Criminal Cross Correlation Mining and Visualization

Peter Phillips and Ickjai Lee

Discipline of IT
James Cook University
Australia
{peter.phillips,ickjai.lee}@jcu.edu.au

Abstract. Criminals are creatures of habit and their crime activities are geospatially, temporally and thematically correlated. Discovering these correlations is a core component of intelligence-led policing and allows for a deeper insight into the complex nature of criminal behavior. A spatial bivariate correlation measure should be used to discover these patterns from heterogeneous data types. We introduce a bivariate spatial correlation approach for crime analysis that can be extended to extract multivariate cross correlations. It is able to extract the top-k and bottom-k associative features from areal aggregated datasets and visualize the resulting patterns. We demonstrate our approach with real crime datasets and provide a comparison with other techniques. Experimental results reveal the applicability and usefulness of the proposed approach.

1 Introduction

Since criminals are creatures of habit, law enforcement agencies can be more effective if they learn from historical data to better understand perpetrators habits and locations they choose to commit crimes. Police and policy makers need an intelligent crime analysis machine that is able to capture the connections that exist between places and events based on past crimes. These patterns can then be used to characterize criminal behavior and discover where, when and why particular crimes are likely to occur. Discovering correlations between crime and spatial features is a core component of intelligence-led policing, and allows for a deeper insight into the complex nature of criminal behavior. Crime activities are geospatially, temporally and thematically correlated therefore a variety of factors can contribute to the formulation of crime. These factors need to be considered when interpreting crime datasets.

Several crime data mining techniques have been developed over recent years [1,2,3,4], however reasoning about crime datasets has received less attention [5,6]. One of the drawbacks of these approaches is that they can typically only reason about positive associative features whereas negative associative features can be just as important. Several works using spatial association rules mining have been proposed in order to mine spatial associations in geospatial databases [7,8,9]. The

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 2–13, 2009.

main drawback of these approaches is that they capture point-to-point association but ignore spatial association.

The increasing availability of heterogeneous datasets, such as those incorporating socio-economic and socio-demographic factors, geospatial features and crime datasets, has increased the need for intelligent analysis. To discover correlations in these heterogeneous data types, a spatial bivariate correlation measure should be used. Standard bivariate techniques do not account for spatial relations in their computation and are therefore not well suited to identifying spatial structures [10]. Previous efforts in developing a bivariate association measure have been proposed by Wartenberg [11] who based their approach on Moran's I statistic. However, Lee [12] identifies drawbacks with Wartenberg's approach and instead proposes his own bivariate spatial association measure.

In this paper we rank crime and geospatial feature datasets by their spatial co-patterning (similarity) using bivariate correlation measures. This results in the discovery of patterns in the form $crimeX$ is more similar to $featureY$ than $featureX$. We also show that while Lee's L index [12] is an effective spatial correlation measure for crime analysis, it achieves inconsistent results when used with areal aggregated datasets over regions of irregular size and shape. To overcome this we present a more suitable weights matrix where the spatial lag better represents neighboring regions of irregular size. Using this weights matrix, we are then able to extract the top-k and bottom-k associative features. We also present a graph based visualization method that allows the user to easily extract patterns of interest.

Research in the area of crime analysis and reasoning, such as the approach outlined here, can be applied to a wide range of other applications that may lead to economic, social and environmental advantages. This technique could be applied to geospatial datasets in several fields including disaster management, epidemiology, business intelligence, geology, environmental monitoring, marketing and e-commerce. For example, in epidemiology, interesting cross correlations between cancer hotspots and environmental features can be identified.

In Section 2 we review existing bivariate spatial association measures, and perform comparisons between them. Section 3 compares two popular correlation measures and describes the problem statement. In Section 4 we detail the special properties of areal aggregated crime data and show how the choice of the spatial weights matrix affects the bivariate correlation measure. Section 5 provides experimental results with real crime datasets and our visualization technique. We conclude with final remarks and ideas for future work in Section 6.

2 Bivariate Spatial Association

Spatial association is the degree to which a set of univariate observations are similarly arranged over space. It quantifies the distribution of patterns among a dataset, with a strong spatial association occurring when two distributions are similar, weak association describes little similarity and no association occurs

when two distributions are dissimilar. Bivariate spatial association measures quantify the spatial relations among many variables in a set of observations.

Many processes, including crime, involve more than one variable, so allowing for their dependence on each other is essential in modeling and in understanding their covariance [13]. Hubert *et al.* [14] make a distinction between the relationship within a pair at each location (point-to-point association) and the relationship between distinct pairs across locations (spatial association). Pearson's r is a common point-to-point association measure, while spatial association is often measured by Moran's I. Pearson's correlation coefficient r for variables X and Y is computed by:

$$r_{X,Y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}}, \tag{1}$$

and Moran's I is given by:

$$I_X = \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum (x_i - \overline{x})^2}, \tag{2}$$

where i and j are two spatial objects and w_{ij} is a general spatial weights matrix of their locational similarity. The choice of the weights matrix is a critical step in computing the spatial association [10,15].

Several works using spatial association rules mining have been proposed in the data mining community in order to mine spatial associations in massive geospatial databases [7,8,9]. In general, these approaches extend traditional association rule mining [16,17] to the spatial and temporal context. One drawback of these approaches is that they typically capture point-to-point association but ignore spatial association. It is argued that to capture spatial co-patterning, a bivariate spatial association measure should be the combination of both the point-to-point association and spatial association measures [12]. That is, the bivariate measure should be a composite of the univariate spatial associations of two variables and their point-to-point association in a certain form.

There are two main approaches in this regard [11,12]. Wartenberg [11] developed a multivariate extension of Moran's I univariate spatial autocorrelation measure to account for the spatial dependence of data observations and their multivariate covariance simultaneously. Wartenberg's approach can be defined as:

$$I_{X,Y} = \frac{\sum_i (x_i - \overline{x})(\widetilde{y}_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \cong \sqrt{SSS_Y} \cdot r_{X,\widetilde{Y}}. \tag{3}$$

Lee [12] utilizes the concept of spatial lag (SL), which is defined as the composed weighted averages of neighbors defined by the spatial weights matrix, and is given as:

$$\widetilde{x} = \sum_j w_{ij} x_j \tag{4}$$

The SL is then used to introduce the concept of a spatial smoothing scalar (SSS) that can reveal substantive information about the spatial clustering of a variable. The SSS ranges from 0 to 1, where more clustered variables have higher SSS values. The SSS is given by:

$$SSS_x = \frac{n}{\sum_i (\sum_j w_{ij})^2} \cdot \frac{\sum_i (\sum_j w_{ij}(x_j - \overline{x}))^2}{\sum (x_i - \overline{x})^2}. \quad (5)$$

Lee's L index between two variables can then be defined as the Pearson's correlation coefficient between their SL vectors multiplied by the square root of the product of their SSSs:

$$L_{X,Y} = \sqrt{SSS_X} \cdot \sqrt{SSS_Y} \cdot r_{\widetilde{X},\widetilde{Y}}. \quad (6)$$

3 Correlation Measures and Problem Statement

We evaluate both Lee's L and Wartenberg's I bivariate correlation measures with a synthetic dataset that shows increasing dissimilarity. We use the same base region of 37 regular sized hexagons that Lee [12] uses to illustrate his bivariate spatial correlation measure. Lee compares his L index to Wartenberg's using simple dataset patterns and argues that Cross-MC should not be used as a bivariate spatial association measure.

Figure 1 starts with 4 clusters in $dataset_a$, and for each subsequent dataset we remove 1 cluster. As clusters are removed, $dataset_{b-d}$ become less similar to the original $dataset_a$. The density for non cluster regions are uniformly distributed random values and are the same among datasets. Table 1 and Fig. 1(e) illustrate the expected result; when we rank the similarity for $dataset_a$, the bivariate correlation measure decreases with the number of clusters removed. For this dataset, both Lee's L and Wartenberg's approach show correct results, however only Lee's L shows the expected linear decrease in correlation value.

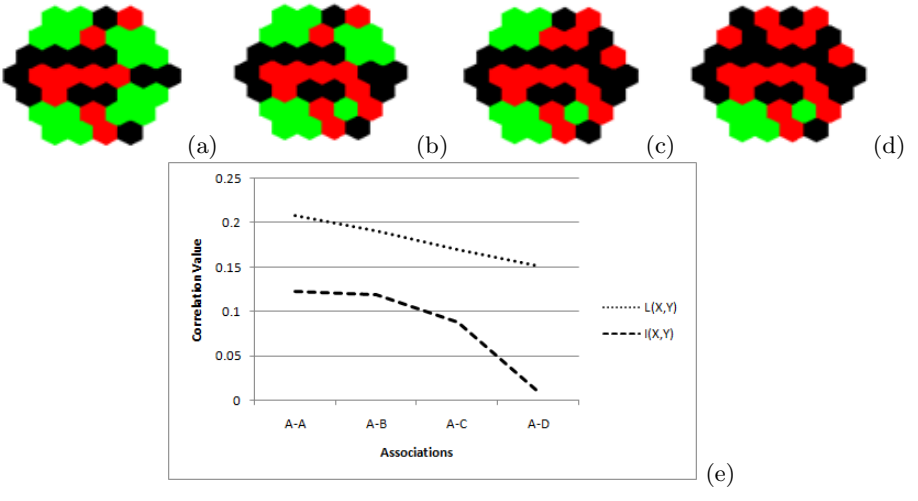

Fig. 1. Synthetic hexagon dataset: (a-d) Datasets $a-d$; (e) Comparison of Lee's L and Wartenberg's I

Table 1. Cross correlations from synthetic hexagon dataset

	SSS_X	SSS_Y	$r_{\tilde{X},\tilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$	$I_{X,Y}$
a-a	0.208	0.208	1.000	1.000	0.208	0.123
a-b	0.208	0.236	0.863	0.896	0.191	0.119
a-c	0.208	0.336	0.642	0.740	0.170	0.089
a-d	0.208	0.387	0.532	0.595	0.151	0.009

4 Shared Border Aware Weight Matrix

Crime, census and geospatial features are often recorded for regions of irregular size and shape. Crime datasets can be aggregated to suburb or police districts and census data using census tracts or collection districts. When the study region contains such irregular regions, the choice of weight matrix is a critical step in the computation of spatial association [15].

Figure 2 shows synthetic areal aggregated datasets over suburbs from Brisbane, Australia. Regions of grey color are low density while green colored regions are high density. This is a subset of the base regions we use in our real world experiments in Section 5. To allow for a deeper insight into the complex question of crime analysis we need to discover crime and spatial features that exhibit strong correlation (similarity). From visual analysis of our synthetic datasets, it can be argued that the ranking of similarity for $dataset_a$ is $dataset_a - dataset_b, dataset_c, dataset_d$. That is, $dataset_a$ is more similar to $dataset_b$ than $dataset_c$ and $dataset_d$. This is because these datasets show a similar density (green high density regions surrounded by grey low density regions) in a similar spatial neighborhood.

Table 2 shows that Lee's L index incorrectly determines that the similarity ranking is $dataset_a - dataset_d, dataset_c, dataset_b$. This is because Lee uses a simple weights matrix that is defined as the row standardized version of the binary connectivity matrix where elements that are neighbors have a value of 1 or otherwise a value of 0. Each neighbor is given the same weighting when calculating the spatial lag vector but with irregular regions this is often not desirable. For example in Fig. 2, $dataset_b$ and $dataset_c$ are both neighbors (share

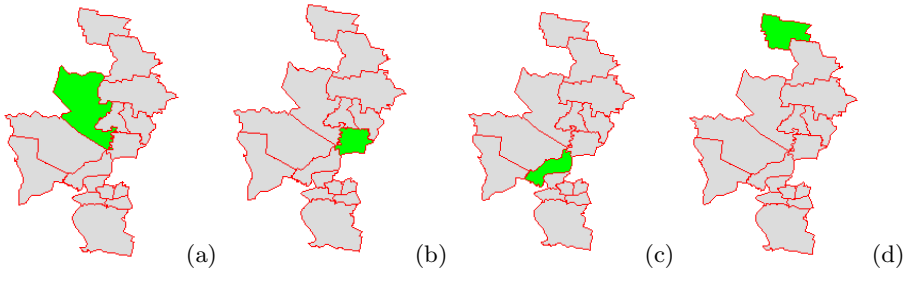

Fig. 2. Synthetic irregular region dataset: (a-d) Datasets $a - d$

Table 2. Cross correlations from irregular dataset

	SSS_X	SSS_Y	$r_{\widetilde{X},\widetilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$
a-a	1.000	1.000	1.000	1.000	1.000
a-b	1.000	0.865	0.028	0.833	0.027
a-c	1.000	1.000	0.040	0.833	0.041
a-d	1.000	0.741	0.448	0.833	0.388

Table 3. Cross correlations from irregular dataset with modified weights matrix

	SSS_X	SSS_Y	$r_{\widetilde{X},\widetilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$
a-a	0.994	0.994	1.000	1.000	0.994
a-b	0.994	0.817	0.152	0.833	0.137
a-c	0.994	0.929	0.112	0.833	0.108
a-d	0.994	0.712	-0.107	0.833	-0.0901

a border) with $dataset_a$, however $dataset_c$ shares only a very small border. The spatial weights matrix should be defined so that the common border reflects the weighting and thus the spatial lag. If i and j are neighbors, we define w as:

$$w(i,j) = sharedBorder_{ij}/totalBorder_i, \tag{7}$$

where $sharedBorder_{ij}$ is the shared border between i and j and $totalBorder_i$ is the total border of i.

Given this new spatial weights matrix, Table 3 shows that Lee's L index correctly identifies the similarity ranking of $dataset_a$ as $dataset_a - dataset_b$, $dataset_c$, $dataset_d$. Discovering patterns of similar crime and geospatial/census features is a key component to intelligence-led policing and crime analysis.

5 Experimental Results

This section shows the results collected from real crime datasets of urban suburbs of Brisbane, Australia. A steadily growing trend in recorded violent crime in Australia [18] has been a major concern not only to policing agencies, but also tourism agencies and the public. The Brisbane study region (Fig. 3) is highly dynamic and active and continues to experience significant and sustained population growth and an increase in various criminal activities [19]. The Queensland Police Service (QPS) releases crime data in areal aggregated format due primarily to privacy concerns. We combine these crime datasets with spatial feature datasets and census datasets so that associative patterns can be discovered. The smallest geographic unit at which census data is available from the Australian Bureau of Statistics (ABS) is the Collection District. The ABS produces statistics for other geographic areas by allocating collection districts to each spatial unit in the particular classification [20].

We use a total of 108 datasets in this experiment; 38 crime datasets, 7 geographic features (caravan parks, railway stations, reserves, schools, hospitals,

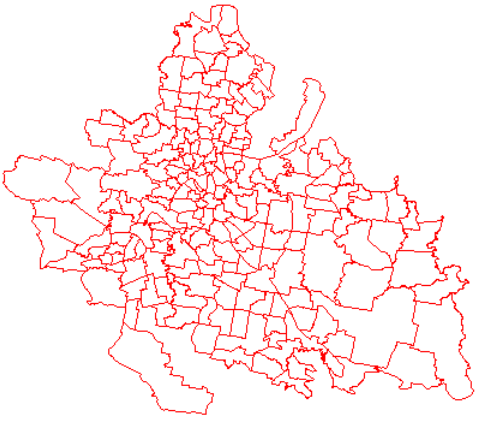

Fig. 3. Brisbane study region

Table 4. Cross correlations from the Brisbane crime dataset

top-k	$L_{X,Y}$
Unlawful Entry With Intent - UNOCCUPIED	0.603
Unlawful Entry With Intent - AGED_25_29	0.580
Unlawful Entry With Intent - UNEMPLOYED	0.564
Other Stealing - OVERSEAS_VISITOR	0.551
Other Theft - OVERSEAS_VISITOR	0.535
bottom-k	
Arson - AGED_85_89	-0.154
Arson - AGED_90_94	-0.152
Liquor (excl. Drunkenness) - SEPARATE_HOUSE	-0.139
Arson - FLAT_UNIT	-0.139
Rape and Attempted Rape - reserve	-0.135

university/colleges and parks) and 63 census classifications. The crime dataset from the QPS has three main categories: personal safety (offences against person), property security (offences against property) and other offences. The census classifications we use include age, dwelling structure, employment status, weekly income, level of education completed and household mobility indicator (change of address). The study region encompasses 186 suburbs of Brisbane that had crime and census data available.

We utilize Lee's L index with the shared border length weights matrix as described in Section 4. We mine the dataset to extract the top-k and bottom-k spatial associative patterns involving salient features or census classifications (for this experiment $k = 5$). We restrict the patterns to those involving at least one type of crime. From the extracted associations shown in Table 4 we can see that there is a strong correlation between the crime *Unlawful Entry With Intent* and *Unoccupied Dwellings* and *Unemployed 25-29 year old persons*. It can be seen that there is also a weak association between *Arson* and *persons aged 85-94*. This information can

then be used to further investigate the cause of these specific associative patterns and also allow for targeted crime prevention efforts.

5.1 Visualization

While the associations shown in Table 4 may be common sense, there may be interesting, unknown patterns in the larger subset of top-k and bottom-k associative patterns. The problem becomes one of information overload; how can the user find interesting patterns hidden amongst the other patterns.

Figure 4 shows the visualization that we developed to help the user extract interesting patterns. For this example we use the same datasets as described in Section 5 and extract the top-50 results. Each node (feature) is labeled and its size is a depiction of the correlation strength, where the size of the circle represents the largest L index score of its relationships. The edges between nodes in our visualization depict the top-k associative patterns. The correlation strength is depicted by the edge thickness and color.

Figure 4 clearly shows that the top-50 associative patterns form two clusters, one revolving around *Overseas Visitors, Other Dwelling - Improvised home* and the second cluster around *Unemployed, Age 25-29, Unoccupied Dwelling, Unlawful Entry With Intent, Unlawful Entry Without Violence*. The visualization environment is dynamic so that the user can move and delete any associations that are not of interest. Figure 5 shows the visualization of the bottom-50 associative patterns. In this case the, the visualization easily allows the user to see that *Arson* is not caused by *People aged 80-100*, locations with *Flat or Units* and locations of *Universities and Colleges*.

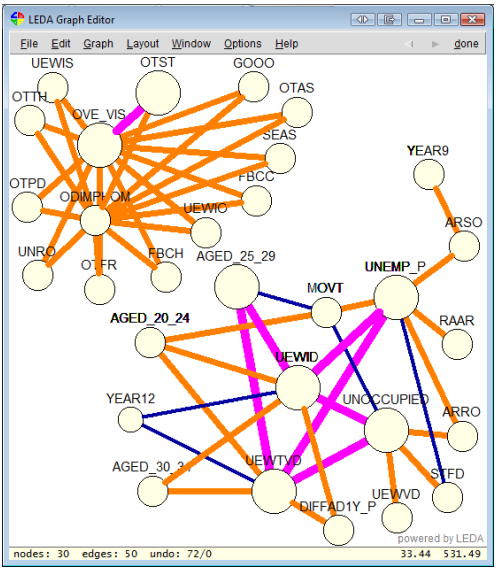

Fig. 4. Visualization of top-50 associative patterns

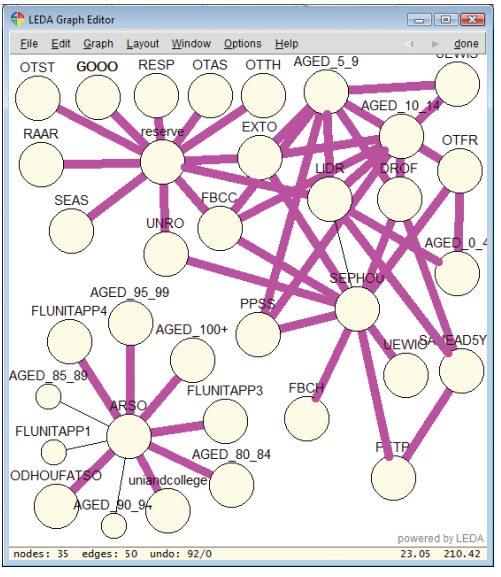

Fig. 5. Visualization of bottom-50 associative patterns

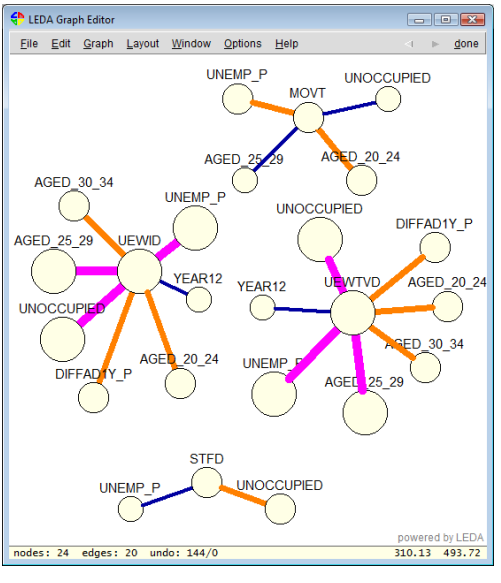

Fig. 6. Visualization of selected top-k crime associative patterns

Users can also select specific crime features to visualize from the top-k/ bottom-k results. Figure 6 shows the associative patterns when the crimes *Unlawful Entry With Intent, Unlawful Entry Without Violence, Motor Vehicle Theft and Stealing from Dwellings* from the top-50 are selected. In this visualization, duplicate geospatial and census features are not removed. As part of a crime decision support system these associative patterns can be used to further investigate and explain the behavior of crime within the study region.

5.2 Comparison with Density Tracing

We compare the spatial associative patterns discovered by the bivariate spatial association approach with our previous Density Tracing approach [21]. Density tracing allows autonomous exploratory analysis and knowledge discovery in areal aggregated crime datasets by transforming the dataset into a density trace that shows density change between regions. The results from the two approaches are not directly comparable, however the associative patterns in the form $crimeX$ is more similar to $featureY$ than $featureX$ are comparable.

Table 5. Cross correlations using Density Tracing

Patterns	*dissimilarity*
Unlawful Entry With Intent - UNEMPLOYED	0.425599
Unlawful Entry With Intent - AGED_90_94	0.45303
Unlawful Entry With Intent - AGED_20_24	0.459116
Other Stealing - OVERSEAS_VISITOR	0.115436
Other Theft - OVERSEAS_VISITOR	0.116533

We examine the top-5 patterns from Table 4 against Density Tracing. We use each crime in the top-5 as a reference feature f in the Density Tracing algorithm and select the most similar non-crime feature. For *Unlawful Entry With Intent* we report the three most similar features. From the results in Table 5 it can be seen that for *Other Stealing* and *Other Theft* both approaches report the same most similar associative patterns. The results for *Unlawful Entry With Intent* are due to the differences in the way the two approaches calculate similarity.

6 Final Remarks

Since crime activities are geospatial phenomena, they are geospatially, thematically and temporally correlated. Thus, crime datasets must be interpreted and analysed in conjunction with various factors that can contribute to the formulation of crime. We propose a bivariate spatial association approach for crime analysis using an enhanced Lee's L index [12] to extract the top-k and bottom-k associative features. We have shown why the choice of spatial weight matrix is an important consideration for spatial association analysis of crime datasets of

irregular size and shape. In addition, we introduced a visualization approach that helps users find interesting patterns.

This research is part of an ongoing project with the aim to build a crime knowledge discovery machine, as a crime decision support system for large areal-aggregated crime datasets, that explains the behavior of crime with the first order effect of crime (concentrations and deviations) and the second order effect of crime (links and associations). Future work includes incorporating temporal data into the bivariate approach. We wish to compare the results from this study to census and crime data for other years. We are also investigating graph based implementations for computational and memory efficiency. A comparison between this technique and an Association Rules Mining approach such as [8] is also planned.

References

1. Chen, H., Atabakhsh, H., Zeng, D., Schroeder, J., Petersen, T., Casey, D., Chen, M., Xiang, Y., Daspit, D., Nandiraju, S., Fu, S.: Coplink: visualization and collaboration for law enforcement. In: Proceedings of the 2002 annual national conference on Digital government research, pp. 1–7 (2002)
2. Craglia, M., Haining, R., Wiles, P.: A Comparative Evaluation of Approaches to Urban Crime Pattern Analysis. Urban Studies 37(4), 711–729 (2000)
3. Hirschfield, A., Brown, P., Todd, P.: GIS and the Analysis of Spatially-Referenced Crime Data: Experiences in Merseyside. U. K. Journal of Geographical Information Systems 9(2), 191–210 (1995)
4. Ratcliffe, J.: The Hotspot Matrix: A Framework for the Spatio-temporal Targeting of Crime Reduction. Police Practice and Research 5, 5–23 (2004)
5. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime Data Mining: A General Framework and Some Examples. Computer 37(4), 50–56 (2004)
6. Oatley, G., Ewart, B., Zeleznikow, J.: Decision Support Systems for Police: Lessons from the Application of Data Mining Techniques to Soft Forensic Evidence. Artificial Intelligence and Law 14(1), 35–100 (2006)
7. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Proceedings of the 4th International Symposium on Large Spatial Databases, Portland, Maine. LNCS, pp. 47–66. Springer, Heidelberg (1995)
8. Lee, I., Phillips, P.: Urban crime analysis through areal categorized multivariate associations mining. Applied Artificial Intelligence 22(5), 483–499 (2008)
9. Shekhar, S., Huang, Y.: Discovering Spatial Co-location Patterns: A Summary of Results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 236–256. Springer, Heidelberg (2001)
10. Dray, S., Saïd, S., Débias, F.: Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. Journal of Vegetation Science 19, 45–56 (2008)
11. Wartenberg, D.: Multivariate spatial correlation: A method for exploratory geographical analysis. Geographical Analysis 17, 263–283 (1985)
12. Lee, S.: Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. Journal of Geographical Systems 3(4), 369–385 (2001)
13. Morrison, D.F.: Multivariate Statistical Methods, 2nd edn. McGraw-Hill, New York (1976)

14. Hubert, L.J., Golledge, R.G., Costanzo, C.M., Gale, N.: Measuring association between spatially defined variables: an alternative procedure. Geographical Analysis 17, 36–46 (1985)
15. Tiefelsdorf, M., Griffith, D.A., Boots, B.: A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A 31(1), 165–180 (1999)
16. Agrawal, R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the ACM SIGMOD 1993 International Conference on Management of Data, pp. 207–216. ACM Press, Washington (1993)
17. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2000)
18. Australian Institute of Criminology: Comparing International Trends in Recorded Violent Crime. In: Crime Facts Info No. 115 (2006), `http://www.aic.gov.au/publications/cfi/cfi115.html`
19. Murray, A.T., McGuffog, I., Western, J.S., Mullins, P.: Exploratory Spatial Data Analysis Techniques for Examining Urban Crime. British Journal of Criminology 41, 309–329 (2001)
20. Australian Bureau of Statistics: Australian Standard Geographical Classification (ASGC) (2005)
21. Phillips, P., Lee, I.: Areal Aggregated Crime Reasoning through Density Tracing. In: International Workshop on Spatial and Spatio-temporal Data Mining in conjunction with IEEE International Conference on Data Mining, Omaha, NE, USA (October 2007)