2)	*layer*(3)	*layer*(4)
loc(1)	1	1	0	0
loc(2)	1	1	1	1
loc(3)	1	0	1	1
loc(4)	1	1	1	1

In the table, *layer*(j) ($0 \leq j \leq n$) in columns denotes j-th geographical layer, *loc*(i) ($1 \leq i \leq m$) in rows denotes i-th cell (attribute cube) in S (numbered in Morton order, Z). t[*loc*(i) , *layer*(j)] is 1 if (clustered) event in the j-th layer occurs in *loc*(i) cell, and t[*loc*(i) , *layer*(j)] is 0 otherwise. For instance, t[*loc*(1) , *layer*(1)] = 1 because members of clusters of railway stations lie within the top-left cell as depicted in Figure 2:(e). We mine multivariate associations from this relational table. One of association rules is as follows:

layer(1) $\wedge$ *layer*(2) $\Rightarrow$ *layer*(4) (66.7%).

With 50% *support* (*loc*(2) and *loc*(4) out of 4 attribute cubes), 66.7% of locations, that are near-by railway stations and have crime incidents, fall within urban areas. In this approach, the granularity of cells plays a critical role.

Figure 3: Vertical-view approach with 16 cells.

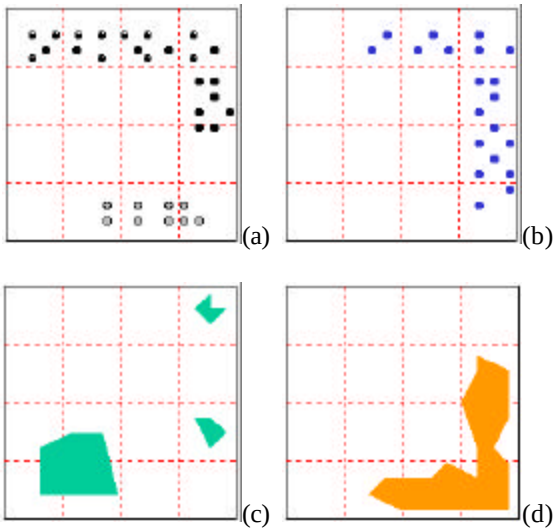


Table 2: 16×4 table.

	layer(1)	*layer*(2)	*layer*(3)	*layer*(4)
loc(11)	1	0	0	0
loc(12)	1	1	0	0
loc(13)	0	0	0	0
loc(14)	0	0	0	0
loc(21)	1	1	0	0
loc(22)	1	1	1	0
loc(23)	0	0	0	0
loc(24)	1	1	0	1
loc(31)	0	0	1	0
loc(32)	0	0	1	0
loc(33)	0	0	1	0
loc(34)	1	0	1	1
loc(41)	0	0	0	0
loc(42)	1	1	1	1
loc(43)	1	0	0	1
loc(44)	1	1	0	1

Figure 3: and Table 2 illustrate the vertical-view approach with a 4×4 grid for the same dataset shown in Figure 2:. Now, the bottom-right cell (attribute cube *loc*(44)) in Figure 3: has values (1, 1, 0, 1) since the cell contains (clustered) events of layers *layer*(1), *layer*(2), and *layer*(4), but *layer*(3). With 50% *support* constraint, now we are not able to find the rule "*layer*(1) $\wedge$ *layer*(2) $\Rightarrow$ *layer*(4)", since its *support* is only 18.8% (3/16). However, the *confidence* is now 100%, since all the locations *loc*(24), *loc*(42) and *loc*(44) satisfying *layer*(1) and *layer*(2), also satisfy *layer*(4).

One of advantages of the vertical-view approach is that it is easy to apply transactional association-rule mining techniques (Agrawal *et al*., 1993; Agrawal and Srikant, 1994; Fu and Han, 1995), since the vertical-view approach uses relational tables. However, it has a main drawback. That is, rules discovered by the vertical-view approach are heavily dependent on the granularity that is difficult to determine. Similar difficulties are found when modeling point-data using raster-like approach.

3.2 Definitions

The following definitions are made to explain the unique process of association-rule mining using regions of clusters. Let P be a data layer storing point data and R a real value in [0,1].

Definition 1. Clusters with Ratio R of P [denoted $CwR(P)$] are the clusters C detected by a clustering approach (algorithm) whose normalized sizes (the number of points in a cluster in C / the total number of points in P) are equal to or greater than R.

Let X and Y be sets of layers (for example $X_1 = \{P_1\}$ is a one layer expression). The expression X will typically identify the *antecedent* while we use Y for the *consequent*. In our approach, $clusters\_areas(X_1)$ is a set of polygonized clusters (regions of clusters) of a point-data layer X_1 (the point-data is converted to area-data). The function $clusters\_areas(X)$ (also sometimes we will denote this as $clusters\_areas(antecedent)$) is the total area of the regions that result of the intersection of $clusters\_areas(X_i)$, for all X_i in X. That is, consider the overlay of $clusters\_areas(X_i)$, for all X_i in X and find the regions that correspond to points $clusters\_areas(X_i)$, for all X_i in X. The total area of these regions is $clusters\_areas(X)$.

Definition 2. The Clustered Support (CS) is the ratio of the area that satisfy both the *antecedent* and the *consequent* to the area of study region S. That is,

$$CS = (clusters\_areas(antecedent) \cap clusters\_areas(consequent)) / area(S).$$

Definition 3. The Clustered Confidence (CC) for a rule $X \Rightarrow Y$ is the conditional probability of areas of CwR of the *consequent* given areas of CwR of the *antecedent*. That is,

$$CC = clusters\_areas(X \cup Y) / clusters\_areas(X).$$

Note that, $clusters\_areas(X \cup Y)$ is the area of the regions where points are in a cluster for all layers. Thus, they are in the (vertical) intersection of the layers. But, the set of layers is the union of the layers in X and the layers in Y.

Definition 4. A Clustered Spatial Association Rule ($CSAR$) is an expression in the form of

$$X \Rightarrow Y\ (CC\%), \text{ for } X \cap Y = f\ .$$

The interpretation is as follows: CC% of areas of clusters of X intersect with areas of clusters of Y.

3.3 Horizontal-view Approach

Since the vertical-view approach needs a parameter to determine the granularity, it is regarded as an argument-dependent approach. Argument-tuning to find best-fit values is not only difficult task, but very expensive in terms of time consumption in data-rich environments. Thus, KD-SDM favors argument-less or argument-free approaches in order to reduce preprocessing time and to minimize inconsistency of results. This is obvious from the fact that the argument-free modeling, the Voronoi diagram and its dual the Delaunay triangulation, has gained popularity in point-data clustering (Eldershaw and Hegland, 1997; Estivill-Castro and Lee, 2000a; Estivill-Castro and Lee, 2000b; Kang *et al*., 1997) as an alternative to overcome drawbacks of typical argument-dependent modeling methods such as k-nearest neighbor neighboring, d-distance neighboring and raster-like cell-based modeling. In this subsection, we propose another approach for mining multivariate associations that minimizes the need for user supplied parameters.

The algorithmic procedure of the horizontal-view approach is as follows:

1. Find $CwR(P)$ for point-data layers P in X and Y.
2. Extract cluster boundaries of each CwR for point-data layers in X and Y.
3. Compute the value of the areas of CwR for point-data layers and the areas of area-data layers.
4. Overlay the *antecedent* and the *consequent*.
5. Apply association-rule mining to detect $CSAR$s.

The vertical-view approach approximates shapes of clusters and then polygonizes clusters with their boundaries. Thus, it requires an effective clustering and a robust cluster-to-area transformation to generate accurate association rules.

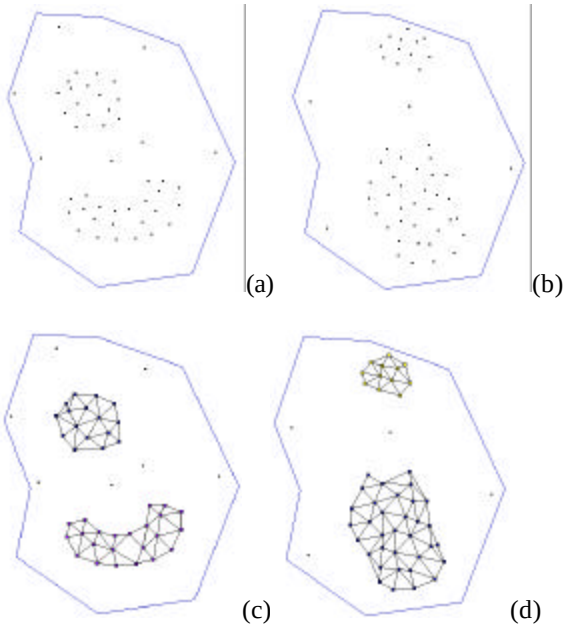
Recently, Estivill-Castro and Lee (Estivill-Castro and Lee, 2000b) proposed a boundary-based clustering that utilizes dynamic thresholds rather than static thresholds. It requires the Delaunay diagram as an underlying proximity graph and performs clustering on the proximity graph. It removes inconsistently long Delaunay edges for all $p \in P$ and removes inter-cluster Delaunay edges to detect various types of clusters. It detects quality clusters including non-convex clusters (unlike partitioning clusterings (Estivill-Castro and Houle, 1999; Ng and Han, 1994)) and clusters with heterogeneous densities (unlike density-based clusterings (Ester *et al*., 1996; Openshaw, 1987) and grid-based clusterings

(Wang *et al*., 1997; Wang *et al*., 1999)). It requires one control value to explore the structure of distribution of *P*. Smaller values of the control value produce strongly cohesive clusters (smaller) while larger values of the control value provide relatively less cohesive clusters (larger). Typically, setting the control value to 1 produces quality clustering in most cases for two-dimensional point-data. Thus, we also use 1 as the control value in this paper unless otherwise noted.

Lee (Lee, 2001) proposed an algorithm that extracts cluster boundaries and polygonizes clusters with their extracted boundaries. The cluster-to-area transformation does not demand user-supplied parameters to detect shapes of clusters, but derives boundaries of clusters from the distribution of *P*. Points within clusters are not the only contributor to the shape of clusters, but points belonging to other clusters affect the shape of the cluster. It is able to polygonize not only non-convex clusters, but clusters with holes (voids). Now, what is available is summarized area-data rather than point-data (Lee, 2001). This area-data approximates shapes of clusters of point-data and represents spatial concentrations where most points are aggregated. For these reasons discussed above, we use the boundary-based clustering (Estivill-Castro and Lee, 2000b) and the cluster boundary extraction approach (Lee, 2001) for mining associations with the horizontal-view approach.

Now, we illustrate definitions made in Section 3.2 and the procedure of the horizontal-view approach with synthetic datasets shown in Figure 4:.

Figure 4: The process of the horizontal-view approach.



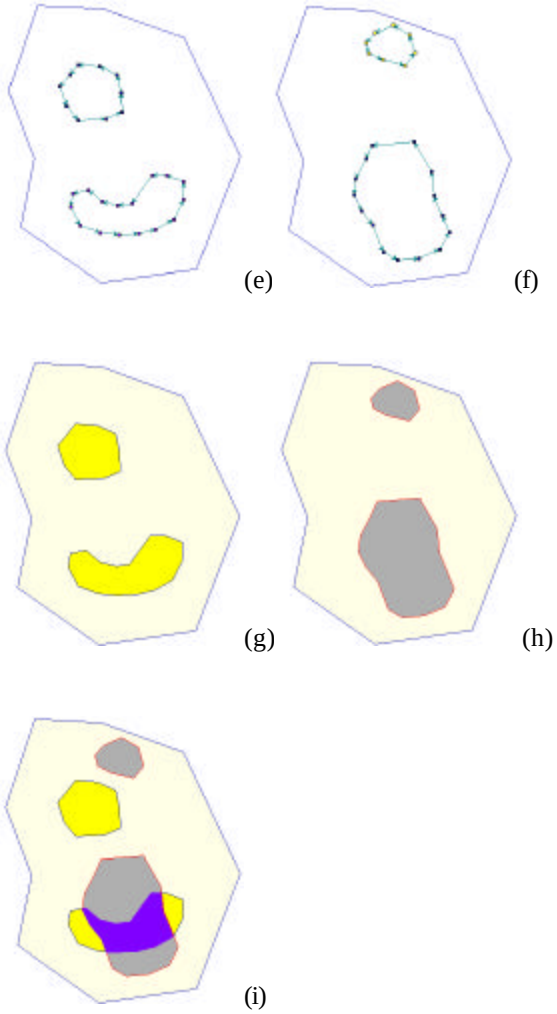


Figure 4:(a) illustrates a dataset I (the number of points = 46) while Figure 4:(b) presents dataset II (the number of points = 50) with a common study region. Both datasets have two spatial concentrations detected by the boundary-based clustering and those are illustrated in Figure 4:(c) and Figure 4:(d), respectively. We now apply the cluster boundary extraction process (Lee, 2001) and cluster boundaries that approximate shapes of clusters are illustrated in Figure 4:(e) and Figure 4:(f). We polygonize clusters with boundaries and their regions of clusters are depicted in Figure 4:(g) and Figure 4:(h), respectively. Figure 4:(i) illustrates an overlay of regions of clusters of dataset I and those of dataset II. Visual inspection indicates that *clusters_areas*(dataset I) and *clusters_areas*(dataset II) intersect. Thus, we may notice that dataset I and dataset II are somehow correlated although we are not able to quantitatively define the association between dataset I and dataset II solely based on visual inspection. Quantitative analysis is displayed with some numerical indicators in Table 3.

Table 3: Statistics of dataset I and dataset II.

	clusters_areas	*CS*(%)	*CC*(%)
S	6940.14	100.0	N/A

dataset I	992.04	14.29	N/A
dataset II	1312.21	18.91	N/A
dataset I $\Rightarrow$ dataset II	401.46	5.78	40.47
dataset II $\Rightarrow$ dataset I	401.46	5.78	30.59

The area of study region *area*(*S*) is 6940.14, the total area of the regions covered by clusters of dataset I is denoted by *clusters_areas*(dataset I) and its value is 992.04. For the other dataset, *clusters_areas*(dataset II) is 1312.21. The intersection area of *clusters_areas*(dataset I) and *clusters_areas*(dataset II) is 401.46 and thus *CS* of dataset I and dataset II is 5.78% (401.46/6940.14).

With 5% of *CS*, we are able to derive two association rules. They are as follows:

Rule 1: dataset I $\Rightarrow$ dataset II (40.47% *CC*),

Rule 2: dataset II $\Rightarrow$ dataset I (30.59% *CC*).

Rule 1 indicates that around 40% (401.46/992.04) of locations belonging to regions in clusters of dataset I also belong to regions in clusters of dataset II. That is, around 40% of incidents of dataset I occur where incidents of dataset II take place. **Rule 2** is similarly interpreted, but with 30.59% clustered confidence. We see that the vertical-view approach quantitatively defines asymmetric associations that suggest highly likely and plausible hypotheses.

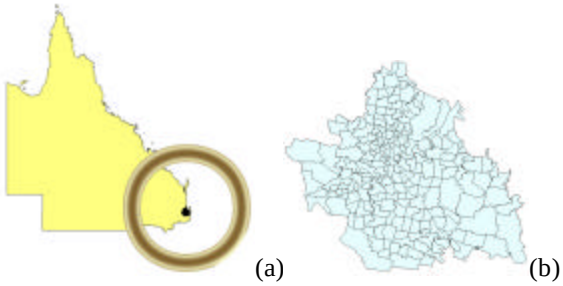
Since the horizontal-view approach is autonomous, it is better suited for mining massive databases than the vertical-view approach. Thus, the horizontal-view does not necessitate domain knowledge, but maximizes user friendliness. We examine complex real crime data with the horizontal-view approach in next section to confirm the virtue of the approach.

4. PERFORMANCE EVALUATION WITH REAL DATASETS

Similar to most urban areas, understanding of crime activity in the south east Queensland region[1] of Australia, where the capital city of Brisbane is located, is important for regional planners and criminologists as well as policing agencies. Figure 5: displays the state of Queensland in Australia and the south east of Queensland that continues to experience significant and sustained population growth (Murray *et al*., 2001; Stimson and Taylor, 1999). However, raw crime data[2] in this region are too complex and extremely huge, thus it seems to be a difficult task even for domain experts to detect valuable patterns of crime incidents.

Crime statistics provided by Queensland Police Service have three main categories: offences against the person, offences against property and other offences. Offences against the person consist of subcategories: homicide, assaults, sexual offences, robbery, extortion, kidnapping and other offences against the person. Offences against property are composed of breaking and entering, arson, other property damage, motor vehicle theft, stealing, fraud and other offences against property. Other offences include drug offences, prostitution, liquor, gaming offences, trespassing and vagrancy, good order offences, traffic and related offences and miscellaneous offences. In addition, the subcategories could have several subsubcategories. For instance, homicide consists of attempted murder, conspiracy to murder, driving causing death and manslaughter.

Figure 5: Queensland in Australia and the south east study region of Queensland with 217 suburbs.



The complex structure of crime data is not the only concern for crime activity analysis. We must consider feature data (parks, railway stations, schools etc.) along with crime data to detect the relationship of crime activity to salient features. The volume of data becomes easily beyond the capability of human analysis if temporal crime data are considered. Thus, sophisticated data mining tools, that autonomously suggest highly likely plausible hypotheses, are greatly demanding to deal with explosive databases.

Geographic information systems and crime mapping software have been dominant tools for exploring crime activity (Murray and Shyy, 2000). However, these tools are not quantitatively exploratory, but rather visually assistant with choropleth mapping, buffering, overlaying, intersecting and containment operations. Although these naïve operations may provide valuable insights into the complex raw crime data, these are not suitable for data-rich environments.

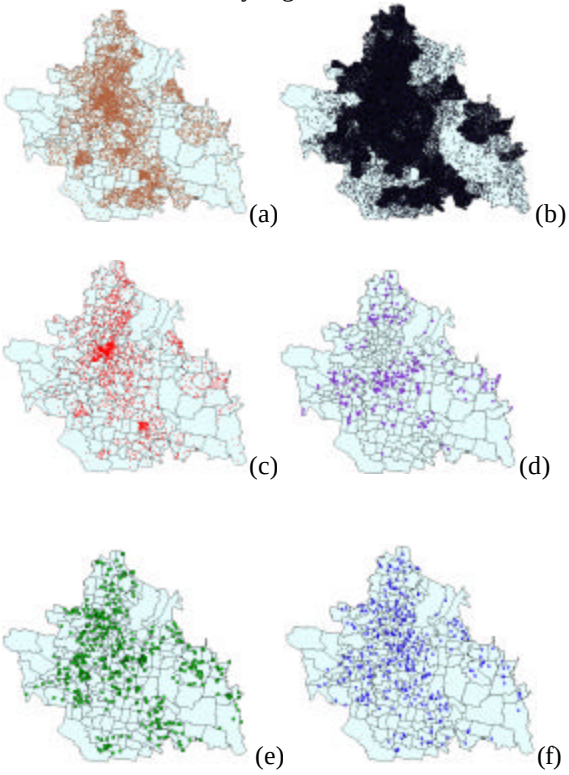
Figure 6: depicts three main categories of crime occurred in the year of 1997 and three feature data in study region. Although visual displays provide general

[1] We consider 217 suburbs around Brisbane as a study region in this paper.

[2] We use crime data recorded by Queensland Police Service in the year of 1997 in the study region.

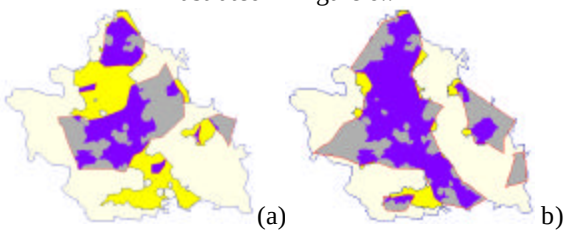
patterns, too much information causes confusion. For illustration purpose, we only display three main types of crime and three feature data in this paper. Figure 6:(a) depicts 9,618 incidents of offences against the person. And, Figure 6:(b) displays 113,618 incidents of offences against property while Figure 6:(c) shows 2,124 incidents of other offences in the region. Feature data, reserves (249), parks including caravan parks (462) and schools (306), are shown from Figure 6:(d) to Figure 6:(f).

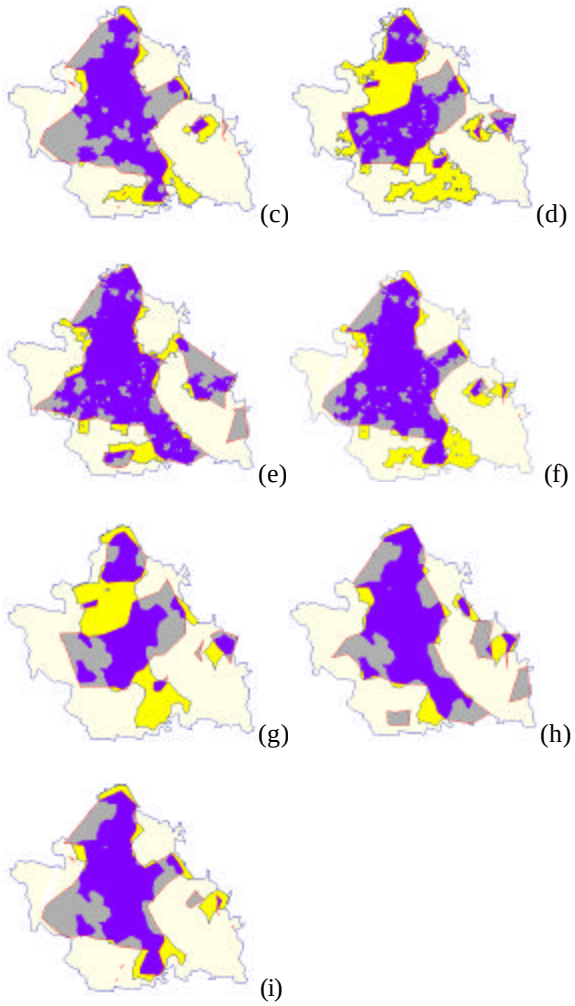
Figure 6: Three main categories of crime incidents and three feature data in study region.



Clustering summarizes complex distributions of raw crime data and polygonization of clusters provides regions of concentrations as hot spots. Figure 7: illustrates the horizontal-view association-rule mining with cluster regions of real data depicted in Figure 6:.

Figure 7: Illustration of association-rule mining with the horizontal-view approach with the real crime data illustrated in Figure 6:.





Figures from Figure 7:(a) to Figure 7:(c) display overlaps between *clusters_areas*(offences against the person) and *clusters_areas*(feature data). Similarly, overlaps between *clusters_areas*(offences against property) and *clusters_areas*(feature data) are described from Figure 7:(d) to Figure 7:(f) and overlaps between *clusters_areas*(other offences) and *clusters_areas*(feature data) are depicted from Figure 7:(g) to Figure 7:(i). Visual inspection interprets parks and schools are more associated with crime than reserves.

This is quantitatively described in Table 4. Clustered confidence (*CC*) of a *CSAR* "Offences against the person $\Rightarrow$ Reserves (44.93%)" is much lower than *CC* of *CSAR*s "Offences against the person $\Rightarrow$ Parks (85.29%)" and "Offences against the person $\Rightarrow$ Schools (77.5%)". These rules imply that more than 70% of locations, where offences against the person occur, are around parks and schools.

Table 4: *CS* and *CC* of *CSAR*s of the real crime data.

	CS(%)	*CC*(%)
Offences against the person $\Rightarrow$ Reserves	15.40	44.93

Reserve ⇒ Offences against the person	15.40	50.99
Offences against the person ⇒ Parks	29.23	85.29
Parks ⇒ Offences against the person	29.23	57.33
Offences against the person ⇒ Schools	26.56	77.50
Schools ⇒ Offences against the person	26.56	59.85
Offences against property ⇒ Reserves	20.83	47.44
Reserves ⇒ Offences against property	20.83	68.99
Offences against property ⇒ Parks	36.25	82.56
Parks ⇒ Offences against property	36.25	71.10
Offences against property ⇒ Schools	33.42	76.11
Schools ⇒ Offences against property	33.42	75.31
Other offences ⇒ Reserves	17.81	50.47
Reserves ⇒ Other offences	17.81	58.97
Other offences ⇒ Parks	29.90	84.74
Parks ⇒ Other offences	29.90	58.64
Other offences ⇒ Schools	28.35	80.36
Schools ⇒ Other offences	28.35	63.89

Possible *CSAR*s will increase exponentially as the number of layers under study (in this case, crime types and features) grows. Thus, it is almost impossible for analysts to find interesting associations manually. The horizontal-view approach automatically generates strong *CSAR*s with user-specified *CC* and *CS*. For instance, with 30% *minimum CC* and 75% *minimum CS*, three strong *CSAR*s, when the size of the *antecedent* and the size of *consequent* are 1, are discovered. These are as follows:

Crime rule 1: Offences against property ⇒ Parks

(36.25% *CS*, 82.56% *CC*),

Crime rule 2: Offences against property ⇒ Schools

(33.42% *CS*, 76.11% *CC*),

Crime rule 3: Schools ⇒ Offences against property

(33.42% *CS*, 75.31% *CC*).

These rules imply that most offences against property (more than 75%) are taking place around parks and schools. Further, with more than 75% *CC*, locations of schools imply occurrences of offences against property. Many hypothesis can now be derived for input into a confirmatory analysis. For example, this very simple illustration already suggests that possibly residents living around schools shall consider additional actions for reducing crime against their property.

5. FINAL REMARK

For the analysis of crime data, Knowledge Discovery and Data Mining (KD-DM) techniques have been applied by the FBI as a part of the investigation of the Oklahoma City bombing the Unabomber case, and many lower-profile crimes (Berry and Linoff, 1997). KD-DM has also been used by Treasury Department of the US to hunt for suspicious patterns in international funds transfer records; patterns that may indicate money laundering or fraud (Berry and Linoff, 1997). Australian examples are the Health Insurance Commission of Australia using KD-DM to investigate fraud within the Medicare system (He *et al*., 1998) and NRMA Insurance Ltd (Williams, 1999). However, in the above examples the techniques have been limited to the attribute-oriented side of the data and KD-DM has been attempted to the spatio-temporal dimensions.

The where and when are crucial for patterns of crime. Recent progress towards developing GIS for crime analysis emphasizes the relevance that geo-reference has to understanding patterns in crime (Hirschfield *et al*., 1995). In a limited exploratory approach (Openshaw, 1994), systems have been developed to geographically visualize patterns of vehicle crime, domestic burglaries, drug-related crime and public disorder in inner city areas.

For example, hypotheses are typically human-generated and reinforced with the display of the most contrasting socio-demographic characteristics of areas with high levels of criminal activity. This type of approach may lead to hypotheses that relate the location patterns of crime to the negative socio-demographic characteristics of areas associated with these patterns (Gandhi and Grubbs, 1996). The application of GIS technology to crime analysis by mapping and display may use statistical methods to produce maps delineating risk surfaces for crime data (Lee, 1995). These are confirmatory analyses that for example, have contributed to the consideration of geographic distribution of criminal behavior as a primary factor for planning residential neighborhoods (Gandhi and Grubbs, 1996).

However, today vast collections of spatio-temporal data are gathered without any previous hypothesis, and less as parts of a structured experiment. Moreover, it is impossible that trained researchers examine all possible interesting patterns in such huge amounts of data. We require an intelligent assistant to process the data and to autonomously (or at least with very little guidance) analyze data. We have presented methods and algorithms that increase even more the capacity of GIS to become intelligent pattern spotters beyond just as repositories to store, manipulate and retrieve data. Note that, this direction complements traditional statistical analyses on geo-referenced data. The intent is to design and deploy autonomous partners in suggesting hypothesis in the knowledge discovery process. We have developed algorithms that will enable this exploratory capability in computers within the context of spatial data and crime data.

We have developed a supporting application[3]. It supports visualization of cluster boundaries and mining association rules.

REFERENCES

Agrawal, R., Imielinski, T. and Swami, A. (1993) "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of the 1993 ACM SIGMOD Conference*, pp. 207-216.

Agrawal, R. and Srikant, R. (1994) "Fast Algorithms for Mining Association Rules", *Proceedings of 1994 International Conference on Very Large Data Bases*, pp. 487-499.

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive Spatial Analysis*. Wiley: New York.

Berry, M.J.A. and Linoff, G. (1997) *Data Mining Techniques --- for Marketing, Sales and Customer Support.* Wiley: New York.

[3] The application is implemented in the C++ programming language using LEDA (Library of Efficient Data types and Algorithms) version 4.2. For documentation and code see http://www.algorithmic-solutions.com/.

Besag, J. E. and Newell, J. (1991) "The Detection of Clusters in Rare Diseases", *Journal of the Royal Statistical Society A*, vol. 154, pp. 143-155.

Eldershaw, C. and Hegland, M. (1997) "Cluster Analysis using Triangulation", in Noye, B. J, Teubner, M. D. and Gill, A. W. (eds.), *Computational Techniques and Applications: CTAC97*. World Scientific: Singapore, pp. 201-208.

Ester, M., Kriegel, H. –P., Sander, J. and Xu, X. (1996) "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226-231.

Estivill-Castro, V. and Murray, A. T. (1998) "Discovering Associations in Spatial Data – An Efficient Medoid based Approach", *Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 110-121.

Estivill-Castro, V. and Houle, M. E. (1999) "Robust Clustering of Large Geo-referenced Data Sets", *Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 327-337.

Estivill-Castro, V. and Lee, I. (2000a) "AMOEBA: Hierarchical Clustering Based on Spatial Proximity Using Delaunay Diagram", *Proceedings of the 9th International Symposium on Spatial Data Handling*, pp. 7a.26-7a.41.

Estivill-Castro, V. and Lee, I. (2000b) "AUTOCLUST: Automatic Clustering via Boundary Extraction for Mining Massive Point-Data Sets", *Proceedings of the 5th International Conference on Geocomputation*, GC049.

Estivill-Castro, V. and Lee, I. (2000c) "AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles", *Proceedings of the International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining*, Springer Verlag Lecture Notes in Artificial Intelligence 2007, pp. 131-144.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Fu, Y. and Han, J. (1995) "Meta-Rule-Guided Mining of Association Rules in Relational Databases", *Proceedings of the 1st Workshop on Integration Knowledge Discovery with Deductive and Object-Oriented Databases*, pp. 39-45.

Gandhi, S. and Grubbs, J.W. (1996) "Crime analysis using GIS at the City of Orlando, Florida: implications for housing policies", *Proceedings of the URISA Annual Conference*, pp. 122-132.

He, N., Graco, W. and Yao, X. (1998) "Applications of Genetic Algorithms and k-Nearest Neighbour Methods in Medical Fraud Detection", *Proceedings of the 2nd Asia-Pacific Conference on Simulated Evolution and Learning*

SEAL-98, Springer Verlag Lecture Notes in Computer Science 1585, pp. 74-81.

Hirschfield, A., Brown, P. and Todd, P. (1995) "GIS and the analysis of spatially-referenced crime data: Experiences in Merseyside, UK", *International Journal of Geographic Information Systems*, vol. 9, no. 2, pp. 191-210.

Kang, I., Kim, T. and Li, K. (1997) "A Spatial Data Mining Method by Delaunay Triangulation", *Proceedings of the 5th International Workshop on Advances in Geographic Information Systems*, pp. 35-39.

Karypis, G., Han, E. and Kumar, V. (1999) "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", *IEEE Computer: Special Issue on Data Analysis and Mining*, vol. 32, no. 8, pp. 68-75.

Knorr, E. M., Ng, R. T. and Shilvock, D. L. (1996) "Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Minnings", *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 884-897.

Knorr, E. M., Ng, R. T. and Shilvock, D. L. (1997) "Finding Boundary Shape Matching Relationships in Spatial Data", *Proceedings of the 5th International Symposium on Spatial Databases*, pp. 29-46.

Koperski, K. and Han, J. (1995) "Discovery of Spatial Association Rules in Geographic Information Databases", *Proceedings of the 4th International Symposium on Large Spatial Databases*, pp. 47-66.

Lee, I. (2001) "Polygonization of Point Clusters through Cluster Boundary Extraction for Spatial Data Mining", *Technical Report 2000-08*, The University of Newcastle, Australia.

Lee, J. (1995) "Point pattern analysis with density estimation for frequency data", GIS/LIS-95 Annual Conference and Exposition, *American Soc. Photogrammetry and Remote Sensing*, pp. 598-607.

Levine, N. (1999) "CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations", *Proceedings of the 4th International Conference on Geocomputation.*

Marshall, R. J. (1991) "A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease", *Journal of the Royal Statistical Society, A*, vol. 154, pp. 421-441.

Miller, H. J. and Han, J. (2001) *Geographic Data Mining and Knowledge Discovery: An Overview*. Cambridge University Press: Cambridge, in press.

Mehlhorn, K. and Naher, S. (1999) *LEDA A platform for combinatorial and geometric computing*. Cambridge University Press: Cambridge.

Murray, A. and Shyy, T. (2000) "Integrating Attribute and Space Characteristics in Choropleth Display and Spatial Data Mining", *International Journal of Geographical Information Science*, vol. 14, pp. 649-667.

Murray, A., McGuffog, I., Western, J. and Mullins, P. (2001), "Exploratory Spatial Data Analysis Techniques for Examining Urban Crime", *British Journal of Criminology*, vol. 41, pp. 309-329.

Ng, R. T. and Han, J. (1994) "Efficient and Effective Clustering Method for Spatial Data Mining", *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 144-155.

Openshaw, S. (1987) "A Mark 1: Geographical Analysis Machine for the Automated Analysis of Point Data Sets", *International Journal of Geographic Information Science*, vol. 1, no. 4, pp. 335-358.

Openshaw, S. (1994) "Two exploratory space-time-attribute patterns analysers relevant to GIS", in Fotheringham, S. and Rogerson, P. (eds.), *Spatial Analysis in GIS*, Taylor and Francis: IK, pp. 83-104.

Openshaw, S. (1999) "Geographical Data Mining: key design issues", *Proceedings of the 4th International Conference on Geocomputation*.

Openshaw, S. and Alvanides, S. (1999) "Applying Geocomputation to the Analysis of Spatial Distributions", in Longley, P. A., Goodchild, M. F., Maguire, D. J. and Rhind, D. W. (eds.), *Geographical Information Systems: Principles and Technical Issues*, John Wiley & Sons: New York, vol. 1, pp. 267-282. second edition.

Stimson, R. and Taylor, S. (1999) "City profile: Brisbane", *Cities*, in press.

Wang, W., Yang, J. and Muntz, R. (1997) "STING: A Statistical Information Grid Approach to Spatial Data Mining", *Proceedings of the 23rd International Conference on Very Large Data Bases*, pp. 144-155.

Wang, W., Yang, J. and Muntz, R. (1999) "STING+: An Approach to Active Spatial Data Mining", *Proceedings of the International Conference on Data Engineering*, pp. 116-125.

Williams, G. J. (1999) "Evolutionary Hot Spots Data Mining", *Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-99*, Springer Verlag Lecture Notes in Artificial Intelligence 1574, pp. 184-193.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996) "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 103-114.

Zhang, B., Hsu, M. and Dayal, U. (2001) "K-Harmonic Means – A Spatial Clustering Algorithm with Boosting", *Proceedings of the International Workshop on Temporal Spatial and Spatio-Temporal Data Mining*, pp. 31-45.