This file is part of the following reference:


Access to this file is available from:


The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owner of any third party copyright material included in this document. If you believe that this is not the case, please contact ResearchOnline@jcu.edu.au and quote [http://researchonline.jcu.edu.au/31323/](http://researchonline.jcu.edu.au/31323/)
Mining Cross-Patterning in Large Areal Aggregated Spatial Datasets

PETER PHILLIPS

This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

School of Business
Discipline of Information Technology
James Cook University
Australia

13 January 2011
Statement of Access

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Theses network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;

I do not wish to place any further restriction on access to this work.

Signature

27-02-2011

Date
Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Signature

27-02-2011

Date
Abstract

Intelligent crime analysis allows for a greater understanding of the dynamics of unlawful activities, providing possible answers to where, when and why certain crimes are likely to happen. With the growth of geo-referenced data and the sophistication and complexity of spatial databases, data mining and knowledge discovery techniques have become essential tools for the successful analysis of large spatial datasets. Crime analysis requires a combination of heterogeneous data, such as socio-economic and socio-demographic factors, geospatial features and crime datasets so that interesting patterns can be discovered.

The Queensland Police Service (QPS) and Australian Bureau of Statistics (ABS) record crime and census information in areal aggregated datasets. The primary reason for this is to protect the privacy of individuals. To enable intelligent crime analysis with these areal aggregated spatial datasets, the cross-patterning relationship between different spatial layers across locations needs to be modelled and quantified.

This thesis focuses on developing a framework for the discovery and visualisation of cross-patterning in areal aggregated spatial datasets. We show that cross-patterning can be modelled in three ways. Cross-association patterns that model the relationship between multiple datasets for each region while ignoring the effect of neighbouring regions, cross-varying patterns that take into consideration the effect of all local neighbouring regions and cross-distribution patterns that consider one local neighbouring region and the global distribution of the dataset. With areal aggregated datasets, local neighbours are defined as those that share a boundary.

The most suited type of cross-patterning is dependent on the application and dataset. We present a generic framework that can discover all types of cross-patterning from areal aggregated spatial datasets. The user can either select a specific type of pattern they wish to discover, or can use the framework to discover all cross-patterning and then use the visualisation environment to highlight patterns of interest.
We develop algorithms to discover each type of cross-patterning. To model cross-association we propose an Association Rules Mining (ARM) based approach. To model cross-varying we extend Lee’s bivariate spatial correlation approach [51], and to model cross-distribution we propose two new techniques based on the spatial distribution of the dataset. The framework also includes a visualisation environment that easily allows the user to interpret discovered patterns and highlight patterns that show overlapping cross-patterning (for example both cross-association and cross-distribution).

The algorithms developed for discovering cross-patterning relationships are designed so they can be applied to areal aggregated spatial datasets from a wide variety of disciplines. We test the framework with synthetic datasets to verify their correctness and then present a case study using real crime data from Brisbane, Australia. The computational performance of the algorithms is analysed and is acceptable for large datasets.
Acknowledgements

The work conducted in this thesis is due to the support and encouragement of many staff members, friends and family, and I am grateful to all those who have helped me complete my research.

First and foremost, I would like to thank my supervisor, Ickjai Lee, for providing guidance and support throughout this research project. His knowledge, assistance and feedback throughout my research has been invaluable and very much appreciated.

I would also like to thank the staff and students of the Discipline of IT for providing an enjoyable and educational research environment. They gave me valuable feedback that contributed towards the quality of my research. A special thanks also to Bruce Litow who helped to expand my knowledge of other topic areas relevant to my research.

For proofreading and providing comments and feedback, I would like to thank Elizabeth Wu. Her comments have helped me to refine and enhance the presentation of my work.

Finally, I would like to thank my parents for supporting me throughout my research.
Publications

International Journals with full paper refereed


International conferences and workshops with full paper refereed


Glossary

The following key concepts and terms are used in this thesis, and have been listed here for ease of reference:

- **Areal Aggregated Data**: A type of dataset that consists of combined (aggregated) values of all observations in a given region, for all regions in the dataset.
- **Autocorrelation**: Loosely defined as how a single variable correlates to itself between pairs of observations.
- **Cross-Patterning**: Describes patterns that happen together in some geospatial region across multiple layers of areal aggregated spatial data.
- **Correlation**: An association or relationship between variables.
- **Crime**: An act or conduct prejudicial to the community, in violation of and punishable by the law.
- **Crime Analysis**: The process of analysing crime data and identifying patterns.
- **Enviromental Criminology**: A branch of criminological theory that aims to understand the various aspects of a criminal event in order to identify patterns of behaviour and environmental factors that create opportunities for crime.
- **Geographic Knowledge Discovery (GKD)**: The process of extracting information and knowledge from large geo-referenced databases.
- **Geospatial Data**: Data containing observations that are located over a geographical space. See Spatial Data.
- **Moran’s I Statistic**: A statistical measure used to quantify spatial autocorrelation.
- **Repeat Victimisation**: The occurrence where a criminal will offend against the same victim repeated times.
- **Spatial Association**: The degree to which a set of univariate observations are similarly arranged over space.
- **Spatial Autocorrelation**: Correlation of a variable with itself through space.
• **Spatial Cluster:** A group of spatial objects whose properties are similar to other objects in the same group and dissimilar to the properties of objects in different groups.

• **Spatial Data:** Data containing observations that are located over a geographical space.

• **Spatial Data Mining:** A variation of data mining to enable the mining of patterns from spatial data.

• **Spatial Neighbourhood:** The area that lies in close proximity to a given object in a spatial dataset, defined by a topological, distance or direction relationship.

• **Tobler’s first law of geography:** ‘Everything is related to everything else, but near things are more related than distant things’.
CONTENTS

Statement of Access ii

Declaration iii

Abstract iv

Acknowledgements vi

Publications vii

Glossary ix

List of Figures xiv

Chapter 1 Introduction 1

1.1 Motivation ................................................................ 4
1.2 Contributions ............................................................. 7
1.3 Thesis Structure ............................................................ 9

Chapter 2 Related Work 11

2.1 Overview ............................................................... 12
2.1.1 Geographic Information Systems ................................. 13
2.1.2 Spatial Data ........................................................ 14
2.2 Spatial Data Mining and Geographic Knowledge Discovery ............. 16
2.2.1 Crime Data Mining ................................................... 18
2.3 Spatial Pattern Mining Techniques ....................................... 19
2.3.1 Hot Spot Analysis ................................................... 20
2.3.2 Spatial Association Rules Mining .................................. 24
2.3.3 Co-location Rules Mining ............................................. 25
2.3.4 Summary .......................................................... 27
2.4 Data Visualisation ................................................................. 28
2.5 Summary ........................................................................... 34

Chapter 3 Datasets 36
3.1 Crime Data from the Queensland Police Service ......................... 37
3.2 Census Data from the Australian Bureau of Statistics .................... 40
3.3 Summary ........................................................................... 41

Chapter 4 Cross-Association Patterns 43
4.1 Areal Categorised Multivariate Association Mining ....................... 44
  4.1.1 Data Preprocessing and Aggregation ................................... 44
  4.1.2 Categorisation and Classification ...................................... 47
  4.1.3 ACMAM ..................................................................... 48
4.2 Summary ........................................................................... 52

Chapter 5 Cross-Varying Patterns 55
5.1 Spatial Association .................................................................. 56
5.2 Problem Formulation and Statement ........................................ 58
  5.2.1 Removing Clusters ......................................................... 59
  5.2.2 Moving Clusters ............................................................. 62
  5.2.3 Randomly Distributed Clusters ........................................ 63
5.3 Irregular Areal Datasets .......................................................... 64
5.4 Summary ........................................................................... 68

Chapter 6 Cross-Distribution Patterns 70
6.1 Density Tracing .................................................................... 71
  6.1.1 Working Principle .......................................................... 72
  6.1.2 Algorithm ..................................................................... 74
  6.1.3 Time Complexity Analysis .............................................. 87
  6.1.4 Optimal Spatially Aware Ordering .................................... 87
  6.1.5 Validation with Synthetic Datasets ................................... 90
6.2 Graph Mining ...................................................................... 95
  6.2.1 Graph Based Representation .......................................... 95
CONTENTS

6.2.2 Working Principle ..................................................... 96
6.2.3 Algorithm ...........................................................102
6.2.4 Validation with Synthetic Datasets ...............................106
6.3 Summary ......................................................................110

Chapter 7 Cross-Patterning Visualisation 112
7.1 Single Result Set Visualisation ........................................113
7.2 Cross-Patterning Visualisation .........................................117
7.3 Summary ......................................................................120

Chapter 8 Case Study: Crime Data Analysis 124
8.1 Cross-Patterning Relationships ........................................124
  8.1.1 Cross-Association Patterns .........................................124
  8.1.2 Cross-Varying Patterns ..............................................132
  8.1.3 Cross-Distribution Patterns .........................................136
  8.1.4 Cross-Patterning Visualisation ......................................142
8.2 Computational Comparison ............................................145
8.3 Summary ......................................................................147

Chapter 9 Discussion 149
9.1 Summary of Contributions ..............................................149
9.2 Future Work .............................................................152

References 154
## List of Figures

1.1 Example cross-patterning: (a) sexual assault; (b) university locations; (c) overseas visitor population. .............................................................. 2  
1.2 Activities performed for spatial analysis of areal aggregated crime datasets. .... 7  
1.3 Cross-Patterning discovery techniques. .................................................. 8  

2.1 Example of Layers in a GIS ............................................................... 15  
2.2 Geographic Knowledge Discovery ..................................................... 17  
2.3 Snow’s Cholera Map (Broad Street Pump circled, with black areas indicating Cholera victims) ................................................................. 21  
2.4 Common classification techniques provided by ArcGIS: (a) Raw data; (b) Equal area classification; (c) Equal interval classification; (d) Natural breaks classification; (e) Quantile classification; (f) Standard deviation classification. 23  
2.5 Example Co-locations .................................................................. 26  
2.6 An example of a traditional pin map.................................................. 29  
2.7 An example of a modern pin map of crime in an area of Houston, Texas. ... 30  
2.8 An example of a modern pin map of fires in Victoria, Australia. ............ 31  
2.9 COPLINK display of criminal activity. ............................................. 32  
2.10 COPLINK Spatio-Temporal Visualizer showing Bank Robberies (September-December, 2001). ............................................................... 33  

3.1 Study region: (a) Map of Australia; (b) Map of the greater Brisbane area. ... 37  
3.2 Study region: Urban suburbs of Brisbane, Australia. ............................ 38  
3.3 A crime taxonomy tree illustrating various crime types in Queensland. ... 40  
4.1 The framework of areal categorised geospatial knowledge discovery..... 45
4.2 Area-to-area aggregation: (a) A base map; (b) A target map labelled with corresponding percentages of offences against person over total offences; (c) An overlay of the base map and the target map. 47

4.3 Categorisation with natural breaks: (a) A base map with national parks; (b) The base map with histograms; (c) A choropleth map by natural breaks with 5 classes. 48

4.4 Synthetic datasets for ACMAM: (a-c) Datasets dataset4.1a–c; (d) Base region. 49

5.1 Example cross-varying: (a) Murder (left) and Reserves (right); (b) Reserves (left) and Hospitals (right). 55

5.2 Drawback of Pearson’s correlation coefficient with spatial data: (a-c) Datasets dataset5.1a–c. 57

5.3 Synthetic hexagon dataset I: (a-d) Datasets dataset5.2D–D3. 59

5.4 Comparison of Lee’s L and Wartenberg’s I with synthetic hexagon dataset I (dataset5.2D–D3). 60

5.5 Synthetic hexagon dataset II: (a) Dataset dataset5.3a; (b) Dataset dataset5.3t. 61

5.6 Synthetic hexagon dataset III: (a-d) Datasets dataset5.4E–E3. 62

5.7 Comparison of Lee’s L and Wartenberg’s I with synthetic hexagon dataset III (dataset5.4). 63

5.8 Synthetic hexagon dataset IV: (a-b) Datasets dataset5.5a–b. 64

5.9 Irregular region dataset: (a-c) Datasets dataset5.6a–c; (d) Region outline. 65

5.10 Spatial lag of irregular dataset datasets5.6: (a-b) Unmodified weight matrix; (c-d) Modified weight matrix. 66

5.11 Synthetic dataset removing clusters: (a-d) Datasets dataset5.7a–d. 67

6.1 Example cross-distribution: (a) Rape and Railway Stations; (b) Fraud and College. 71

6.2 Working principle of our framework: (a-c) Datasets dataset6.1a–c before normalisation; (d) GLS region ordering; (e) Density trace; (f) Density trace with area highlighted. 73
6.3 Spatially aware region ordering: (a) Base map including centre points; (b) GLS; (c) DFS; (d) BFS; (e) NN-5; (f) Variance of edge lengths. ......................... 83
6.4 Density Trace: (a) Two datasets; (b) Subset of all datasets. ......................... 85
6.5 Region Weighted Locality In-between Polylines. ......................................... 85
6.6 A subset of regions from the Brisbane base map with synthetic data: (a-d) Datasets dataset6.2a−d. .................................................. 88
6.7 Optimal ordering: (a) GLS; (b) DFS; (c) BFS; (d) NN-5; (e) Optimal order with same starting region. ........................................ 89
6.8 Comparison of normalised density dissimilarity of synthetic datasets using various orders. ......................................................... 89
6.9 Synthetic experiment removing clusters: (a-d) Datasets dataset6.3a−d; (e) Increasing dissimilarity with respect to number of removed clusters. ............... 91
6.10 Synthetic experiment removing clusters: (a) Initial dataset dataset6.4a; (b) Increasing dissimilarity with respect to number of removed clusters. ................. 92
6.11 Synthetic experiment with moving clusters: (a-e) Datasets dataset6.5a−e; (f) Increasing dissimilarity with respect to number of moved clusters. ................. 93
6.12 Synthetic experiment with moving clusters: (a) Initial dataset dataset6.6a; (b) Final dataset dataset6.6u; (c) Increasing dissimilarity with respect to number of moved clusters. ........................................ 94
6.13 Generated graph of base map. ................................................................. 96
6.14 Synthetic experiment with moving clusters: (a-e) Datasets dataset6.7a−e .......... 97
6.15 MSTs of synthetic dataset: (a) MST of Datasets dataset6.7a−e from Fig. 6.14. 99
6.16 Correlations from synthetic dataset dataset6.7 using minimum edge weight and Pearson’s r ................................................................. 100
6.17 Correlations from synthetic dataset dataset6.7 using MST and Pearson’s r . . . 100
6.18 Correlations from synthetic dataset dataset6.7 using MST and Jaccard’s J . . . 101
6.19 Synthetic experiment with moving clusters: (a) Initial dataset dataset6.8a; (b) Final dataset dataset6.8u; (c) Correlations using MST. ............................. 101
6.20 Synthetic dataset removing clusters: (a-d) Datasets dataset6.9a−d. ................. 107
6.21 Correlations from synthetic dataset dataset 6.9 using minimum edge weight and Pearson’s \( r \) ................................................................. 107
6.22 Correlations from synthetic dataset dataset 6.9 using MST and Pearson’s \( r \) ................................................................. 108
6.23 Correlations from synthetic dataset dataset 6.9 using MST and Jaccard’s \( J \) ................................................................. 108
6.24 Synthetic experiment removing clusters: (a) Initial dataset dataset 6.10; (b) Increasing dissimilarity with respect to number of removed clusters. ................................................................. 109

7.1 Visualisation of synthetic result set 1 - random layout. ................................................................. 114
7.2 Visualisation of synthetic result set 1 - circular layout. ................................................................. 114
7.3 Visualisation of synthetic result set 1 - 2D spring layout. ................................................................. 115
7.4 Visualisation of synthetic result set 1 - 3D spring layout. ................................................................. 116
7.5 Visualisation of synthetic result set 2 - random layout. ................................................................. 117
7.6 Visualisation of synthetic result set 2 - 2D spring layout. ................................................................. 118
7.7 Visualisation of multiple result sets - random layout. ................................................................. 118
7.8 Visualisation of multiple result sets - circular layout. ................................................................. 119
7.9 Visualisation of multiple result sets - 2D spring layout. ................................................................. 119
7.10 Visualisation of multiple result sets - 3D spring layout. ................................................................. 120
7.11 Visualisation of synthetic result set 3 - circular layout. ................................................................. 122
7.12 Visualisation of synthetic result set 3 - 2D spring layout. ................................................................. 123

8.1 Boolean ACMAM with support and confidence values. ................................................................. 126
8.2 Boolean ACMAM with support and lift values. .............................................................................. 126
8.3 Visualisation of Boolean ACMAM - Random layout. ................................................................. 127
8.4 Visualisation of Boolean ACMAM - Circle layout. ................................................................. 127
8.5 Visualisation of Boolean ACMAM - 2D spring layout. ................................................................. 128
8.6 Quantitative ACMAM with support and confidence. ................................................................. 129
8.7 Quantitative ACMAM with support and lift. .............................................................................. 130
8.8 Visualisation of Quantitative ACMAM - Random layout. ................................................................. 130
8.9 Visualisation of Quantitative ACMAM - Circle layout. ................................................................. 131
8.10 Visualisation of Quantitative ACMAM - 2D spring layout. ................................................................. 132
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.11</td>
<td>Number of generated frequent itemsets and rules using boolean ACMAM</td>
<td>133</td>
</tr>
<tr>
<td>8.12</td>
<td>Number of generated frequent itemsets and rules using quantitative ACMAM</td>
<td>133</td>
</tr>
<tr>
<td>8.13</td>
<td>Visualisation of top-50 cross-varying patterns</td>
<td>134</td>
</tr>
<tr>
<td>8.14</td>
<td>Visualisation of bottom-50 cross-varying patterns</td>
<td>135</td>
</tr>
<tr>
<td>8.15</td>
<td>Visualisation of selected top-(k) crime cross-varying patterns</td>
<td>135</td>
</tr>
<tr>
<td>8.16</td>
<td>Visualisation of Density Tracing top-50 cross-distribution patterns</td>
<td>136</td>
</tr>
<tr>
<td>8.17</td>
<td>Visualisation of Density Tracing top-50 cross-distribution patterns</td>
<td>137</td>
</tr>
<tr>
<td>8.18</td>
<td>Visualisation of Density Tracing selected cross-distribution patterns</td>
<td>138</td>
</tr>
<tr>
<td>8.19</td>
<td>Lowering minimum similarity to reduce cross-distribution patterns</td>
<td>138</td>
</tr>
<tr>
<td>8.20</td>
<td>Comparison between GLS, DFS and BFS for kidnapping</td>
<td>139</td>
</tr>
<tr>
<td>8.21</td>
<td>Comparison between weighted and non-weighted regions for reserves</td>
<td>139</td>
</tr>
<tr>
<td>8.22</td>
<td>Visualisation of Graph Mining top-50 cross-distribution patterns</td>
<td>141</td>
</tr>
<tr>
<td>8.23</td>
<td>Visualisation of Graph Mining bottom-50 cross-distribution patterns</td>
<td>141</td>
</tr>
<tr>
<td>8.24</td>
<td>Visualisation of all cross-patterning - Random layout</td>
<td>143</td>
</tr>
<tr>
<td>8.25</td>
<td>Visualisation of all cross-patterning - 2D spring layout</td>
<td>143</td>
</tr>
<tr>
<td>8.26</td>
<td>Visualisation of all cross-patterning - 3D spring layout</td>
<td>144</td>
</tr>
<tr>
<td>8.27</td>
<td>Visualisation of overlapping cross-patterning - Circle layout</td>
<td>144</td>
</tr>
<tr>
<td>8.28</td>
<td>Visualisation of overlapping cross-patterning - 2D spring layout</td>
<td>145</td>
</tr>
<tr>
<td>8.29</td>
<td>Computational comparison of approaches for discovering cross-patterning</td>
<td>146</td>
</tr>
<tr>
<td>8.30</td>
<td>Computational comparison between Density Tracing and Graph based approach</td>
<td>146</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Intelligent crime data analysis facilitates an improved understanding of the dynamics of criminal activities such as robbery or assault. It aims to discover patterns of criminal behaviour that may assist in the discovery of where, when and why particular crimes are likely to occur.

Criminal behaviour is dependent upon situational factors and as such is patterned according to the location of criminogenic environments. Crime will be concentrated around crime opportunities and other environmental features that facilitate criminal activity. These environmental factors may include socio-economic, socio-demographic or geospatial features.

Criminals are creatures of habit and tend to have repetitive patterns of behaviour [94]. For example, some studies suggest that one third of convicted burglars admit to returning to the same residence more than once [35]. Law enforcement agencies can be more effective if they have information about perpetrators’ habits and areas they choose to commit crimes. To fully take advantage of this characteristic, crime data analysis techniques must capture connections between places and events based on past crimes.

Decreases in storage costs and ease of spatial data collection has meant that the amount of features available for analysis has grown substantially over recent years. Each of these features are represented as a layer, defined as a spatial dataset consisting of one thematic topic. For example one layer may contain suburb aggregates for murder while another layer contains aggregates for arson. To enable crime data analysis, data mining and knowledge discovery techniques have become essential tools to capture connections between many hundreds of these spatial layers.
The Queensland Police Service (QPS) and Australian Bureau of Statistics (ABS) record crime and census information in areal aggregated datasets due to limited environmental circumstances and ethical issues such as concern for the privacy of individuals. Areal aggregated spatial datasets are region-based datasets that have aggregate data values for regions, e.g., a particular suburb may have recorded five assaults. Regions that share a boundary are defined as neighbours. To enable intelligent crime analysis with these areal aggregated datasets, the relationship between many different spatial layers across locations needs to be modelled and quantified.

Cross-patterning describes patterns that occur together in some geospatial region across multiple layers. For example, Fig. 1.1 shows three areal aggregated spatial datasets with colour-coded aggregate data values (white=1, blue=2, red=3). An example cross-patterning relationship may be that incidents of sexual assault in region X and neighbouring regions Y and Z show cross-patterning with university locations in region X, Z, and an overseas visitor population in region Y, Z.
This thesis focuses on developing a framework for the discovery and visualisation of cross-patterning in areal aggregated spatial datasets. We identify three ways cross-patterning can be modelled; cross-association, cross-varying and cross-distribution.

Cross-association patterns ignore the effect of neighbouring regions. For example it would model the relationship between \textit{\{robbery in region X, schools in region X\}}.

Cross-varying patterns take into consideration the effect of all neighbours in the local region (i.e. all neighbours that share a border with the region). For example it would model the relationship between \textit{\{robbery in region X, robbery in neighbouring regions of X, schools in region X, schools in neighbouring regions of X\}}.

Cross-distribution patterns consider one neighbouring region and the global distribution of the dataset. For example it would model the relationship between \textit{\{robbery in region X, robbery in one neighbouring region of X, global distribution of robbery, schools in region X, schools in one neighbouring region of X, global distribution of schools\}}.

We consider three ways to model cross-patterning so as to provide a generic framework that can be used within many disciplines and applications. Tobler’s first law of geography states that ‘everything is related to everything else, but near things are more related than distant things’ [84]. When analysing areal aggregated spatial datasets, this means that neighbours (those that share a boundary) should be considered. This is achieved in our framework by discovering cross-varying patterns (all local neighbours) and cross-distribution patterns (one local neighbour and the global spatial distribution). In some applications, for computational efficiency or when performing non-spatial metadata analysis, neighbouring information may be ignored. In our framework this is achieved by cross-association patterns.

In many applications, analysts can benefit from a combination of the three types of cross-patterning. Our framework is able to discover each type and then using a visualisation environment we are able to highlight overlapping patterns. These overlapping patterns indicate that more than one type of cross-patterning has been discovered and may be of particular interest to the user.
The core objective of this thesis is the development of a framework for the discovery and visualisation of cross-patterning in areal aggregated spatial datasets. We identify and define three ways cross-patterning can be modelled to discover interesting patterns across spatial layers. We also develop a visualisation environment that enables users to easily interpret the resulting cross-patterning.

The algorithms we develop for discovering cross-patterning relationships are designed so they can be applied to areal aggregated spatial datasets from a wide variety of disciplines. For example, the framework presented in this thesis can be applied to spatial datasets in disaster management, epidemiology, business intelligence, geology, environmental monitoring, marketing and e-commerce. One of the motivating factors of this thesis is to improve analysis of crime datasets to allow reasoning towards better decision making and as such we provide a case study that shows how these algorithms are suited for such a purpose.

The following section highlights the motivation behind this thesis while Section 1.2 provides an overview of our contributions. Section 1.3 outlines the content of this thesis.

1.1 Motivation

The focus of this thesis is on the development of a framework for the discovery and visualisation of interesting patterns across many spatial layers of areal aggregated data. Our specific application of these tools is to assist in providing a deeper understanding of criminal events and behaviour in areal aggregated crime datasets. Therefore it is necessary that we first describe why crime data analysis is a challenging, interesting and important area of study. We then describe how advances in computer storage and spatial data collection pose challenges to crime data analysis that must be overcome.

Intelligent crime data analysis is of great importance to a number of people and agencies such as regional planners, politicians, police and residents. Regional planners can use information about environmental features that attract crime to design safer living environments (for example, with good lighting). Politicians can develop policies and laws to help deter criminals and
reduce the incidence of crime. Police can use knowledge about criminal patterns to develop preventative crime strategies, and to help catch offenders. While residents can use this information to try to avoid falling victim to crimes.

The costs of crime go far beyond economic, which alone in Australia was estimated at nearly AU$36 billion a year in 2005, to include social, political and environmental factors [74]. Therefore, improving our understanding of crime is of benefit to all members of society.

The distribution of crime in time and space is non-random. Because criminal behaviour is dependent upon situational factors, crime is patterned according to the location of criminogenic environments. Crime will be concentrated around crime opportunities and other environmental features that facilitate criminal activity. For example, the top three locations for armed robbery in Australia during 2006 in decreasing order were on the street and footpath, from a retail store and from a service station. This is largely due to their perception as easy targets [50]. Also, in Queensland, the highest residential burglary rates were in coastal tourist regions, inner city areas of Brisbane and a band of suburbs south of Brisbane between Ipswich and Beenleigh [24]. The purpose of crime analysis is to identify and describe crime patterns such as these [94].

Environmental criminology is a branch of criminological theory that can guide crime analysis and crime prevention efforts. The goal of environmental criminology is to understand the various aspects of a criminal event in order to identify patterns of behaviour and environmental factors that create opportunities for crime [13]. Discovering crime and spatial features that exhibit cross-patterning relationships is a key component to environmental criminology and provides insight into the complex nature of criminal behaviour. These resulting patterns may be possible causal factors that warrant further investigation. For example, determining that criminal offences involving receiving stolen property show a cross-patterning relationship to improvised homes and parks could lead to a targeted crime prevention effort by police in those specific areas. Improvised homes and parks may not be the causal factors of receiving stolen property, but their cross-patterning makes the pattern valuable to police and policy makers.
Crime, socio-demographic information and environmental feature datasets are maintained by different government and civilian agencies. As with traditional databases, the decrease in storage costs and ease of data collection has led to a data rich environment. This growth in data has increased the need to make sense of it. Data mining has therefore gathered significant attention, as it has the capability of providing a better understanding of the information contained within the data.

The aim of data mining is to find previously unknown, nontrivial information from data. This process is automated using algorithms that enable the efficient discovery of such information from the data. Crime has a spatial component (where the crime took place) and as such techniques for crime analysis should consider this spatial information. In addition to the location information contained in such datasets, they also contain textual and/or numeric attributes to describe the feature. For example, the street address of a school is recorded along with information such as the number of pupils, teachers and buildings.

Due to the greater complexity of spatial data compared to traditional data, previous methods for data mining are generally not suitable for application directly on spatial data [78]. As a result, new data mining techniques to discover information from spatial data have been developed. However, many spatial data mining techniques, such as predictive modelling, spatial clustering and spatial outlier detection, typically focus on a single spatial layer. To enable crime data analysis, techniques that are able to capture connections and relationships between many hundreds of spatial layers from a wide variety of sources are needed.

In addition, a framework is needed that can use heterogeneous data types (such as areal aggregated, point and line datasets) from a wide variety of sources (such as police agencies, census agencies, local councils), efficiently mine these datasets for cross-patterning relationships and then present information to the user in a usable fashion. Obtaining useful information from these crime datasets is a non-trivial task due to the high number of datasets to be considered and the high number of possible cross-patterning relationships within these datasets.
1.2 Contributions

There are four major outcomes that this thesis achieves. They are to:

- Identify and define cross-patterning relationships across multiple areal aggregated spatial datasets;
- Develop new techniques to discover and model these cross-patterning relationships;
- Develop a visualisation environment to enable users to easily interpret the discovered relationships; and,
- Use the developed tools for the analysis of crime datasets along with socio-economic, socio-demographic and geospatial features.

As with all data mining tasks, the main goal is to provide an improved understanding of the data in the form of information. Figure 1.2 provides an overview of this; many layers of raw data are processed to identify and model cross-patterning relationships, which are then visualised for easy interpretation by domain experts.
spatial data are available on the left, analysis is performed in the middle section to discover unknown patterns, and then these patterns are presented to the end user in an easy to understand format. This thesis is specifically focused on discovering relationships between different spatial layers across locations. The first part of this work identifies and defines these relationships as cross-patterning relationships. Cross-patterning describes patterns that happen together in some geospatial region across multiple layers.

We identify three ways in which cross-patterning can be modelled. These are; cross-association patterns that ignore the effect of neighbouring regions, cross-varying patterns that take into consideration the effect of all local neighbouring regions and cross-distribution patterns that consider one local neighbouring region and the global distribution of the dataset. The second part of this work is to develop techniques to model and quantify each of these types of patterns, as depicted in Figure 1.3.

![Figure 1.3: Cross-Patterning discovery techniques.](image)
Cross-association captures the relationship between different spatial layers for each region to allow for the extraction of frequent patterns. We propose an ARM based approach for automating the detection of multivariate cross-associations based on a given areal base map. Cross-varying models patterns using a combination of point-to-point association and spatial dependence. We extend Lee’s bivariate spatial correlation approach [51] to consider areal aggregated datasets containing irregular regions (such as suburbs). Cross-distribution models the spatial distribution of datasets to allow for the discovery of datasets that show similar distribution in similar spatial neighbourhoods. We develop two techniques to discover cross-distribution patterns. Density Tracing incorporates both localised clusters (through the use of context sensitive weighting and clustering) and the global distribution trend. We also investigate an efficient graph based dataset representation that is able to mine a set of datasets for cross-distribution patterns.

The visualisation environment developed in the third part of this work serves two purposes. The main purpose is to allow the end user to easily access and interpret cross-patterning information extracted from the datasets. Secondly, it allows for the combination of results from the cross-association, cross-varying and cross-distribution algorithms, in order to identify those patterns that overlap with each other. These overlapping patterns may be of particular interest and are highlighted to the user.

The algorithms developed for discovering cross-patterning relationships are designed so they can be applied to areal aggregated spatial datasets from a wide variety of disciplines. In this thesis we focus on crime analysis and as such we provide a case study that shows how these algorithms are suited for such a purpose.

### 1.3 Thesis Structure

Following this introduction, we first provide the necessary background information and related work that is required in order to understand the work performed in this thesis. The background begins with an overview of Geographic Information Systems (GIS) and spatial data. Following this, information about geographic knowledge discovery and crime data mining is presented to
highlight the need for new algorithms to discover cross-patterning in areal aggregated spatial datasets.

Chapter 4 describes the cross-association algorithm, while Chapter 5 details the cross correlation algorithm to model cross-varying. The two algorithms developed to model cross-distribution, Density Tracing and Graph Mining are described in Chapter 6.

We describe the visualisation environment and show how the results from cross-association, cross-varying and cross-distribution can be combined in Chapter 7.

Chapter 8 presents a case study detailing how cross-patterning can be used for intelligent crime data analysis. We examine each algorithm in detail to analyse crime data from the QPS along with socio-economic, socio-demographic and geospatial features. In this chapter each algorithm’s performance is also compared.

Chapter 9 concludes the findings of our research and summarises potential future work that could take place following on from this work.
CHAPTER 2

Related Work

This chapter provides a review of the relevant literature that describes the current landscape of spatial data analysis and crime data analysis.

Crime analysis is only possible where data about crimes is available. Therefore the first component to consider is how spatial data is stored and analysed currently using GIS, and the benefits and limitations of such systems for analysis. While these systems are usually able to answer basic locational questions, they are typically unable to provide answers to questions relating to patterns of behaviour. Therefore other techniques need to be investigated to discover such patterns.

In developing these techniques it is important to understand that due to the unique features of spatial data, traditional data mining techniques are not always suitable for mining patterns from such data. This has led to the development of new spatial data mining techniques which form part of the process of Geographic Knowledge Discovery (GKD). Specific work has also been conducted in the area of crime data mining which is a subset of spatial data mining.

The data we focus on in this thesis is areal aggregated. There are several existing techniques that can be adapted to discover patterns from areal aggregated spatial data. These include hot spot analysis, Spatial Association Rules Mining (SARM) and Co-Location Rules Mining (CLRM). We describe these techniques and provide examples to illustrate their usage. Techniques that are able to discover patterns across multiple layers of areal aggregated data are less well studied and are the focus of this thesis.

Once relationships within the datasets have been found, the resulting patterns need to be investigated by an analyst for useful information. Due to the size and complexity of spatial data, many
of these techniques discover a large number of complex patterns. Therefore the next challenge is to provide the most important, useful and relevant information to the analysts. Visualisation is a clear and effective technique for communicating these results.

This chapter outlines related work that has contributed and led to the development of the techniques introduced in this thesis, which enable crime analysts to gain greater knowledge and understanding about criminal patterns of behaviour.

2.1 Overview

This section introduces the GIS which is the primary storage location for most geographic and spatial data. While GISs have been around since the early 1960s [71], they were not widely available until the early 1980s. This was helped along by the increasing adoption of Personal Computers such as that by IBM in 1982. Improvements in technology eventually lead to a number of GIS vendors producing commercial GIS products, such as those released by the Environmental Systems Research Institute (ESRI) and MapInfo. Microsoft has also recently started to provide some spatial tools as part of it’s SQL Server 2008 suite at a more affordable price than some of the dedicated GIS vendors, allowing more users to store and use spatial data.

In more recent years, increasing amounts of spatial data are becoming available due to developments in both collection and storage technology. For example, improvements in the performance of telescopes have enabled the Sloan Digital Sky Survey (SDSS) [7] to collect data and map a large number of galaxies, quasars and stars. On a smaller scale, Google is encouraging organisations and individuals to provide and access spatial information through their Google Maps technology, which is widely available to the public [33].

Different types of spatial data are also increasingly being captured and in some cases being made publicly available either directly or through applications that are built from them [6, 29, 85]. Governments and organisations are coming to realise the benefits of spatial mapping where both types of enterprises can use spatial information for the provision of services and for planning.
As the size and types of spatial data is increasing, the ability to analyse and understand it is a growing challenge. Additionally, spatial data contains some complexities over traditional trans-actional data that must be taken into account when developing pattern discovery techniques.

In this section we introduce GISs, their benefits and limitations and then describe the complexities of the spatial data that they store. A more detailed description of the crime data used in this thesis is provided in Chapter 3.

2.1.1 Geographic Information Systems

Most spatial data is stored in a GIS. A GIS is defined as a tool for the manipulation of geographic data [11]. The GIS performs a great diversity of functions; some of which include compilation, verification, storage, retrieval, manipulation, update and visualisation of geographic data.

The adoption and storage of data into GISs has allowed users to find answers to spatially-related queries. Some such queries may be, ‘What are the names of all customers who live within 10km of my store?’, or ‘What are all the crimes that have occurred within a 5km radius of my service station?’. Another example may be a query from a mobile device user who may ask ‘Where is the nearest Automatic Teller Machine (ATM)?’. The latter example is already being performed through the use of a GIS by Mastercard [57].

However, while the advances in spatial GISs are significant and important, realising the full value of spatial data is not yet complete. Growth in the use of such systems has led to an increase in the storage of spatial data. As with traditional databases, the decrease in storage costs and ease of data collection has meant that gathering useful and interesting spatial information is a difficult but important task.

Most GISs, without the help from spatial data mining techniques, are not able to mine spatial data to find interesting and previously unknown information. Some examples of such information are ‘What services are frequently requested together from a mobile device?’ or ‘Around which spatial features is crime likely to occur near?’.
One of the reasons spatial information discovery is challenging is that GISs often store multiple layers of spatial data. Each layer represents a different feature set. Figure 2.1 illustrates some examples of layers that can be stored in a GIS. The layers in this example include census information along with locations of crimes, parks, schools and train stations. Being able to find relationships in the data between layers is important in understanding patterns of criminal behaviour. In this example, we may be able to find relationships between the locations of particular crimes and the locations of parks or schools.

Layers can also be at different granularities and consist of different spatial data types (lines, points, grids, regions) which can also make pattern discovery more challenging. Each layer can also hold additional information about the data points or regions captured. For example, crime data can indicate the type and time of a particular crime incident, and park data could include the size or activity level of a park.

Spatial data mining as described in the subsequent sections, is able to provide a solution to this problem. It can be used to discover patterns from multiple layers of spatial data, so that inferences can be drawn about relationships between the location of particular events and other features.

### 2.1.2 Spatial Data

Spatial data is data that contains thematic and geographic attributes. For example, the thematic attribute could be the type of crime committed (such as arson) or the land feature (such as a park). The geographic attribute is the location of the thematic attribute, that is, the location where the crime occurred, or the location of the park.

With an increasing amount of spatial data available, there has been a growing need to mine information about the types of patterns that are prevalent in the data. However, this is a difficult task as there are many features of spatial data that must be considered.

One of the most cited differences between spatial and non-spatial data is that observations are usually not independent and identically distributed (i.i.d.). Statistical independence
means that the occurrence of one event has no effect on the likelihood of the occurrence of another event. An example of how spatial data is not independent is that weather conditions in one region is influenced by weather conditions in neighbouring regions. Identical distribution means that each observation should have the same probability distribution as any other observation. However in the case of spatial data, the probability of certain values occurring is influenced by the values of neighbouring observations.

As stated in Chapter 1, Tobler’s first law of geography states that ‘everything is related to everything else, but near things are more related than distant things’ [84]. The term ‘spatial autocorrelation’ is also often used to describe this phenomenon. Spatial autocorrelation is
defined by Goodchild [31] as “the degree to which objects or activities at some place on the Earth’s surface are similar to objects or activities nearby”.

Another consideration when mining spatial data is that there are at least two types of information for any given spatial data point [31]. That is, their location, which may be specified using co-ordinates for example, and their attributes which are the variable values taken for that location, such as the number of crimes committed.

Such features mean that traditional approaches to mining transactional data are not easily applied to spatial data. In order to handle such data, spatial data mining techniques need to be developed. Some such techniques are highlighted in Section 2.2.

### 2.2 Spatial Data Mining and Geographic Knowledge Discovery

Geographic Data Mining (GDM) is a component of GKD. GKD is the process of extracting information and knowledge from massive geo-referenced databases [62]. GDM involves the use of computational techniques and tools to discover patterns that are distributed over geographic space and time [62]. It is also sometimes referred to as spatial data mining within the data mining community [47, 76].

The GKD process is shown in Figure 2.2. The first step is to extract the raw spatial data from a spatial data store. The extracted data may not be in the desired form for GDM, therefore some preprocessing may be required. As with data in traditional databases, this data may contain some inconsistencies, such as missing data that should be dealt with prior to the GDM process. The layers that are to be mined should also be specified and those layers extracted. As spatial data layers may be stored in different granularities, any aggregations should also be made to this data to simplify GDM where necessary.

Once the preprocessing is complete, GDM algorithms may be run to discover patterns from the data. Some examples of GDM techniques are outlined in Section 2.3. Once patterns are
retrieved, they are often returned in plain text format, or other formats that are difficult to interpret. Therefore a postprocessing step must take place to attempt to transform the data into more useful information. This step involves filtering out any uninteresting patterns and providing a graphical visualisation to the end user expert. The final stage is for the expert to investigate possible relationships further and potentially gain further knowledge about the discovered patterns.

Geoinformation and spatial data have unique characteristics that some data mining techniques do not take into consideration [28, 69]. Openshaw [69] defines GDM as ‘a special type of data mining that seeks to perform similar generic functions as conventional data mining tools, but modified to take into account the special features of geoinformation, the rather different styles and needs of analysis and modelling relevant to the world of GIS, and peculiar nature of geographical explanation.’

Therefore, in order to detect geographically interesting patterns, GDM techniques must consider the unique features of geographic data in order to find potentially interesting geographical patterns. The discovered patterns can be either descriptive or prescriptive. Descriptive pattern mining techniques summarise and characterise the data. Prescriptive pattern mining techniques attempt to predict patterns by drawing inferences from the data.

Geographic Data Mining, as a quantitative study of georeferenced phenomena, explicitly focuses on the geospatial features of the data such as location, distance, interaction and geospatial arrangement [5, 10].
2.2 Spatial Data Mining and Geographic Knowledge Discovery

2.2.1 Crime Data Mining

Crime data mining has gained significant attention among academics over the past decade as terrorist attacks have become more widespread and as data mining techniques have become more mature. Crime data mining builds on traditional data mining techniques and applies them to different types of crimes \([17, 18]\).

Most research in crime data mining has focused on the application side. Dedicated novel crime data mining methods have been relatively less studied. Although several efforts have been made to develop variants of dedicated crime data mining techniques, they are task-specific and still lack a unifying crime data mining framework \([17]\).

For the analysis of crime data, data mining techniques have been applied by the US Federal Bureau of Investigation (FBI) as a part of the investigation of the Oklahoma City bombing (Unabomber case), and many lower-profile crimes \([12]\). Data mining has also been used by the US Treasury Department to hunt for suspicious patterns in international funds transfer records; patterns that may indicate money laundering or fraud \([12]\). Examples also include the Health Insurance Commission of Australia using data mining to investigate fraud within the Medicare system \([42]\) and National Roads and Motorists’ Association (NRMA) Insurance Ltd \([91]\).

However, in the above examples, the techniques have been limited to the thematic attribute-oriented side of the data, while often ignoring the spatial and temporal attributes. The where and when are crucial for patterns of crime.

Recent progress towards developing GIS for crime analysis emphasises the relevance that georeference has to understand patterns in crime \([44]\). In a limited exploratory approach, systems have been developed to geographically visualise patterns of vehicle crime, domestic burglaries, drug-related crime and public disorder in inner city areas \([68]\), and violent crimes in Brisbane \([64, 65]\). The COPLINK project \([17, 40, 41]\) developed by the University of Arizona Artificial Intelligence Lab in collaboration with the Tucson Police Department is the most comprehensive crime data mining suite. It consists of several different modules conducting various
crime data mining functions including link analysis visualisation, incident analysis though clustering, visualisation through geospatial mappings, and geospatial/temporal analysis.

Although this suite provides a list of comprehensive data mining techniques, there are several limitations associated with it. Firstly, it mainly focuses on link analysis to find strong associations in organised crimes. Incident analysis is based on point data not areal aggregated data. Socio-demographic data is not properly considered in the geospatial domain, and the level of detail in data significantly differs from the level of detail in areal aggregated crime data available from the QPS.

2.3 Spatial Pattern Mining Techniques

The most common techniques used to extract patterns from spatial data include hot spot analysis, SARM and CLRM.

Hot spot analysis is the discovery of areas with particularly high levels of the occurrence of some spatial feature. A crime hot spot can be generally defined as ‘an area containing dense clusters of criminal incidents’ [3]. It is a descriptive mining tool that is able to summarise and characterise the data. Hot spot analysis can is also one way of performing deviation detection, and can also be used as a type of trend detection.

Spatial association rules and co-location rules are both predictive pattern mining tools that can draw inferences from the data for predictions. A spatial association rule is a rule of the form \( A \rightarrow B \), where \( A \) and/or \( B \) are a set of non-spatial and sometimes spatial predicates. For example, a strong spatial association rule may be that 90% of sexual assaults may occur in parks. Co-location rules are defined as non-spatial items that frequently occur together over space [79].

These techniques are discussed in further detail in the following subsections.
2.3.1 Hot Spot Analysis

Hot spot analysis is the discovery of areas with particularly high levels of the occurrence of some spatial feature. For example, there may be a particularly high rate of illness or crime in one particular area. One of the earliest and most commonly cited hot spot analysis was performed by John Snow in Soho, England in response to a Cholera outbreak during 1854 \[80\]. By plotting the location of cholera victims on a map of London city shown in Figure 2.3 he showed that many of the victims were situated around a particular water pump location on Broad Street. The discovery of this cholera ‘hot spot’ led authorities to remove the handle from the water pump.

The availability of spatial data and the introduction of storage systems such as GISs has enabled hot spot analysis to be performed on much larger sets of spatial data, and on a variety of different types of data.

In crime analysis, the use of hot spots to determine policing and crime prevention strategies has grown over recent years \[23, 44, 56, 64, 72\]. Crime hot spots have appealed to both crime prevention practitioners and police managers \[72\]. Current GISs such as ArcGIS provide a limited range of data classification methods for hot spot analysis such as natural breaks, quantiles, equal area, equal interval and standard deviations \[27\]. Figure 2.4 shows various classification methods available within ArcGIS for the crime dataset, Fraud by Cheque, from Brisbane, Australia.

Natural breaks (Figure 2.4(d)) divide the data by identifying natural groupings and ‘break points’ in the data that best group similarities within a class and maximise differences between classes. That is, boundaries are set where there are large jumps between neighbouring values. This method is best used when there is uneven distribution of values.

In quantile classification (Figure 2.4(e)), the number of values is the same in each class. This is achieved by dividing the range of possible values into unequal-sized intervals. Classes at the extremes have a wider interval range. Therefore this method is generally used to highlight variations in the middle-values of the distributions.
The equal interval technique (Figure 2.4(c)) is similar to quantile classification, except that interval ranges are equal for each class. As there are usually less values at the extremes, this technique is generally used to highlight changes in the extreme range of values.

Equal area classification (Figure 2.4(b)), also known as Geometrical Interval classification, was designed primarily for use with continuous data. Breaks between classes are based on class intervals that have a geometrical series. A geometric series is created by multiplying all the values in the series by a constant coefficient.
The standard deviation classification (Figure 2.4(f)) method shows how much an attribute value deviates from the mean. It places class breaks above and below the mean value at intervals of 1, 0.5, or 0.25 standard deviations.

The correct choice of classification method has a great impact on generated hot spots, and thus any reasoning based on these hot spots. Choosing the ‘best’ classification method is heavily dependent on the user’s domain knowledge and even then may require a number of iterations to determine the most suitable method. In data-rich environments, these iterations are time-consuming and expensive. Most importantly, the boundaries of the base map dominate shapes of crime hot spots and we are not able to find data-oriented natural shapes of crime hot spots with the current techniques available in GISs. Thus, subsequent reasoning about hot spots can be misleading.

Several approaches have been made to overcome these inherent limitations of GISs and to identify data-oriented shapes of crime hot spots in the geographical data mining community [23, 56, 64, 72]. These approaches identify base-map-free hot spots based on various clustering criteria. However, relatively little research has been conducted on postclustering and reasoning about crime hot spots for answers to questions such as “why are they there?”.

Estivill-Castro and Murray [64] present a partitioning clustering algorithm using a \( k \)-medoid type optimisation function. The algorithm uses hill-climbing heuristics to efficiently allow the exploratory analysis of large spatial datasets. The novel idea of stopping these heuristics early is analysed and shown to offer large computational savings while keeping the quality of the clustering high. They show that clustering in respect to a given feature set (such as schools and parks) allows positive associations to be discovered. The downside to their approach is that the clustering algorithm only produces single level clusters and uses \( k \)-partitioning clustering. It is known that criminal activities are inherently hierarchical and the number of clusters is unknown [52, 56]. Choosing the best value for \( k \) is labourious and non-trivial.

Lee and Estivill-Castro [53] present a framework for mining interesting associative patterns in complex crime datasets. They use clustering to identify hot spots, transform them into polygons, and then find possible associative relations with ARM. Their framework considers layers
of different spatial features and how they impact on each other. The performance of the framework is heavily dependent on the clustering technique and the boundary extraction technique used. The major drawback to this technique is that it uses single-level clustering which does not fully reveal the true structure of complex crime data.

A similar framework using hierarchical Short-Long clustering and Region Connection Calculus (RCC) is proposed by Lee and Williams [54]. They use cluster boundary extraction and the resulting polygons to detect relationships between multi-level clusters and to suggest their
positive associations. For massive crime datasets, the computational burden of computing and extracting cluster boundaries limits the usability of both of these techniques.

Although there exist a number of techniques available to conduct hot spot analysis, most of these techniques focus on point data. In this thesis we focus on discovering patterns from areal aggregated data.

### 2.3.2 Spatial Association Rules Mining

Traditional ARM [1, 2, 37, 43] has been a powerful tool for discovering positive associations among a set \( \mathcal{I} = \{I_1, I_2, \ldots, I_n\} \) of items in a transactional database \( \mathcal{D} \). Here, each transaction \( T \in \mathcal{D} \) is a subset of \( \mathcal{I} \). An association rule is an expression in the form of \( X \Rightarrow Y \ (c\%) \), \( X \in \mathcal{I}, Y \in \mathcal{I} \), and \( X \cap Y = \emptyset \) where \( X \) is called antecedent and \( Y \) is called consequent. It is interpreted as “\( c\% \) of transactions in \( \mathcal{D} \) that satisfy \( X \) also satisfy \( Y \)”.

Typically, support and confidence are two measures of a rule’s interestingness. Support ensuring statistical significance is the probability that \( X \) and \( Y \) exist in a transaction \( T \in \mathcal{D} \). Confidence indicating the rule’s strength is the probability that \( Y \) exists in a transaction \( T \) given that \( T \) contains \( X \). That is, support is an estimate for \( \text{Prob}(X \cup Y) \) and confidence is an estimate for \( \text{Prob}(Y|X) \). Two user defined thresholds, minimum support and minimum confidence, are used for pruning rules to find only interesting rules. Itemsets satisfying the required minimum support constraint are named frequent while rules satisfying the two thresholds are called strong.

Another measure of a rules interestingness is lift [90]. The lift value is given by the ratio of confidence to expected confidence, where the expected confidence is the number of transactions that include the consequent divided by the total number of transactions. Lift indicates the increase in probability of \( X \) given \( Y \).

Support, confidence and lift are used to filter out uninteresting association rules that are less probable, and thus potentially provide analysts with previously unknown but useful strong association rules that can assist with decision making.
Mennis [61] applied traditional ARM to socioeconomic and land cover change analysis. It ignores the spatial dimension and simply converts the spatial database to relational tables to which traditional ARM can be applied. Typically, ARM has been applied to geospatial data in two different ways. This is parallel to the fact that there are two popular views of geographical space in a GIS: vector and raster. The vector view regards geographical space as a set of objects while the raster view sees it as a set of locations. Spatial Association Rules Mining (SARM) [48] is similar to the raster view in the sense that it tessellates a study region $S$ into discrete groups based on spatial or aspatial predicates derived from concept hierarchies. For instance, a spatial predicate $\text{close_to}(\alpha, \beta)$ divides $S$ into two groups, locations close to $\beta$ and those not. So, $\text{close_to}(\alpha, \beta)$ can be either true or false depending on $\alpha$’s closeness to $\beta$. A spatial association rule is a rule that consists of a set of predicates in which at least one spatial predicate is involved. For instance, $\text{is_a}(\alpha, \text{house}) \land \text{close_to}(\alpha, \text{beach}) \Rightarrow \text{is_expensive}(\alpha)$.

Although this approach efficiently mines complex geospatial datasets, it has several drawbacks. It is not directly applicable to datasets in which transactions can be infinite or indefinite. That is, situations where the traditional interesting measures cannot be directly measured. Also, a great degree of human involvement is required to find association rules. In this approach, the predefined concept hierarchy drives the human-oriented rule discovery which does not comply with exploratory data analysis and the principle to “let the data speak for themselves” [34, 68].

### 2.3.3 Co-location Rules Mining

Co-Location Rule Mining [79] discovers a subset of features given a set of point features frequently located together in a geographic space. An example of this is shown in Figure 2.5, where green circles and red squares are frequently located together. These shapes could represent any two items in space. For example these points could indicate the co-location of public recreation areas and theft, chemical storage facilities and cancer or puma and deer.

It extends traditional ARM by providing a transaction free approach of mining co-location rules, using the concept of neighbourhoods without having to define a reference feature. This
avoids potential loss of proximity relationship information in partitioning continuous geographic space into transactions. Thus, this approach focuses on features, similar to the vector approach. Here, transactions correspond to point locations and items correspond to features. Since features have different occurrences, the size of transactions vary with features. Because of this it is impossible to use the traditional interestingness measures. To tackle this problem, Shekhar et al. [79] introduces two new interestingness measures (prevalence and conditional probability) that can be used in a dynamic situation where transactions are not fixed to a constant. Prevalence indicates the minimum participation ratio of a feature in either the antecedent and consequent of a rule. The conditional probability is the probability of finding the consequent in a neighbourhood of antecedent.

Huang et al. [45] improve upon the event centric model and present a generalised algorithm to discover co-location patterns from point spatial datasets. It includes a multi-resolution filter that increases the performance of the algorithm when the features of the dataset are naturally clustered. One of the major problems with the Co-location Miner is that it cannot handle extended spatial data types such as line segments, polygons and circles. Xiong et al. [95] propose the EXCOM algorithm which effectively solves this problem using the notion of zones.
around spatial objects called buffers. This algorithm integrates the best features of the event-centric model and applies a statistically consistent definition for the conditional probability measure. The authors suggest that the EXCOM algorithm could also be extended to mine co-incidence events, which extend the concept of co-location events.

Co-location rules mining is able to find interesting rules and does not require a centric feature to be examined for related items, as with SARM. However, in order to find objects with relationships, a distance measure must be defined. The value of this distance measure can have a significant impact on the rules that are found, and it may be difficult from a user perspective to know which rules are the most interesting and important.

### 2.3.4 Summary

The spatial data mining techniques discussed in the previous sections can provide interesting insights into patterns that occur in spatial datasets.

Hot spot analysis discovers high density regions for a particular feature, such as robbery. The limitation however is that it only considers one feature layer, and does not look for relationships between various features. In addition, earlier work focuses primarily on point-based data, while we focus on discovering patterns from areal aggregated data.

Spatial association rules mining finds ‘strong’ rules, where the presence of one feature is highly associated with the presence of other features. However it is not directly applicable where traditional interestingness measures can’t be measured directly, and it can require a lot of human involvement. These approaches typically focus on frequent patterns, and can thus be dominated by uninteresting frequent patterns, such as traffic lights are associated with roads.

Another limitation of SARM is that it is performed by searching for associations using a centric feature. Co-location rules mining overcome this by searching for all features that are located within a user-defined distance from each other. However setting different distances can change the rules that are found. These may not always be the ‘best’ rules, and it may be difficult to know which rules are more interesting unless they are investigated further.
While these techniques are able to discover some relationships in spatial data, they are not directly applicable to discovering relationships across many layers of areal aggregated spatial datasets such as those studied in this thesis. We propose new techniques to overcome these limitations and provide a method of discovering patterns across layers of areal aggregated spatial datasets.

## 2.4 Data Visualisation

Visualisation is ‘a computer-aided process that aims to reveal insights into an abstract phenomenon by transforming abstract data into visual spatial forms’ [16]. In order for data mining techniques to be effective, systems should be able to display discovered patterns in a variety of forms [37, p 157].

The use of visualisations can improve our ability to understand and interpret data that would not easily be understood in its plain textual or numeric form. Using a visualisation tool can make interpretation much easier and faster for users. It is also able to capture and convey complex relationships and information that would be difficult to convey in text or other forms. This is particularly important when visualising results from a data mining technique as often a large number of complex relationships are discovered.

Representing crime data visually is not a new phenomenon. Early crime analysis involved the use of physical maps of particular jurisdictions with pins stuck in them [39]. The usage of crime maps by the New York Police Department for example, has been traced back to as early as 1900. An example of a physical pin map is shown in Figure 2.6. Limitations of pin-mapping, including loss of historical pin locations, difficulties with displaying different types of crime, and the large amount of physical space required for such maps eventually led to the development of ‘virtual’ computer assisted map representations. In more recent years, this has been vastly facilitated by increased computing power to aid such visualisations [17]. An example of a modern virtual map made available by The City of Houston [20] is shown in Figure 2.7. There are several advantages of virtual maps over pin maps, including:
2.4 Data Visualisation

- A much larger number of points can be plotted and analysts can zoom in or out on the image;
- Search facilities can be made available;
- Plotting different types of items with a variety of features is possible;
- A large amount of metadata can potentially be stored about each point;
- Historical values can be stored, and the visualisation can be made to select different time periods under analysis;
- The spatial data can be backed-up so it will not be lost in the case of a disaster; and
- Data is easy to share between analysts and any other individuals with access.

While pin maps provide effective visualisation of historical events, it is difficult to discover trends and potential locations of future events. This is true for both crime visualisation and other types of pin map visualisations such as fire events. An example of a hot spot fire map produced by the Victorian Country Fire Authority (CFA) [86] is shown in Figure 2.8. In attempt to overcome this limitation, other maps have been developed to visualise spatial data.
Spatial visualisations can portray a range of different information about crime. This could include geographical crime mapping of actual locations of individual crimes as previously described or the collection of all crimes committed in any given region. It could also include temporal information about when particular crimes occurred. Other visualisations may include features of a specific crime incident and the display of this in a graphical format. The COPLINK project [16] includes many such visualisations. An example of this is shown in Figures 2.9 [22] and 2.10 [19].

While mapping and visualising information about raw data is useful to an extent, it does not necessarily provide deeper insight into the possible relationships between multiple features. Rather, it is used as a tool to display trends on a map. To display more complex information about relationships such as feature-crime, feature-feature, and crime-crime, such patterns first
need to be extracted from the raw data using techniques such as those described in the previous sections.

Visualisation of data mining results is an important task to improve the knowledge we can absorb from the data. Earlier work into non-spatial data mining visualisations has been widely studied. Data mining analysts typically investigate extremely large datasets that may generate a lot of results and patterns from the data mining algorithms. For example, an algorithm such as traditional ARM may produce a large number of rules, and a lot of interrelated items. In a supermarket basket analysis situation, often the analyst will first obtain raw data containing millions of rows. Within these rows there may be thousands of products. After running the
ARM procedure, the analyst will likely end up with many thousands of rules, depending on the pruning procedure. While it is possible to rank these rules from highest to lowest confidence (how correlated they are to one another), it is more difficult to see complex relationships spanning over more than two products. While the association rules mined are useful in their own form, more information can be gained by adopting a visualisation tool, and often other interesting knowledge can be gained by doing so. Visualisation of traditional association rules has been investigated by many researchers including [14, 15, 81, 82].

There exist many other data mining techniques, and there has been significant research into visualisation of the results of these mining models [67]. Visualisation of data mining outputs has proven to be a very important tool for visual feedback of the data mining algorithms. This is helped by better interpretation of the results from an improved understanding of how the algorithms work.

Visualisation of spatial data mining algorithm outputs is less studied. Compared to the traditional data mining techniques described above, spatial data mining is a relatively young field of research. Andrienko [4] present a variety of visualisation techniques, including the use of maps.
FIGURE 2.10: COPLINK Spatio-Temporal Visualizer showing Bank Robberies (September-December, 2001).

to show the accuracy of a classification algorithm and the distribution of female death rate and life expectancy. While their technique is useful for a number of analysis scenarios, it does not have the ability to show a large number of relationships as it focuses on just a few at a time.

The COPLINK system [16] is able to display information about a specific crime incident and its associated information mined from the crime data. This is particularly useful for investigation into particular crimes. However, it does not portray information about larger trends in criminal activity which are especially useful for prevention efforts. Such trends can provide a better ‘big picture’ view of places and events that are likely to occur together, and are the output of the cross-patterning techniques presented in this thesis. Representing these patterns visually can provide a great deal of insight into criminal behaviours.

The cross-patterning visualisation technique presented in this thesis can display several degrees of relationships, and multiple relationships on the one screen. It can also visually display the strength of the relationships though varying mechanisms such as the thickness of an edge or
line between two graphically represented features. It is then very quick to see the complex network of relationships between multiple entities and notice different trends. Many of these would be very difficult to identify in their plain textual form.

Crime data analysis is the first step in finding interesting facts about criminal behaviour. Visualisation is the next step, allowing domain experts to interpret a large amount of this valuable mined information which may be otherwise too difficult to absorb.

2.5 Summary

Geographic Information Systems have facilitated the collection of spatial data, including crime data. Although these systems are able to provide answers to some spatially oriented questions, they are unable to provide information about spatial patterns across multiple layers.

The need for such spatial pattern discovery and the unique nature of spatial data have encouraged the development of GDM techniques. In the field of crime data mining, efforts have been made to use spatial data for crime analysis. However current techniques face a variety of limitations. Some of the existing crime data mining research focuses on the application of existing data mining methods, rather than new approaches. Often this means they are task specific, and do not lead to the development of a unifying crime data mining framework. Other techniques may also focus on the thematic attributes within the data while ignoring the spatial attributes. Knowing the location of a crime is important for understanding crime patterns. Another issue is that often crime data is not point-based data, and may be censored by using areal aggregated statistics. Existing techniques may not be available to handle data in such a format.

There are a number of spatial data mining techniques that can be used to find patterns in geographic data. These include hot spot analysis, SARM and CLRM. Each of these techniques faces some limitations. Hot spot analysis only investigates one spatial layer, looking for regions of higher density. Spatial association rules mining discovers rules using a centric feature, and may require some interpretation of the rules. Co-location rules mining requires some user
input to decide on the distance that objects can be from one another in order to be considered as having some spatial relationship.

Our technique aims to enable the discovery of patterns across many spatial layers, taking into account the unique features of spatial data, while letting the data ‘speak for itself’.
CHAPTER 3

Datasets

Crime data is data recorded about criminal activities that have occurred in the past. The type of crime that is studied in this thesis is crime that has a physical component, such as robbery, assault or murder.

Crime data consists of three main components - the type of crime committed, where it was committed and when it occurred. That is, it consists of a thematic attribute that relates to the type of data - the type of crime committed, a spatial attribute - the location, either in geographical coordinates or in a geographic region and a temporal attribute - when the crime occurred.

The granularity of this information can vary depending on privacy laws and the individual policies of each police force. The crime data used in our study from the QPS contains information about the crime and when it occurred with the location aggregated by suburb area.

As crime incidents are a geospatial phenomena, they must be interpreted along with socio-demographic/socio-economic census data, geospatial feature data and textual metadata. Census data are mostly areal aggregated based on various areal units while feature data may be in point, line and area formats. For many reasons, crime datasets are recorded in areal aggregated formats. In particular, this is the case when limited environmental circumstances, ethical issues, or topological/geometrical computational considerations are in place. Environmentally, this is the case when the areal base map is of main interest (for instance, state division boundary map for state government). Ethically, this is when protecting the criminal’s or victim’s privacy from crime data mining and knowledge discovery is important. With geospatial crime data, crime incidents can be aggregated on an area basis to avoid delicate privacy and security issues,
however it is noted that area aggregates at quite small geographical scales may remove privacy issues while still providing useful information [10].

Computationally, it is when efficient and effective management of topological operations is necessary. Topological operations with areal objects have been identified as relatively easy and more established than other primitive data types [58]. Geometrically, converting primitive data types to more complex data types is well supported by existing GISs and computing tools through intersection, containment and buffering operations, and it is considered to be topologically restorable. However, transforming complex data types to primitive data types may cause loss of geometrical and topological information. For these reasons, the datasets we use from the QPS is recorded as areal aggregated crime information. The large areal aggregated geospatial crime datasets collected by the QPS need to be properly analysed along with possible causal socio-demographic geospatial factors.

The work in this thesis focuses on discovering cross-patterning from the thematic and spatial properties of spatially distributed crime data. By discovering information from combined sources of data, including crime data and data about geographical features such as parks and schools, crime analysts can gain a deeper insight into criminal patterns of behaviour.

### 3.1 Crime Data from the Queensland Police Service

**Figure 3.1:** Study region: (a) Map of Australia; (b) Map of the greater Brisbane area.
The crime datasets we analyse in this thesis are from the urban suburbs of Brisbane Australia. Steady increases in recorded violent crime in Australia [9] have been a major concern to policing agencies, tourism agencies and the general public. Figure 3.1(a) depicts the location of Brisbane within the state of Queensland in Australia while Fig. 3.1(b) outlines the suburbs of the greater Brisbane region. The Brisbane study region is highly dynamic and active and continues to experience significant and sustained population growth which has led to an increase in various criminal activities [64].

The distribution of crime in time and space is non-random. Criminal activity is patterned according to the location of criminogenic environments (for example crime opportunities or other environmental features). In Queensland, the highest residential burglary rates were in coastal tourist regions, inner city areas of Brisbane and a band of suburbs south of Brisbane between Ipswich and Beenleigh [24]. In this thesis we focus on the inner city areas of Brisbane as depicted in Figure 3.2. The crime data is released by the QPS in areal aggregated format due primarily to privacy concerns. There are 186 regions for which 50 crime types are recorded, and aggregations are made.

Different crime types can be the responsibility of different levels of law enforcement [17], and so data is recorded about crimes for which each enforcement group has responsibility. In
Queensland, the QPS records information from three major groups. These three main categories are shown in Table 3.1. The first category is ‘Personal Safety’ offences, which are offences made against a person such as homicide or assault. The second is ‘Property Security’ offences, which are offences made against a property, such as breaking and entering or arson. The final category ‘Other Offences’ consists of all other offences which do not fit into the previous two categories, such as drug offences or prostitution.

In addition, the subcategories could have several subsubcategories. For instance, homicide consists of attempted murder, conspiracy to murder, driving causing death and manslaughter. The crime tree in Fig. 3.3 depicts this hierarchical nature of the crime data while Table 3.2 lists all crime types and their abbreviations that are used in this study.
Figure 3.3: A crime taxonomy tree illustrating various crime types in Queensland.

3.2 Census Data from the Australian Bureau of Statistics

The crime data from the QPS is combined with spatial feature and census datasets, to allow for the discovery of cross-patterning relationships across multiple layers. This census data is obtained from the ABS.

The collection district is the smallest geographic unit at which census data is provided by the ABS. The ABS also produces statistics for other geographic areas by allocating collection districts to each spatial unit in the particular classification [8]. In our case the collection districts correspond to the suburb regions captured in the QPS dataset.

The data from the ABS consists of 63 census classifications. These include age, dwelling structure, employment status, weekly income, level of education completed and household mobility indicator (change of address). Table 3.3 details the census abbreviations used in this study.
3.3 Summary

We supplement the census data provided by the ABS with 7 geographic feature sets. These include caravan parks, railway stations, reserves, schools, hospitals, university/colleges and parks. This data is collected by a number of private companies and local government agencies with all datasets being converted into areal aggregated format as per the preprocessing outlined in Chapter 4.

### Table 3.2: Crime Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOAPE</td>
<td>Total Personal Safety</td>
</tr>
<tr>
<td>TOHO</td>
<td>Total Homicide</td>
</tr>
<tr>
<td>MURD</td>
<td>Murder</td>
</tr>
<tr>
<td>ATMU</td>
<td>Attempted Murder</td>
</tr>
<tr>
<td>COMU</td>
<td>Conspiracy to Murder</td>
</tr>
<tr>
<td>MSLT</td>
<td>Manslaughter</td>
</tr>
<tr>
<td>DRCD</td>
<td>Driving Causing Death</td>
</tr>
<tr>
<td>TOAS</td>
<td>Total Assault</td>
</tr>
<tr>
<td>SEAS</td>
<td>Serious Assault</td>
</tr>
<tr>
<td>OTAS</td>
<td>Other Assault</td>
</tr>
<tr>
<td>TOSO</td>
<td>Total Sexual Offences</td>
</tr>
<tr>
<td>RAAR</td>
<td>Rape/Attempted Rape</td>
</tr>
<tr>
<td>OTSO</td>
<td>Other Sexual Offences</td>
</tr>
<tr>
<td>TORO</td>
<td>Total Robbery</td>
</tr>
<tr>
<td>ARRO</td>
<td>Armed Robbery</td>
</tr>
<tr>
<td>UNRO</td>
<td>Unarmed Robbery</td>
</tr>
<tr>
<td>EXTO</td>
<td>Extortion</td>
</tr>
<tr>
<td>KAAE</td>
<td>Kidnapping &amp; Abduction</td>
</tr>
<tr>
<td>OFAP</td>
<td>Other Personal Safety</td>
</tr>
<tr>
<td>TOAPR</td>
<td>Total Property Security</td>
</tr>
<tr>
<td>TOUE</td>
<td>Total Unlawful Entry</td>
</tr>
<tr>
<td>UEWID</td>
<td>Unlawful Entry - Dwelling</td>
</tr>
<tr>
<td>UEWIS</td>
<td>Unlawful Entry - Shop</td>
</tr>
<tr>
<td>UEWIO</td>
<td>Unlawful Entry - Other</td>
</tr>
<tr>
<td>ARSO</td>
<td>Arson</td>
</tr>
<tr>
<td>OTPD</td>
<td>Other Property Damage</td>
</tr>
<tr>
<td>MOVT</td>
<td>Motor Vehicle Theft</td>
</tr>
<tr>
<td>OTTH</td>
<td>Other Theft</td>
</tr>
<tr>
<td>STFD</td>
<td>Stealing from Dwellings</td>
</tr>
<tr>
<td>SHST</td>
<td>Shop Stealing</td>
</tr>
<tr>
<td>STST</td>
<td>Stock Stealing</td>
</tr>
<tr>
<td>OTST</td>
<td>Other Stealing</td>
</tr>
<tr>
<td>TOFR</td>
<td>Total Fraud</td>
</tr>
<tr>
<td>FBCH</td>
<td>Fraud by Cheque</td>
</tr>
<tr>
<td>FBCC</td>
<td>Fraud by Credit Card</td>
</tr>
<tr>
<td>THSG</td>
<td>Total Handling Stolen Goods</td>
</tr>
<tr>
<td>PPSS</td>
<td>Possess Stolen Property</td>
</tr>
<tr>
<td>RESP</td>
<td>Receiving Stolen Property</td>
</tr>
<tr>
<td>PETP</td>
<td>Tainted Property</td>
</tr>
<tr>
<td>OHSG</td>
<td>Other Handling Stolen Goods</td>
</tr>
<tr>
<td>TOOO</td>
<td>Total Other Offences</td>
</tr>
<tr>
<td>DROF</td>
<td>Drug Offences</td>
</tr>
<tr>
<td>PROF</td>
<td>Prostitution Offences</td>
</tr>
<tr>
<td>LIDR</td>
<td>Liquor</td>
</tr>
<tr>
<td>GAOF</td>
<td>Gaming Offences</td>
</tr>
<tr>
<td>TRAV</td>
<td>Trespassing and Vagrancy</td>
</tr>
<tr>
<td>GOOO</td>
<td>Good Order Offences</td>
</tr>
<tr>
<td>TARO</td>
<td>Traffic and Related Offences</td>
</tr>
<tr>
<td>MIOF</td>
<td>Miscellaneous Offences</td>
</tr>
</tbody>
</table>

This chapter has introduced the type of datasets that are used throughout this thesis. As crime incidents are a geospatial phenomena, techniques for intelligent crime data analysis must also
### 3.3 Summary

#### Table 3.3: Census Abbreviations

<table>
<thead>
<tr>
<th>Household Mobility Indicator (change of address)</th>
<th>SAMEAD1Y_P</th>
<th>DIFFAD1Y_P</th>
<th>SAMEAD5Y_P</th>
<th>DIFFAD5Y_P</th>
<th>P_OVER_VIS</th>
<th>OVE_VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of Education Completed</td>
<td>YEAR8</td>
<td>YEAR9</td>
<td>YEAR10</td>
<td>YEAR11</td>
<td>P_1-39</td>
<td></td>
</tr>
<tr>
<td>Weekly Income</td>
<td>P_40-79</td>
<td>P_80-119</td>
<td>P_120-159</td>
<td>P_160-199</td>
<td>P_200-299</td>
<td>P_300-399</td>
</tr>
<tr>
<td></td>
<td>P_400-499</td>
<td>P_500-599</td>
<td>P_600-699</td>
<td>P_700-799</td>
<td>P_800-999</td>
<td>P_10001499</td>
</tr>
<tr>
<td></td>
<td>P_1500MORE</td>
<td>P_NOT_STA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment Status</td>
<td>EMP_TOT_P</td>
<td>UNEMP_P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dwelling Structure</td>
<td>SEPHOU</td>
<td></td>
<td>SDRTETWN1</td>
<td>SDRTETWN2</td>
<td>FLUNITAPP</td>
<td>FLUNITAPP1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FLUNITAPP2</td>
<td>FLUNITAPP3</td>
<td>FLUNITAPP4</td>
<td>ODCARAVAN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ODIMPHOM</td>
<td>ODHOUFATSO</td>
<td>OD_TOTAL</td>
<td>NOTSTATED</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UNOCCUPIED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>AGED_0_4</td>
<td>AGED_10_14</td>
<td>AGED_5_9</td>
<td>AGED_20_24</td>
<td>AGED_15_19</td>
<td>AGED_30_34</td>
</tr>
<tr>
<td></td>
<td>AGED_25_29</td>
<td>AGED_30_34</td>
<td>AGED_35_39</td>
<td>AGED_40_44</td>
<td>AGED_45_49</td>
<td>AGED_50_54</td>
</tr>
<tr>
<td></td>
<td>AGED_55_59</td>
<td>AGED_60_64</td>
<td>AGED_65_69</td>
<td>AGED_70_74</td>
<td>AGED_75_79</td>
<td>AGED_80_84</td>
</tr>
<tr>
<td></td>
<td>AGED_85_89</td>
<td>AGED_90_94</td>
<td>AGED_95_99</td>
<td>AGED_100+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analyse socio-demographic/socio-economic census data, geospatial feature data and textual metadata.

In this study we combine ABS census data with QPS crime data to provide more meaningful and useful insights into criminal patterns of behaviour. We present a case study using these datasets in Chapter 8.
CHAPTER 4

Cross-Association Patterns

Geographic Information Systems are used extensively in crime analysis as they allow us to model the real world along with many geographical layers (themes or coverages) represented by point, line and area data types. In such a multi-layered architecture, it is important to find cross-patterning relationships among layers for the discovery of thematically, geospatially and temporally correlated events. Detected patterns are a key resource for geospatial decision-making, such as planning, precaution, prediction and policy making.

In this chapter we introduce one way in which this cross-patterning can be modelled; cross-association patterns. Cross-association models the relationship between multiple datasets for each region, that is, it does not consider the influence of neighbouring regions in its calculation.

We propose a framework for automating the detection of these cross-associations based on a given areal base map. To achieve this we investigate a series of geospatial preprocessing steps involving data conversion and classification so that traditional ARM can be applied. ARM is a popular and well researched method for discovering interesting relations between variables in large databases and the resulting cross-associations can be used as the basis for decisions about crime prevention and planning.

Our framework can be fully integrated into traditional GISs using a dynamic link library and scripting languages such as AVENUE for ArcView, Python for ArcGIS, or MapBasic for MapInfo. This allows for greater automation of both the preprocessing and data mining phases and thus provides greater ease of use for users. In the following sections we describe the Areal Categorised Multivariate Association Mining (ACMAM) and provide examples of how it can be used to discover cross-association patterns from areal aggregated spatial datasets.
4.1 Areal Categorised Multivariate Association Mining

The ACMAM framework is a three-phase knowledge discovery process. It initially involves two preprocessing phases called aggregation and categorisation (classification). ARM is then applied to the processed datasets. Both of these steps are fully implemented within the GIS to allow greater automation of the process. In principle, it transforms all geospatial layers into areal aggregated data types based on a given areal base map. Areal aggregates are numerical values that will be further categorised into several groups. GISs provide a number of built-in categorisation schemes and we use these functions for this framework. Figure 4.1 describes the overall framework involving data aggregation, categorisation and ARM. The discovery process is interactive and iterative. Each phase will be discussed further in the subsequent subsections.

4.1.1 Data Preprocessing and Aggregation

Spatial datasets can be recorded at a number of different granularities. For example, crime data from the QPS is recorded as suburb aggregates, census data as collection district aggregates, schools as point locations and railway tracks as line objects. We must preprocess these heterogeneous data types before we can mine for cross-patterning. As the majority of our data is areal aggregated (crime and census datasets) we choose to transform all the datasets into area form.

In geometrical and topological transformations, the intersection operation is a basis to other geospatial operations such as meet, overlap, cover and contain [93]. In addition, it has been used as a basis for qualitative spatial reasoning [21]. The intersection operation is therefore a fundamental operation and is used as the default for aggregation and conversion in this framework. In the following subsections we discuss how to perform each type of dataset transformation.

4.1.1.1 Point-to-Area Aggregation

Point-to-area aggregation has been widely used in GISs. The containment or within function, so called point-in-polygon operation, can be used to aggregate point data. Aggregates are
Figure 4.1: The framework of areal categorised geospatial knowledge discovery.

typically visualised through choropleth maps using graduated colour to represent density. In this chapter, point-in-polygon operation is used in point data aggregation.

4.1.1.2 Line-to-Area Aggregation

Similar to point-to-area aggregation, we can use the intersection topological relationship in line-to-area conversion. Areal units that intersect with a target line object would represent the presence of the line object. These areal units imply proximity to the line object. This approach would work well when areal units are represented on a large scale to capture the details. One alternative is to use buffering. Buffering creates an enclosing polygon of a line
4.1 AREAL CATEGORIZED MULTIVARIATE ASSOCIATION MINING

object at a specified distance around the line. Once transformed, we can apply area-to-area data aggregation that will be discussed in the next subsection. In this chapter, we adopt the former as default.

4.1.1.3 Area-to-Area Aggregation

The intersection operation is still a solid candidate for this transformation. We adopt an approach somewhat similar to the areal-stealing (or areal-weighting) interpolation technique [30, 32]. Let \( B = \{b_1, b_2, \ldots, b_n\} \) be an areal base map with a set \( n \) of areal units. Let \( T = \{t_1, t_2, \ldots, t_m\} \) be an areal target map with a set \( m \) of areal units. Then, the target map is converted based on the following rules.

\[
b_i.area = \sum_{j=1}^{n} b_i.IntersectionArea(t_j). \tag{4.1}
\]

\[
b_i.value = \sum_{j=1}^{n} (b_i.IntersectionArea(t_j) \times t_j.value) / b_i.area. \tag{4.2}
\]

Formula 4.1 denotes the area of an areal unit in the base map is the sum of its intersection areas with target areal units. Formula 4.2 represents an aggregate value of the area in the base map is the sum of the products of stolen areas and corresponding values in the target map.

This is clarified by Figure 4.2. Figure 4.2(a) displays an areal base map while Figure 4.2(b) shows a target areal map with different boundaries. Let us assume that the area highlighted in Figure 4.2(a) is of interest and we denote this shaded area as \( b_1 \). It is labelled with the percentage of offences against a person over total offences in the same study region. Our task is to aggregate the target layer based on the areal base map shown in Figure 4.2(a). Figure 4.2(c) depicts an overlay of \( b_1 \) and the target map labelled with area intersecting with areal units in the target map. According to Formula 4.1, the area of \( b_1 \) (16.34) is the sum of stolen areas (3.37, 4.98, 6.15 and 1.84) shown in Figure 4.2(c). The value of \( b_1 \) (\( b_1.value \)) can be computed as follows based on Formula 4.2: \( b_1.value = (3.37 \times 0.27 + 4.98 \times 0.24 + 6.15 \times 0.51 + 1.84 \times 0.67)/16.34 \approx 40\% \). Thus, 40\% of offences in \( b_1 \) are regarded as offences against a person.
4.1 AREAL CATEGORISED MULTIVARIATE ASSOCIATION MINING

4.1.2 Categorisation and Classification

Before traditional association rules mining can be applied to areal aggregated data it must first be categorised into ranges. This discretisation should be made such that it will identify clearly the ranges that are present in the dataset. GISs provide various built-in categorisation schemes that can be used for highlighting dense areas including hot spots and local excesses.

ArcGIS offers natural breaks, quantile, equal area, equal interval and standard deviations (see Section 2.3.1 for definitions). Natural breaks is the default categorisation method in ArcGIS and this method identifies natural breakpoints between classes using a statistical formula called Jenk’s optimisation [25]. This method minimises the sum of the variance within each of the classes. It is noted that natural breaks find groupings and patterns inherent in the data. Because of this, it is chosen as the default method used in our approach. However, other classification

FIGURE 4.2: Area-to-area aggregation: (a) A base map; (b) A target map labelled with corresponding percentages of offences against person over total offences; (c) An overlay of the base map and the target map.
methods (or hot spot analysis techniques) can easily be integrated into the framework. Figure 4.3 shows an example of natural break categorisation. Figure 4.3(c) depicts a choropleth map generated by the default categorisation method with 5 classes.

**Figure 4.3:** Categorisation with natural breaks: (a) A base map with national parks; (b) The base map with histograms; (c) A choropleth map by natural breaks with 5 classes.

### 4.1.3 ACMAM

The last step of the framework for discovering cross-associations in areal aggregated datasets is ACMAM. Given an areal base map and preprocessed datasets we perform ARM to discover these cross-associations. ARM is a popular and well researched method for discovering interesting relations between variables in large databases and there are a number of algorithms available.
The Apriori algorithm has been dominant in ARM since it was first introduced in 1993 [1]. Many apriori-like algorithms [2, 43] have been suggested using different indexing schemes or pruning techniques [37]. The Apriori algorithm employs a candidate generation method that requires an iterative level-wise search where \( k \)-items are used to explore \((k+1)\)-items. It first finds frequent 1-itemsets that satisfy the minimum support constraint. These sets are then used to find frequent 2-itemsets which will be used to find frequent 3-itemsets. The downward closure property (all non-empty subsets of a frequent item must also be frequent) is used to improve the efficiency of the level-wise generation of frequent items. Once all frequent itemsets are generated, a minimum confidence is then used to eliminate low confidence rules, and return only strong rules. Readers can refer to Section 2.3.2 and the original paper by Agarwal [1, 2] for further details. A FP-Tree [38] approach that does not need to generate candidates has also been proposed and has proven to be more time efficient than the candidate generation approach and produces similar frequent patterns. We implement both algorithms in ACMAM and perform an experimental comparison in Chapter 8.
In our approach, spatial layers (features) correspond to items while areal units correspond to transactions. Featuresets denote sets of features. ACMAM is used to find frequent featuresets and strong rules based on a given areal base map. ARM is used to compute all strong rules.

We investigate two ways in which these datasets can be transformed for use with traditional ARM, these are boolean and quantitative classification. Figure 4.4(a-c) depict three areal aggregated datasets that will be used to describe the use of ACMAM. The areal base map is shown in Fig. 4.4(d). The simplest transformation possible to preprocess the datasets into a form suitable for ARM is the boolean approach. Table 4.1 shows this transformation for dataset4.1a–c. There are five areal regions labelled r1–5 in Fig. 4.4(d) and hence five transactions in the boolean classification. For each transaction (or region) we record a value of 0 if there were no recorded events (an aggregate of 0 in dataset4.1a–c) or a value of 1 if there were recorded events (an aggregate of more than 0 in dataset4.1a–c). For example, in Fig. 4.4(a), region r1 has a value of 4, in Fig. 4.4(b), region r1 has a value of 6 and in Fig. 4.4(c), region r1 has

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequent</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>A</td>
<td>0.6</td>
<td>0.75</td>
<td>0.94</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>0.6</td>
<td>0.75</td>
<td>0.94</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>0.4</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>0.4</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>0.4</td>
<td>0.50</td>
<td>0.83</td>
</tr>
<tr>
<td>C</td>
<td>B</td>
<td>0.4</td>
<td>0.50</td>
<td>0.83</td>
</tr>
<tr>
<td>B,C</td>
<td>A</td>
<td>0.2</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>A,B</td>
<td>C</td>
<td>0.2</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>B</td>
<td>A,C</td>
<td>0.2</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>A,C</td>
<td>B</td>
<td>0.2</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>C</td>
<td>A,B</td>
<td>0.2</td>
<td>0.25</td>
<td>0.63</td>
</tr>
<tr>
<td>A</td>
<td>B,C</td>
<td>0.2</td>
<td>0.25</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 4.2: Boolean ACMAM of datasets dataset4.1a–c.
4.1 AREAL CATEGORISED MULTIVARIATE ASSOCIATION MINING

### TABLE 4.3: Quantitative classification of datasets dataset4.1_a-c.

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequent</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_low</td>
<td>B_med</td>
<td>0.4</td>
<td>1.0</td>
<td>2.50</td>
</tr>
<tr>
<td>B_med</td>
<td>C_low</td>
<td>0.4</td>
<td>1.0</td>
<td>2.50</td>
</tr>
<tr>
<td>B_low</td>
<td>A_low</td>
<td>0.2</td>
<td>0.5</td>
<td>5.00</td>
</tr>
<tr>
<td>A_low</td>
<td>B_low</td>
<td>0.2</td>
<td>0.5</td>
<td>5.00</td>
</tr>
<tr>
<td>B_low,C_high</td>
<td>A_low</td>
<td>0.2</td>
<td>1.0</td>
<td>5.00</td>
</tr>
<tr>
<td>A_low,C_high</td>
<td>B_low</td>
<td>0.2</td>
<td>1.0</td>
<td>5.00</td>
</tr>
<tr>
<td>A_low,B_low</td>
<td>C_high</td>
<td>0.2</td>
<td>1.0</td>
<td>2.50</td>
</tr>
<tr>
<td>C_high</td>
<td>A_low,B_low</td>
<td>0.2</td>
<td>1.0</td>
<td>2.50</td>
</tr>
<tr>
<td>B_low</td>
<td>A_low,C_high</td>
<td>0.2</td>
<td>0.5</td>
<td>2.50</td>
</tr>
<tr>
<td>A_low</td>
<td>B_low,C_high</td>
<td>0.2</td>
<td>0.5</td>
<td>2.50</td>
</tr>
</tbody>
</table>

### TABLE 4.4: Quantitative ACMAM of datasets dataset4.1_a-c.

A value of 0. Therefore, the boolean transformation provides the values 1, 1 and 0 respectively for each of the features in region $r_1$.

Table 4.2 details the cross-associations generated by the Apriori ARM algorithm for this boolean classification. For demonstration purposes we do not constrain the results by minimum support, confidence or lift. For dataset4.1_a-c the strongest cross-association is found to be $C \Rightarrow A$ due to the fact that both dataset4.1_a and dataset4.1_c have one region of no recorded events (a value of 0) while dataset4.1_b has two regions of no recorded events.

The second way in which we can apply traditional ARM to areal aggregated datasets is quantitative classification. We use the same three synthetic datasets from Fig. 4.4(a-c) for explanation. Table 4.3 shows this transformation for dataset4.1_a-c. For this synthetic dataset we have chosen to categorise the data aggregates ($d$) into three ranges: low ($d < 2.5$), medium ($2.5 < d < 7.5$) and high ($d > 7.5$). Table 4.4 details the cross-associations generated by the Apriori ARM algorithm for this classification. In this case we have limited the results to
the first 10 cross-associations with a minimum support of 0.2. The strongest cross-association discovered is $C_{low} \Rightarrow B_{med}$. This is because regions $r_1$ and $r_2$ of dataset $A.1.c$ are both classified as low while regions $r_1$ and $r_2$ of dataset $A.1.b$ are both classified as medium. Care must be taken when selecting the classification method as it has a great impact on the discovered cross-associations.

Users can specify must-contain constraints in ACMAM to tailor their search. Constraints can be placed on the antecedent and/or consequent. The former (a must-contain featureset $\rightarrow ^*$) is used to explore possible associations of the featureset. For instance, a city council may want to explore the possible consequences of introducing a new casino into the area (casino $\rightarrow ^*$). If casino has a high cross-association with many criminal activities, then the council may not want to introduce one into the region. The latter ($^* \rightarrow$ a must-contain constraint) is used to explore possible stimuli of the featureset. In particular, this constraint is useful for crime data mining as it may reveal “possible lures” that attract a particular crime type.

### 4.2 Summary

Cross-patterning allows relationships between different spatial layers across locations to be modelled and quantified. One way in which cross-patterning can be modelled is by cross-association. Cross-association models the relationship between multiple datasets for each region, that is, it does not consider the influence of neighbouring regions in its calculation.

In this chapter we have proposed a framework for automating the detection of these cross-associations based on a given areal base map. We investigate a series of geospatial preprocessing steps involving data conversion and classification so that traditional ARM can be applied to areal aggregated spatial datasets. We then demonstrate how our ACMAM framework is able to discover cross-associations using a synthetic dataset with boolean and quantitative classification.
All of these individual pieces of our framework - the GIS data aggregation/overlay, the data classification, and the rule mining algorithm - are well established techniques. Our contribution is combining these techniques so that cross-associations can be discovered from areal aggregated datasets. It is the automated combination of these techniques that allows the integration of all heterogeneous data types that is not trivial. We are able to automate the procedure more fully, and potentially provide greater user interaction.

One strength of using ARM is the ability to discover multivariate cross-associations. Multivariate cross-associations are those that contain more than one antecedent or consequent, for example $A, B \Rightarrow C$ from Table 4.2. This is useful as it allows us to capture complex patterns within the areal aggregated datasets that might be easily missed by an analyst using visual inspection within a GIS.

By combining traditional ARM with quantitative classification we are able to discover cross-associations with discretised intervals. For example, we may discover that high classification of the crime ‘Drug Offences’ leads to a high classification of the crime ‘Stealing from Dwellings’. Discretised intervals can reveal more specific cross-associations and thus leads to better information for police and other groups. Once information such as this is discovered it can be used as the basis for decisions about crime prevention and planning, for example city councils may try to reduce drug use by increasing the number of rehab centres available.

One drawback of using ARM is that often a large amount of uninteresting patterns are discovered. Many ARM algorithms are designed to find frequent patterns and can thus be dominated by uninteresting frequent patterns. For example, frequent features such as streets, bus stops and intersections dominate the patterns, but we are more interested in a pattern of schools and break-in than one of streets and break-in, since streets are almost everywhere.

Cross-association patterns do not consider the influence of neighbouring regions. That is, it models the relationship between multiple layers for each region. Cross-associations are one type of pattern that comprise cross-patterning relationships and the combination of this technique and techniques presented in following chapters allow for a better understanding of areal aggregated spatial data and the patterns that may be contained within them.
In this chapter we introduced ACMAM which is a geospatial knowledge discovery framework that utilises ARM based on a given areal base layer to discover cross-association patterns. In Chapter 8 we show how ACMAM can be used to discover cross-association relationships for crime analysis and we also compare the computation aspect against the other techniques for discovering cross-patterning presented in this thesis.
Tobler’s first law of geography emphasizes an important property of spatial data in that we should carefully measure and analyse relations among near things. Cross-varying models cross-patterning over multiple spatial layers while considering the interaction with all neighbours in the local region (i.e. all neighbours that share a border with the region). For example, the number of vandalism incidents in two neighbouring regions may be related to a particular socio-demographic characteristic of all neighbouring regions.

Figure 5.1 shows two pairs of aggregated spatial datasets, \{Murder, Reserves\} and \{Reserves, Hospitals\} with colour coded data values (white=1, blue=2, red=3). The pair \{Murder, Reserves\} has a stronger visual correspondence than the pair \{Reserves, Hospitals\}. A cross-varying technique is able to model and quantify this visual correspondence.

Figure 5.1: Example cross-varying: (a) Murder (left) and Reserves (right); (b) Reserves (left) and Hospitals (right).
In this chapter we introduce the concept of spatial bivariate association. We analyse two such
techniques, Wartenberg’s Cross-MC (Moran Coefficient) and Lee’s $L$ index. These measures
are able to account for spatial relations in their computation and are thus able to identify spatial
structure within datasets. We then demonstrate how a spatial bivariate association measure
can be used to rank crime and geospatial feature datasets based on their spatial cross-varying.
This allows for the discovery of patterns in the form $\text{crime}_X$ is more similar to $\text{feature}_Y$
than $\text{feature}_X$. We show that Lee’s $L$ index [55] is an effective spatial bivariate association
measure for cross-varying analysis, however when used with areal aggregated datasets over
regions of irregular size and shape the results can be inconsistent. To overcome this we present
a more suitable weights matrix so that the spatial lag better represents neighbouring regions
of irregular size. We can then use this new weights matrix to extract the top-$k$ and bottom-$k$
cross-varying patterns.

5.1 Spatial Association

Spatial association is the degree to which a set of univariate observations are similarly arranged
over space. Spatial association 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 modelling and in understanding their
covariance [63]. Hubert et al. [46] 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:

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

(5.1)
and Moran’s $I$ is given by:

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

(5.2)

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 a measure of spatial association [26, 83] and we investigate this further in Section 5.3.

Figure 5.2 shows three hexagon shaped areal aggregated datasets (shaded regions indicate density: black=3, grey=2, white=1). The three pairs of dataset 5.1, A-B, B-C and A-C show identical point-to-point association as all show a Pearson’s correlation coefficient of 0.422 (each dataset has 5 black, 17 grey and 13 white regions). When analysing spatial data we must account for the spatial neighbourhood information that is not captured by the point-to-point association measure. For example, we can see in Figure 5.2 that dataset 5.1 A-B have a stronger visual correspondence than dataset 5.1 A-C.

It is argued that to capture spatial cross-varying, a bivariate spatial association measure should be the combination of both the point-to-point association and spatial association [55]. 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 [55, 89]. Wartenberg [89] develops a multivariate extension of the 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:

$$I_{X,Y} = \frac{\sum_i (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

(5.3)

Lee [55] utilises the concept of Spatial Lag (SL), which is the composed weighted averages of neighbours given by the spatial weights matrix, and is defined as:

$$\bar{x} = \sum_j w_{ij} x_j.$$ 

(5.4)

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:

$$SSS_X = \frac{n}{\sum_i (\sum_j w_{ij})} \cdot \frac{\sum_i (\sum_j w_{ij}(x_i - \bar{x}))^2}{\sum (x_i - \bar{x})^2}.$$ 

(5.5)

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:

$$L_{X,Y} = \sqrt{SSS_X} \cdot \sqrt{SSS_Y} \cdot r_{\bar{X},\bar{Y}}.$$ 

(5.6)

Given this definition of SSS we can now restate Wartenberg’s approach as:

$$I_{X,Y} = \frac{\sum_i (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \approx \sqrt{SSS_Y} \cdot r_{X,\bar{Y}}.$$ 

(5.7)

In the following sections we investigate the suitability of both Wartenberg’s Cross-MC and Lee’s $L$ index for discovering cross-varying relationships.

5.2 Problem Formulation and Statement

Lee [55] illustrates his bivariate spatial association measure with a small synthetic dataset consisting of 37 regular sized hexagons. We expand on that dataset by increasing the number of
hexagon regions to 397 to better represent the size of real world crime datasets. We evaluate both Lee’s $L$ and Wartenberg’s $I$ bivariate association measures with various datasets to demonstrate the potential suitability of these bivariate measures for discovering cross-varying relationships. We experiment with synthetic datasets that show increasing dissimilarity by first removing clusters from datasets and then moving clusters within a dataset. We also formulate and state problems of these measures in the following subsections.

### 5.2.1 Removing Clusters

![Figure 5.3: Synthetic hexagon dataset I: (a-d) Datasets dataset5.2D–D3.](image)

Figure 5.3(a-d) show synthetic datasets exhibiting increasing dissimilarity used for this experiment. We start with $n = 8$ clusters in $\text{dataset5.2}_D$, and for each subsequent dataset we remove 2 clusters. As clusters are removed, $\text{dataset5.2}_{D1–D3}$ become less similar to the original $\text{dataset5.2}_D$. Table 5.1 and Figure 5.4 illustrate this expected result; when we rank the similarity for $\text{dataset5.2}_D$, the bivariate association measure decreases with the number of clusters removed. For this dataset, both Lee’s $L$ and Wartenberg’s Cross-MC (noted as $I_{X,Y}$) show a similar pattern.
5.2 PROBLEM FORMULATION AND STATEMENT

As witnessed by Figure 5.4, datasets’ bivariate association values decrease as they become more dissimilar to dataset5.2_D. A linearly decreasing pattern is evident from (dataset5.2_D, dataset5.2_D) to (dataset5.2_D, dataset5.2_D3). A similar pattern can also be seen in Table 5.1 among (dataset5.2_D, dataset5.2_D1), (dataset5.2_D1, dataset5.2_D2) and (dataset5.2_D2, dataset5.2_D3).

We further examine the suitability of both Lee’s L and Wartenberg’s I with the synthetic hexagon dataset I (dataset5.2_D−D3).

We further examine the suitability of both Lee’s L and Wartenberg’s Cross-MC by expanding the synthetic 397 hexagon dataset to 817 hexagon regions. Figure 5.5(a) depicts the first dataset dataset5.3_a with n = 20 clusters, for each subsequent dataset we remove 1 cluster. We assert that as clusters are removed, dataset5.3_{b−t} become less similar to the original dataset5.3_a. Figure 5.5(b) depicts the final dataset dataset5.3_t (we do not show...
Table 5.2 shows that Lee’s $L$ index captures this decreasing similarity as clusters are removed from the datasets. Wartenberg’s Cross-MC does not achieve this expected result, for example ($\text{dataset5.3}_a$, $\text{dataset5.3}_i$) is shown to be more similar (has a higher $I$ value) than ($\text{dataset5.3}_a$, $\text{dataset5.3}_e$).
This discrepancy in the results is highly dependent on the dataset under analysis. Wartenberg’s Cross-MC is vulnerable to a reverse of the direction of association. For example, when an area $i$ with a higher than average value for both $X$ and $Y$ are surrounded by lower than average values, the numerator value in (5.3) could be given a negative value, because the $SL$ of $Y$ for the area is negative (the right part of the numerator), with the left part being necessarily positive. For these reasons we recommend Lee’s $L$ index as a more robust bivariate spatial association measure.

### 5.2.2 Moving Clusters

![Fig 5.6](image)

**Figure 5.6:** Synthetic hexagon dataset III: (a-d) Datasets $dataset_{5.4_{E-E3}}$.

Figure 5.6(a-d) depict the second set of synthetic datasets used for experimentation. Figure 5.6(a) shows two circular clusters in the bottom-left of the study region. Its rotated and moved clusters are shown in Figure 5.6(c). Figure 5.6(b) illustrates two elongated clusters in the bottom-right of the study region. Its rotated and moved clusters are depicted in Figure 5.6(d). As the datasets do not share any common clusters, there is no discernible spatial
association among the datasets. Table 5.3 and Figure 5.7 show that both Lee’s L and Wartenberg’s Cross-MC quantify this by calculating the spatial association between \((\text{dataset5.4}_E, \text{dataset5.4}_E)\), \((\text{dataset5.4}_E, \text{dataset5.4}_E)\) and \((\text{dataset5.4}_E, \text{dataset5.4}_E)\) to be near zero.

![Table 5.3: Cross-varying patterns from synthetic hexagon dataset III (dataset5.4).](image)

### Table 5.3: Cross-varying patterns from synthetic hexagon dataset III (dataset5.4).

<table>
<thead>
<tr>
<th></th>
<th>(SS_X)</th>
<th>(SS_Y)</th>
<th>(r_{X,Y})</th>
<th>(r_{X,Y})</th>
<th>(L_{X,Y})</th>
<th>(I_{X,Y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-E</td>
<td>0.471</td>
<td>0.471</td>
<td>1</td>
<td>1</td>
<td>0.471</td>
<td>0.555</td>
</tr>
<tr>
<td>E-E1</td>
<td>0.471</td>
<td>0.371</td>
<td>-0.091</td>
<td>-0.037</td>
<td>-0.038</td>
<td>-0.037</td>
</tr>
<tr>
<td>E-E2</td>
<td>0.471</td>
<td>0.471</td>
<td>-0.081</td>
<td>-0.037</td>
<td>-0.038</td>
<td>-0.037</td>
</tr>
<tr>
<td>E-E3</td>
<td>0.471</td>
<td>0.371</td>
<td>-0.091</td>
<td>-0.037</td>
<td>-0.038</td>
<td>-0.037</td>
</tr>
<tr>
<td>E1-E2</td>
<td>0.371</td>
<td>0.471</td>
<td>-0.091</td>
<td>-0.037</td>
<td>-0.038</td>
<td>-0.037</td>
</tr>
<tr>
<td>E1-E3</td>
<td>0.371</td>
<td>0.371</td>
<td>-0.101</td>
<td>-0.037</td>
<td>-0.037</td>
<td>-0.037</td>
</tr>
<tr>
<td>E2-E3</td>
<td>0.471</td>
<td>0.371</td>
<td>-0.091</td>
<td>-0.037</td>
<td>-0.038</td>
<td>-0.037</td>
</tr>
</tbody>
</table>

**Figure 5.7:** Comparison of Lee’s L and Wartenberg’s I with synthetic hexagon dataset III (dataset5.4).

### 5.2.3 Randomly Distributed Clusters

Figure 5.8(a-b) show two synthetic datasets illustrating uniformly distributed random values used for this experiment. Table 5.4 shows that both Lee’s L and Wartenberg’s Cross-MC detect no discernible spatial association between \((\text{dataset5.5}_1, \text{dataset5.5}_1)\), \((\text{dataset5.5}_1, \text{dataset5.5}_1)\).

These datasets illustrate the robustness of Lee’s L bivariate statistic for capturing cross-varying patterns involving complex spatial datasets. However, these are only tested with regular hexagonal datasets. The next section examines irregular datasets that better model complex real-world situations and extends them to top-\(k\) and bottom-\(k\) cross-varying mining.
Real world areal aggregated datasets are often recorded over regions of irregular size and shape. Crime datasets can be aggregated to suburbs or police districts and census datasets may be recorded using census tracts. In these cases, the choice of weight matrix is a critical step in the computation of spatial association [83].

Figure 5.9 shows a subset of suburbs from Brisbane, Australia. The QPS record crime statistics on a suburb level. Figure 5.9(a-c) show synthetic datasets over these base regions. We argue that dataset5.6\textsubscript{a} is more similar to dataset5.6\textsubscript{c} than dataset5.6\textsubscript{b} because dataset5.6\textsubscript{a} and dataset5.6\textsubscript{c} show a similar density change in a similar spatial neighbourhood. That is, neighbouring regions (sharing a boundary) in dataset5.6\textsubscript{a} and dataset5.6\textsubscript{c} show a similar density.

Lee’s $L$ index uses a simple weights matrix that is defined as the row standardised version of the binary connectivity matrix where elements that are neighbours have a value of 1 or 0.186.
Figure 5.9: Irregular region dataset: (a-c) Datasets dataset5.6\textsubscript{a-c}; (d) Region outline.

Table 5.6: Cross-varying patterns from irregular dataset dataset5.6 with modified weights matrix.

<table>
<thead>
<tr>
<th></th>
<th>(SSS_X)</th>
<th>(SSS_Y)</th>
<th>(r_{X,Y})</th>
<th>(r_{X,Y})</th>
<th>(L_{X,Y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-a</td>
<td>0.947</td>
<td>0.947</td>
<td>1</td>
<td>1</td>
<td>0.947</td>
</tr>
<tr>
<td>a-b</td>
<td>0.947</td>
<td>1.04</td>
<td>-0.107</td>
<td>0.971</td>
<td>-0.106</td>
</tr>
<tr>
<td>a-c</td>
<td>0.947</td>
<td>0.996</td>
<td>0.152</td>
<td>0.971</td>
<td>0.148</td>
</tr>
</tbody>
</table>

Table 5.6 shows that with this weights matrix, the \(L\) index determines that dataset5.6\textsubscript{a} is more similar to dataset5.6\textsubscript{b} than dataset5.6\textsubscript{c}. We argue that because of the similar density change among neighbours, (dataset5.6\textsubscript{a},dataset5.6\textsubscript{c}) should show a stronger (higher \(L\) value) cross-varying.

The reason for this result is that each neighbour is given the same weighting when calculating the spatial lag vector. With irregular regions this is often not desirable. For example,
Figure 5.10: Spatial lag of irregular dataset datasets 5.6: (a-b) Unmodified weight matrix; (c-d) Modified weight matrix.

Figure 5.9(d) shows two shaded regions that are neighbours, but only share a very small border. Figure 5.10 depicts the spatial lag of dataset 5.6a and dataset 5.6b where these particular regions are most influential. Lee’s SL will give all neighbours a SL value of 0.233, Figure 5.10(a-b). The spatial weights matrix should be defined so that the common border reflects the weighting and thus the spatial lag, Figure 5.10(c-d).

If \( i, j \) are neighbours, we define \( w \) as:

\[
w(i, j) = \frac{\text{sharedBorder}_{ij}}{\text{totalBorder}_i}, \tag{5.8}
\]

where \( \text{sharedBorder}_{ij} \) is the percentage of the distance of shared border between \( i \) and \( j \) and \( \text{totalBorder}_i \) is the total distance of the border of \( i \).

Given this new spatial weights matrix, Table 5.6 shows that Lee’s \( L \) index identifies \( (\text{dataset}5.6_a,\text{dataset}5.6_c) \) as more similar than \( (\text{dataset}5.6_a,\text{dataset}5.6_b) \). Table 5.7 shows the result of using the new weights matrix on the synthetic dataset from Figure 5.3. The same
pattern is detected - decreasing $L$ values as clusters are removed. The dataset from Lee’s [55] original 37 region dataset also produces consistent results (i.e. the same patterns are found).

![Figure 5.11: Synthetic dataset removing clusters: (a-d) Datasets dataset5.7.a–d.](image)

The need for this modified weights matrix is further highlighted when we examine a synthetic dataset covering the urban suburbs of Brisbane, Australia. Figure 5.11(a-d) depicts
5.4 Summary

Discovering cross-patterning in areal aggregated spatial datasets allows for a greater understanding of the various factors that can influence datasets. In this chapter we have introduced cross-varying patterns as one way in which cross-patterning can be modelled in areal aggregated spatial datasets. Techniques to discover these cross-varying relationships across different spatial layers are able to take into consideration the effect of all local neighbouring regions in their calculation.

Table 5.8: Cross-varying patterns from synthetic dataset dataset5.7 with unmodified weights matrix.

<table>
<thead>
<tr>
<th></th>
<th>$SSS_X$</th>
<th>$SSS_Y$</th>
<th>$r_{x,y}$</th>
<th>$r_{x,Y}$</th>
<th>$L_{X,Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-a</td>
<td>0.672</td>
<td>0.672</td>
<td>1</td>
<td>1</td>
<td>0.672</td>
</tr>
<tr>
<td>a-b</td>
<td>0.672</td>
<td>0.685</td>
<td>0.926</td>
<td>0.927</td>
<td>0.628</td>
</tr>
<tr>
<td>a-c</td>
<td>0.672</td>
<td>0.560</td>
<td>0.357</td>
<td>0.472</td>
<td>0.219</td>
</tr>
<tr>
<td>a-d</td>
<td>0.672</td>
<td>0.637</td>
<td>0.441</td>
<td>0.558</td>
<td>0.289</td>
</tr>
</tbody>
</table>

Table 5.9: Cross-varying patterns from synthetic dataset dataset5.7 with modified weights matrix.

<table>
<thead>
<tr>
<th></th>
<th>$SSS_X$</th>
<th>$SSS_Y$</th>
<th>$r_{x,y}$</th>
<th>$r_{x,Y}$</th>
<th>$L_{X,Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-a</td>
<td>0.800</td>
<td>0.800</td>
<td>1</td>
<td>1</td>
<td>0.800</td>
</tr>
<tr>
<td>a-b</td>
<td>0.800</td>
<td>0.798</td>
<td>0.914</td>
<td>0.927</td>
<td>0.730</td>
</tr>
<tr>
<td>a-c</td>
<td>0.800</td>
<td>0.838</td>
<td>0.492</td>
<td>0.558</td>
<td>0.403</td>
</tr>
<tr>
<td>a-d</td>
<td>0.800</td>
<td>0.835</td>
<td>0.430</td>
<td>0.472</td>
<td>0.351</td>
</tr>
</tbody>
</table>

synthetic areal aggregated datasets where the shaded colour of the suburbs indicate an aggregate data value between 0 and 10. We start with 4 clusters of dense suburbs in dataset5.7a, and for each subsequent dataset 1 cluster is removed. We assert that the linear decrease in the number of clusters results in these datasets showing increasing dissimilarity, that is, as clusters are removed, dataset5.7b–d become less similar to the original dataset5.7a. Table 5.8 shows that using the unmodified weights matrix does not achieve this expected result, that is (dataset5.7a,dataset5.7d) is shown to be more similar (has a higher $L$ value) than (dataset5.7a,dataset5.7c). Using the modified weights matrix as described above, Table 5.9 is able to confirm a increasing dissimilarity as clusters are removed from the datasets.
We introduce the concept of bivariate spatial association and show how this can be used to discover cross-varying relationships. We compare and contrast two techniques that are able to quantify this bivariate spatial association, Lee’s $L$ index and Wartenberg’s Cross-MC. We demonstrate that Lee’s $L$ index is more robustly able to capture bivariate spatial association and suggest it as a suitable candidate for cross-varying discovery.

Real world areal aggregated datasets are often recorded over regions of irregular size and shape. For example the QPS crime datasets are recorded for suburb regions. We demonstrate that without modification Lee’s $L$ index cannot reliably capture cross-varying with these irregular regions. We present a modification of the weight matrix used when calculating the spatial lag vector (which in turn is used to calculate Lee’s $L$ index). This modified spatial weights matrix is defined so that the size of the common border between regions reflects the weighting and thus the spatial lag. We then demonstrate that this modification is able to robustly discover cross-varying relationships in spatial datasets recorded over suburb regions of irregular shape.

These discovered cross-varying relationships can be of use to a number of people and agencies such as regional planners, politicians, police and residents. For example, we may discover that a particular suburb with a high number of schools shows a cross-varying relationship with high incidents of vandalism in that suburb and surrounding suburbs. An example use of this information might be that schools in those areas increase preventative strategies such as closely monitoring student attendance. The local council may also use the information to encourage community awareness of the issue and implement neighbourhood watch programs. This cross-varying relationship between schools and vandalism may not be the causal factor of vandalism, but the information it can provide is still valuable.

In Chapter 8 we show how this cross-varying approach can be used for crime analysis and we also compare the computation performance of Lee’s $L$ bivariate spatial association measure against other approaches discussed in this thesis for discovering cross-patterning.
Cross-patterning allows relationships between different areal aggregated spatial layers across locations to be modelled and quantified. The previous two chapters have discussed two ways in which cross-patterning can be modelled; cross-association and cross-varying. Cross-association ignores the effect of neighbouring regions and is useful for discovering multivariate relationships. Cross-varying patterns take into consideration the effect of all local neighbouring regions and is well suited when analysing spatial datasets.

In this chapter we introduce a third way to model cross-patterning in areal aggregated spatial datasets. Cross-distribution uses the spatial distribution of a dataset to model patterns over multiple layers and can consider the spatial distribution of both one local neighbour (i.e. one neighbour that shares a border with the region) as well as the global distribution (modelling the whole study region). Figure 6.1 shows two pairs of aggregated spatial datasets, \(\text{\{Rape, Railway Stations\}}\) and \(\text{\{Fraud, College\}}\) with colour coded data values (white=1, blue=2, red=3). We can see that the spatial distribution of the pair \(\text{\{Rape, Railway Stations\}}\) has a stronger visual correspondence than that of the pair \(\text{\{Fraud, College\}}\). That is, they show a similar density change in the bottom-right regions. A cross-distribution technique is able to model and quantify this visual correspondence.

Cross-varying and cross-distribution patterns are similar in that they are both able to take into consideration the effect of neighbouring regions. The difference between the two is how this is achieved and which neighbours are taken into account. Cross-varying uses the combination of point-to-point association and neighbouring spatial dependence while cross-distribution uses the spatial distribution of the entire dataset.
We introduce two new techniques that can discover these cross-distribution relationships in areal aggregated spatial datasets. Density Tracing models areal aggregated spatial datasets as density traces projected onto the Cartesian plane following a defined spatial ordering of the regions within the dataset. These traces can then be compared and the spatial distribution similarity between datasets quantified and used to extract cross-distribution relationships. We then propose an efficient Graph Mining approach that models the spatial distribution of areal aggregated spatial datasets as weighted directed multigraphs. These graphs are then compared by extracting each vertices degree to compute a correlation coefficient such as Pearson’s $r$ or Jaccard’s Index $J$ so that cross-distribution patterns can be discovered.

### 6.1 Density Tracing

The first technique we propose to capture these cross-distribution relationships utilises a density trace to model the spatial distribution of areal aggregated datasets. We are able to account for local neighbourhood information (capturing nearby neighbours) and the global distribution (modelling the whole study region) within the density trace.

The density traces are projected onto the Cartesian plane following a defined spatial ordering of the regions within the dataset. The region ordering should reflect how density changes from neighbour to neighbour so that we are able to retain as much spatial information as possible.
We examine four popular locational ordering methods to determine the spatial ordering of the areal units in the study region: Guided Local Search (GLS), Depth First Search (DFS), Breadth First Search (BFS) and a Nearest Neighbour (NN) technique. We show that in general, the GLS technique is the most suited method to capture neighbours (i.e. those regions that share a border). The distance (dissimilarity) between two density traces is calculated using a modified Locality In-between Polylines (LIP) distance measure [70]. Intuitively, two density traces are considered spatially similar when they move close (i.e., their traces approximate each other) at the same place. To the best of our knowledge, this is the first attempt at using density traces/routes to represent the spatial distribution of areal aggregated datasets for crime data mining. Our density tracing approach efficiently and robustly reveals the top-$k$ positive and negative cross-distribution relationships.

6.1.1 Working Principle

The working principle of our framework is explained using the example synthetic datasets shown in Fig. 6.2. The three datasets show the centre point and density values of each region. We use the spatial distribution of density values to model the cross-distribution relationship between regions belonging to different datasets across the study region. Datasets dataset6.1$a$ and dataset6.1$b$ show a similar density distribution: higher than normal density in the shaded regions. Dataset dataset6.1$c$ however has a more consistent density pattern with all regions showing a similar density.

We will now present an overview of how Density Tracing discovers cross-distribution relationships. Details of each step will be given in Section 6.1.2. Initially, the three areal aggregated datasets dataset6.1$a$–$c$ are loaded and the centre point of each region is calculated. As we are discovering cross-distribution relationships within a single study region, all the datasets share a common base map, thus the centre point of each region only needs to be calculated once. The density values are then normalised into the range $[0, 1]$. The next step is to determine the spatial ordering of regions. In this example the GLS method is used with a random starting location, Fig. 6.2(d) shows the resulting order with the starting region labelled 1. Once the spatial ordering of regions is determined, the density traces for each dataset can be calculated.
Figure 6.2: Working principle of our framework: (a-c) Datasets dataset\textsubscript{6.1\_a-c} before normalisation; (d) GLS region ordering; (e) Density trace; (f) Density trace with area highlighted.
Figure 6.2(e) depicts these density traces with each line representing one dataset. These density traces represent the spatial distribution of the density values projected onto the Cartesian plane. In this example we wish to find patterns involving dataset6.1_a, thus we select dataset6.1_a as the reference feature \( f \). The similarity value between \( f \) and each dataset is then computed using a modified Locality In-between Polylines technique, with the basic idea being to calculate the area of the shape formed by the two 2D density lines. Figure 6.2(f) highlights the areas that need to be computed to calculate the similarity between dataset6.1_a and dataset6.1_b. The results can then be pruned with a user supplied minimum similarity \( \text{min_sim} \) or the user can simply retrieve the \( k \) Most Similar and/or \( k \) Least Similar results. The textual output using this example dataset is as follows, showing that dataset6.1_a is more similar to dataset6.1_b than dataset6.1_c:

Reference Feature selected: dataset6.1_a
Feature: dataset6.1_b Dissimilarity: 0.735477
Feature: dataset6.1_c Dissimilarity: 1.476240

6.1.2 Algorithm

The Density Tracing approach is detailed in Algorithm 6.1.1. To calculate a density trace for areal aggregated crime and feature datasets, they must first be preprocessed. Line 3 determines the centre point of each region, which can then be used to determine the density trace and neighbouring locations (we define neighbours as those that share a boundary). Areal aggregated spatial datasets are those that record aggregate values for each region in a study area (such as number of school for each suburb in a city). To discover cross-patterning we analyse datasets that cover the same study region, i.e. all datasets share a common base map. This means that the centre points of each region only needs to be calculated once. The centre of a region is determined by taking the centre point of the region’s bounding box. If the centre does not fall inside the region itself the point is moved horizontally on the X-axis until it enters into the region. The point is then moved in the same direction, along the X-axis, until it exits the polygon. The centroid is calculated to be halfway between the two points, on the same X-axis. For complex polygons that have more than one pair of polygon outlines that cross the X-axis,
each pair of outlines is compared to see which pair creates the widest length along the X-axis. Then, the centroid is calculated to be halfway between the points where this pair of outlines crosses the X-axis. This is the same as the standard GIS technique employed by ArcGIS for calculating the centre point of a polygon [73].

Algorithm 6.1.1 Density Tracing for cross-distribution patterns

Input: A set $D = \{d_1, d_2, \ldots, d_n\}$ of spatial areal aggregated datasets, a common base map $B = \{b_1, b_2, \ldots, b_l\}$, the required spatial ordering algorithm $order\ Alg$, a reference feature $f$, and a minimum similarity $minSimilarity$;

Output: A set $S = \{s_1, s_2, \ldots, s_m\}$ of ordered features with $dissimilarity_f < minSimilarity$;

1: LoadDatasets($D$);
2: CalcCenters($B$):
4: for each region $r$ in $B$ do
5: \hspace{1em} $cp \leftarrow centerPointOfBoundingBox(r)$
6: \hspace{1em} if $cp$ is not within $BoundingBox(r)$ then
7: \hspace{2em} point1: Move $cp$ in the X direction until it enters into $r$
8: \hspace{2em} point2: Move $cp$ in the X direction until it exits $r$
9: \hspace{1em} Centroid $cp$ is calculated to be halfway between point1 and point2 on the same X axis
10: \hspace{1em} if More than one pair of polygon outlines cross the X-axis then
11: \hspace{2em} Compare each pair of outlines to see which pair creates the widest length along the X-axis
12: \hspace{2em} Calculate centroid $cp$ as halfway between the points where this pair of outlines crosses the X-axis
13: \hspace{1em} end if
14: end if
15: return $cp$
16: \textbf{end for}
17:
18: NormalizeDensity(D):
19: \textbf{for all} \( d_i \in D \) \textbf{do}
20: \hspace{1em} upperLimit \leftarrow 1
21: \hspace{1em} lowerLimit \leftarrow 0
22: \hspace{1em} Hi \leftarrow d_i.max()
23: \hspace{1em} Lo \leftarrow d_i.min()
24: \hspace{1em} fact \leftarrow (upperLimit - lowerLimit)/(Hi - Lo)
25: \hspace{1em} \textbf{for all} \( x \in d_i \) \textbf{do}
26: \hspace{2em} d_i[x] \leftarrow (d_i[x] - Lo) \times fact + lowerLimit
27: \hspace{1em} \textbf{end for}
28: \textbf{end for}
29:
30: RegionOrdering(B, orderAlg);
31:
32: DT \leftarrow \text{DensityTrace}(D):
    \hspace{1em} We have to project the normalised density onto the Cartesian plane according to the region ordering
33: \textbf{for all} \( d_i \in D \) \textbf{do}
34: \hspace{1em} \textbf{for} iterator \( r \leftarrow \text{regionOrder.begin()} \) \textbf{do}
35: \hspace{2em} Get the corresponding density value (d1) for region \( r \) from \( d_i \)
36: \hspace{2em} point(X, Y) \leftarrow p1((\ast r), d1)
37: \hspace{2em} Get the next region \( r \) from \text{regionOrder}: \( r++ \)
38: \hspace{2em} \textbf{if} \( r! = \text{RouteSet.end()} \) \textbf{AND} \( r! = \text{NULL} \) \textbf{then}
39: \hspace{3em} Get the corresponding density value (d2) for region \( r \) from \( d_i \)
40: \hspace{3em} point(X, Y) \leftarrow p2((\ast r), d2)
41: \hspace{3em} Make segment between \( p1, p2 \)
42: \hspace{3em} Store segment in list: \( DT \)
43: \hspace{1em} \textbf{end if}
for all $d_i \in D$ do
  list similarity $S$
  $localS \leftarrow \text{CalcSimilarity}(f, d_i, \text{minSimilarity})$
  if $localS! = -1$ then
    $S$.append($localS'$)
  end if
end for
$S$.sort()
return $S$

{Note: Library of Efficient Data types and Algorithms (LEDA) utilised to provide geometric algorithms and data types such as intersection of lines, area of polygons, etc.}

Line 18 of the density tracing framework normalises the density values so that a meaningful similarity can be measured. The datasets must be normalised as the density traces are projected onto the Cartesian plane, with density as the Y axis. To measure similarity we compute the area formed between two density traces, so to successfully compare between pairs of datasets the density values must be in a common range. This is achieved by using the *min-max normalisation* [77] technique which retains the original distribution of scores except for a scaling factor and transforms all the density values into the common range $[0, 1]$. Line 30 (detailed in Alg. 6.1.2) of the algorithm is to determine the spatial ordering of the regions so that the density trace can be determined. We must preserve the spatial information contained within the data so that the discovered patterns describe the spatial distribution (the effect of this ordering is investigated in Section 6.1.4). We are concerned with how density changes from location to location, from neighbour to neighbour, and the region ordering should reflect this. Again, as all the datasets are assumed to share a common base map, this spatially aware region ordering only needs to be calculated once for all datasets.
Algorithm 6.1.2 Density Tracing: RegionOrdering

**Require:** Centre point for each region in the base map has been calculated;

**Input:** A common base map $B$ and the required spatial ordering algorithm $orderAlg$;

**Output:** A list of regions $order$, ordered by the spatial ordering algorithm $orderAlg$;

1: Starting region can either be random or user supplied; polygon $p = randomRegion()$
2: if $orderAlg = GLS$ then
3: Use Fast Local Search, the 2-Opt heuristic as the move operator, no construction heuristic and $Alpha \leftarrow 0.167$
4: Invoke GLS algorithm using a random starting solution
5: Refer to Voudouris [87] for complete algorithm
6: else if $orderAlg = DFS$ then
7: list order
8: stack < polygon > $S$
9: $S.push(p)$
10: while !$S.empty()$ do
11: polygon $local \leftarrow S.pop()$
12: $order.append(local)$
13: for Each neighbouring region $n$ of $local$ ordered by distance do
14: if $n$ is not already in $order$ or $S$ then
15: $S.push(n)$
16: end if
17: end for
18: end while
19: return $order$
20: else if $orderAlg = BFS$ then
21: list order
22: queue < polygon > $Q$
23: $Q.append(p)$
6.1 Density Tracing

24: while !Q.empty() do
25: polygon local ← Q.pop()
26: order.append(local)
27: for Each neighbouring region n of local ordered by distance do
28: if n is not already in order or Q then
29: Q.append(n)
30: end if
31: end for
32: end while
33: return order
34: else if orderAlg = NN then
35: list order
36: order.append(p)
37: local base map lb ← B
38: for all region r ← p ∈ lb do
39: list polygon nnList ← nearestNeighbors(r, 5)
40: for all region n ∈ nnList do
41: if n is not already in order then
42: order.append(n)
43: lb.remove(n)
44: end if
45: end for
46: end for
47: return order
48: end if
Algorithm 6.1.3 Density Tracing: CalcSimilarity

Require: Density Trace for all datasets stored in $DT$;
Input: A reference feature $f$, dataset $d$ and minimum similarity $minSimilarity$;
Output: The dissimilarity $dissimilarity$ between traces of $f$ and $d$;

1: segment list $referenceSegment ← DT_f$
2: segment list $datasetSegment ← DT_d$
3: length $l ← 0$
4: point list $areaPoints$
5: last intersection segment/point $lastIntersect,lastIntersectPoint$
6: number regions in polygon $numR ← 0$
7: region weight for polygon $rw ← 0$

8: for $s1 ∈ referenceSegment, s2 ∈ datasetSegment$ do
9:    if $INTERSECTION(s1, s2)$ then
10:       Save the intersection point: $areaPoints.append(INTERSECTION(s1, s2))$
11:       Work back from intersection point of other line ($datasetSegment$):
12:          iterator $it ← s2$
13:          while $it! = NULL$ do
14:             if $it = lastIntersect$ then
15:                Save the intersect point: $areaPoints.append(lastIntersectPoint)$
16:                Save the length of the segment
17:                break
18:          else
19:                Save the start point: $areaPoints.append((∗it).source())$
20:                Save the length of the segment
21:          end if
22:          $it ← −$
23:      end while
24:    Save the length of the segments up to the intersection point
6.1 Density Tracing

25: Increase num regions covered by polygon: \(\text{numR} \leftarrow \text{numR} + 1\)
26: Update region weight covered by polygon: \(\text{rw} \leftarrow \text{rw} + \text{RegionWeight}(s1.\text{source})\)
27: Make polygon: \(\text{polygonP(areaPoints)}\)
28: weight \(w \leftarrow l/((\text{totalSegmentLength}(f) + \text{totalSegmentLength}(d))\)
29: avg region weight \(\text{rw} \leftarrow \text{rw}/\text{numR}\)
30: \(\text{dissimilarity} \leftarrow \text{dissimilarity} + \text{areaPoints} \times w \times \text{rw}\)
31: if \(\text{dissimilarity} > \text{minSimilarity}\) then
   \hspace{1em} return \(-1\)
   end if
32: \hspace{1em} Reset variables: \(\text{areaPoints.clear()}, l \leftarrow 0, \text{numR} \leftarrow 0, \text{rw} \leftarrow 0\)
33: \hspace{1em} \text{lastIntersect} \leftarrow s2, \text{lastIntersectPoint} \leftarrow \text{INTERSECTION}(s1, s2)
34: else if \(s1 = \text{referenceSegment.end()}\) then
35: Make a polygon from intersection point to end of segment
36: Save the length of the segments up to the end of segment
37: Increase num regions covered by polygon: \(\text{numR} \leftarrow \text{numR} + 1\)
38: Update region weight covered by polygon: \(\text{rw} \leftarrow \text{rw} + \text{RegionWeight}(s1.\text{source})\)
39: Make polygon: \(\text{polygonP(areaPoints)}\)
40: weight \(w \leftarrow l/((\text{totalSegmentLength}(f) + \text{totalSegmentLength}(d))\)
41: avg region weight \(\text{rw} \leftarrow \text{rw}/\text{numR}\)
42: \(\text{dissimilarity} \leftarrow \text{dissimilarity} + \text{areaPoints} \times w \times \text{rw}\)
43: if \(\text{dissimilarity} > \text{minSimilarity}\) then
   \hspace{1em} return \(-1\)
   end if
44: \hspace{1em} Reset variables: \(\text{areaPoints.clear()}, l \leftarrow 0, \text{numR} \leftarrow 0, \text{rw} \leftarrow 0\)
45: else
46: \hspace{1em} Continue checking for intersection until last segment
47: Save the length of these segments: \(l \leftarrow l + s1.\text{length}() + s2.\text{length}()\)
48: Add reference segment start/end point:
   \hspace{1em} \text{areaPoints.append}(s1.\text{source}()), \text{areaPoints.append}(s1.\text{target}())
49: end if
We investigate four popular linear techniques that can be used to determine this spatially aware region ordering; greedy GLS, DFS, BFS and a NN technique. As can be seen from Fig. 6.3, the GLS and NN-5 orders are better able to preserve spatial neighbourhood information than both DFS and BFS. Visually, it can be seen that the GLS route is the only region ordering that does not have overlapping edges (that is, edges that cross over another edge), also Fig. 6.3(f) shows that compared with GLS and NN-5, DFS and BFS exhibit a significantly larger edge length variance.

There are a number of different techniques that can be used to determine spatial neighbours and orders. Neighbourhood graphs such as the Relative Neighbourhood Graph (RNG), Gabriel Graph (GG) and the Delaunay Triangulation (DT) [66] cannot be used directly with this approach as they produce a nonlinear ordering. The Density Tracing approach outlined in this chapter projects the normalised density onto the Cartesian plane following a defined spatial ordering. The ordering must be linear as the algorithm computes a dissimilarity score by calculating the area formed between two density traces when overlaid. Each region can only be connected to a maximum of two other regions in the ordering. Similarly, space filling curves such as Morton-order or Hilbert curve [75] also cannot be used directly as the base map contains irregular regions that are not well represented by these curves.

Determining region ordering for use in our framework is similar to the symmetric Travelling Salesman Problem (TSP) where the cost between regions is the Euclidean distance. Starting from Region\textsubscript{x} we wish to order all regions in the study area so that the next region visited is a spatial neighbour (sharing a border). Region\textsubscript{x} can be chosen at random or, perhaps more usefully, it can be chosen based on some real world property (for example the Central Business
6.1 Density Tracing

FIGURE 6.3: Spatially aware region ordering: (a) Base map including centre points; (b) GLS; (c) DFS; (d) BFS; (e) NN-5; (f) Variance of edge lengths.

<table>
<thead>
<tr>
<th>Order</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLS</td>
<td>0.000082</td>
</tr>
<tr>
<td>DFS</td>
<td>0.007476</td>
</tr>
<tr>
<td>BFS</td>
<td>0.006869</td>
</tr>
<tr>
<td>NN-5</td>
<td>0.000766</td>
</tr>
</tbody>
</table>
6.1 DENSITY TRACING

District of a city. GLS is an intelligent search strategy for combinatorial optimisation problems. The technique sits on top of local search procedures and has as the main aim to guide these procedures for exploring efficiently and effectively (see [88] for the complete algorithm). We generate a random ordering starting from Region\(_x\) and then apply GLS to generate the ordering that minimises the Euclidean distance between regions (i.e. ideally we wish to move from neighbour to neighbour). We use the suggested GLS parameters from [88] (Fast Local Search, 2-Opt heuristic move operator, no construction heuristic and \(Alpha = 0.167\)).

DFS starts from Region\(_x\) and pushes each neighbouring region onto a stack. The closest neighbour is then popped from the stack and added to our region ordering. It continues in this fashion before backtracking to Region\(_x\). The process then repeats for the remaining neighbours of Region\(_x\). BFS is similar to DFS except for the use of a queue instead of a stack. It starts at Region\(_x\) and adds all neighbouring regions to the region ordering (in order of distance) and then continues the search outwards.

The NN technique that is used first adds the five (NN-5) closest neighbours of Region\(_x\) to the region ordering. Then the next closest neighbour is selected (i.e. the 6th closest to Region\(_x\)) which becomes the new starting point. This process of adding the closest five neighbours is repeated until there are no regions left. We choose five closest regions as for our suburb dataset it is noted that suburbs have on average five neighbours.

Note that the ordering approaches adopted in this chapter utilise the combination of topological information (neighbouring regions sharing boundaries) and geometric information. Other approaches such as the Minimum Spanning Tree (MST) and cumulative distance ordering could be used with modification, however the former is not linear (that is, it does not visit the different nodes sequentially), whilst the latter only uses topological information. The effect of region ordering is experimentally investigated in Section 6.1.4.

Once the spatial ordering of regions in the study area is determined, Step 32 of Algorithm 6.1.1 is to calculate the density traces of each dataset. This trace projects the normalised density onto the Cartesian plane (following the defined ordering from Step 30) and depicts the spatial
distribution of the density values within the study region. Figure 6.4(a) shows the density trace for two features and Figure 6.4(b) a subset of the density traces for 34 datasets.

To discover patterns of similar spatial distribution the set of density traces needs to be queried for similarity. To measure similarity we need to quantify the distance between two density traces. We modify the Spatial Trajectory Similarity Search technique \cite{70} to incorporate region weights. Given a reference feature $f$, the most similar trace in $D$ with respect to $f$ is the one that minimises the distance measure Region Weighted Locality In-between Polylines (RWLIP). Intuitively, two traces are considered spatially similar when they move close (i.e., their traces approximate each other) at the same place. As such, RWLIP defines a distance function upon the traces (projected on the Cartesian plane) where the idea is to calculate the area of the shape...
formed by the two traces. Note that this distance measure is equivalent to dissimilarity, i.e. a lower distance measure equates to higher similarity.

The distance measure between two traces $Q$ and $S$ is defined as follows:

$$RWLIP(Q, S) = \sum_{\forall \text{polygon}_i} \text{Area}_i \cdot w_i \cdot \text{region}_i,$$

(6.1)

where $\text{polygon}_i$ is a member of the set of polygons formed between intersection points $I$ created when $Q$ and $S$ are overlaid in the 2D plane, $w_i \in [0, 1]$ is a contribution weight and $\text{region}_i \in [0, 1]$ is a region weight. Figure 6.5 illustrates the respective areas that contribute in $RWLIP(Q, S)$. Let $\text{Length}_Q(I_i, I_{i+1})$ and $\text{Length}_S(I_i, I_{i+1})$ be the length of the trace that participates in the construction of a given polygon. The contribution weight can then be defined as follows:

$$w_i = \frac{\text{Length}_Q(I_i, I_{i+1}) + \text{Length}_S(I_i, I_{i+1})}{\text{Length}_Q + \text{Length}_S},$$

(6.2)

that is, the numerator is the perimeter of the polygon in question, while the denominator is the sum of the total length of the routes. It is a weight of how much a certain polygon contributes to the whole trace.

Each region of the reference feature $f$ can have a user specified weight. It is designed to enable the combination of clustering results into the density trace similarity algorithm. The $\text{region}_i$ is defined as the average region weight of all regions that contribute to the polygon $i$. If no region weight is specified, then each areal unit is equally important and the global trend (the entire set of areal units) is used for the similarity calculation. On the other hand, the user can assign different weights to regions so that high peak areal units (clusters) have more effect on the similarity score. This enables the user to incorporate both context sensitive weighting and clustering into our system.

As the RWLIP algorithm traverses the spatial ordering, if there are no intersections between $Q$ and $S$ (the traces are parallel) then the algorithm detects this and closes the segments by connecting the initial points of $Q$ and $S$ and the final points of $Q$ and $S$. The algorithm can
then proceed with this one area. The range of the similarity measure RWLIP is dependent
on the density range (Y axis) and number of regions (X axis). A lower dissimilarity measure
equates to higher similarity between the density traces.

The last step of the algorithm is to save any features that show a similarity to our reference
feature \( f \). We can either use a user supplied minimum similarity \( \text{min}_\text{sim} \) or retrieve the \( k \)
Most Similar and/or \( k \) Least Similar results.

### 6.1.3 Time Complexity Analysis

To analyse the time complexity of our approach each function of Algorithm 6.1.1 is analysed
separately. Given \( n \) as the number of spatial areal aggregated datasets in \( D \) and \( l \) as the number
of spatial regions in the base map \( B \), \( \text{LoadDatasets}(D) \) is linear to \( n \). The time complexity
of \( \text{CalcCenters}(B) \) is dependent on the regions of \( B \). Typically \( \text{CalcCenters}(B) \) is linear
to \( l \), however if \( B \) contains convex polygon regions then the complexity is \( O(l \log l) \) [73]. To
normalise \( D \), for each dataset \( d_n \) we must examine \( l \) values, and thus \( \text{NormaliseDensity}(D) \)
has a time complexity of \( O(n \times l) \). To generate the spatial ordering of our base map \( B \) using the
default GLS method, \( \text{RegionOrdering}(B, \text{orderAlg}) \) typically requires \( O(l \log l) \) [88]. To
generate the density trace of each dataset \( d_n \) we need to examine \( l \) regions, thus \( \text{DensityTrace}(D) \)
requires \( O(n \times l) \). Note that the original \( \text{LIP}(Q, S) \) computation requires \( O(l \log l) \) [70], thus
our RWLIP extension also has a time complexity of \( O(l \log l) \).

\( \text{CalSimilarity}(f, d_i, \text{minSimilarity}) \) requires the comparison of \( n \) datasets using the RWLIP
algorithm and thus has a time complexity of \( O(n \times l \log l) \).

### 6.1.4 Optimal Spatially Aware Ordering

To further investigate the effect of the four linear spatial ordering methods, we generate all
possible orderings for a given base map and find the optimal solution that minimises the dis-
similarity score for a given control dataset. This enables us to compare the optimal linear
ordering to the orders generated by our techniques.
We select a small subset of eight regions from the Brisbane base map as the study area. For this study area there are $8!$ possible spatial orders. We generate all possible orders to determine the optimal spatial order that will minimise the dissimilarity score for the datasets shown in Fig. 6.6(a-d). We assert that dataset6.2a is similar to dataset6.2b as they show a similar high density distribution in the three top left regions and low density distribution in the remaining regions. It can also be seen that dataset6.2a is very dissimilar to dataset6.2c−d. Figure 6.7(a-d) show the GLS, DFS, BFS and NN-5 orders for the study region respectively. To enable comparison we use the same starting region for each ordering. Constraining the optimal ordering to the same starting region as our three techniques reduces the number of possible orders to $7!$.

Figure 6.8 shows a comparison of the results obtained by the density tracing algorithm for each spatial ordering. The dissimilarity values are normalised using the min-max technique described in Section 6.1.2. Both GLS, NN-5 and DFS are good approximations of the optimal ordering. Figure 6.7(e) shows that the optimal order is very similar to the orders generated using the three other techniques with the same starting region. In particular, the GLS order is most similar. It is noted that an optimal order is not needed, only one that can model spatial
6.1 Density Tracing

Figure 6.7: Optimal ordering: (a) GLS; (b) DFS; (c) BFS; (d) NN-5; (e) Optimal order with same starting region.

Figure 6.8: Comparison of normalised density dissimilarity of synthetic datasets using various orders.
neighbourhood information so that the spatial distribution of datasets can be compared. From the comparison of this small subset and visually from Fig. 6.3 it can be seen that GLS is suited to capture spatial information for these study regions.

6.1.5 Validation with Synthetic Datasets

The examples in Section 6.1.1 are illustrative experiments that are designed to explain our technique of using Density Tracing to discover cross-distribution relationships between datasets. We now conduct a number of experiments with synthetic datasets to evaluate and justify this approach. The synthetic datasets are produced using a MATLAB program which generates random point data drawn from a mixture of multivariate gaussians. This point set data is then converted to an areal aggregated dataset by assigning points to regions of the given base map. To aid readability the figures in this section display the point data overlaid onto the base map instead of the density aggregates.

We first evaluate the approach with synthetic datasets that show increasing dissimilarity. Figure 6.9(a-d) show the synthetic datasets used in this first experiment. We start with $n = 4$ clusters in dataset6.3a, and for each subsequent dataset remove $k$ clusters ($0 < k < n$). We assert that the linear decrease in the number of clusters results in these datasets showing increasing dissimilarity, that is, as clusters are removed, dataset6.3b−d become less similar to the original dataset6.3a. Figure 6.9(e) illustrates this; when dataset6.3a is selected as the reference feature the dissimilarity increases with the number of clusters removed. We repeat this experiment on a larger second synthetic dataset. Figure 6.10(a) shows the initial dataset6.4a with $n = 21$ clusters, and again for each subsequent dataset $k$ clusters are removed (we do not show figures for dataset6.4b−u). As clusters are removed from dataset6.4a we assert that subsequent datasets will become less similar. Figure 6.10(b) shows that for these synthetic datasets all four region orderings identify the increasing dissimilarity as clusters are removed from the datasets.

The second experiment we present is a modification of the first removing clusters synthetic experiment. Figure 6.11 starts with $n = 4$ clusters in disjoint regions of the feature space
and for each subsequent dataset the $k^{th}$ cluster is moved to a new location within the space. The main difference between this and the previous experiment is that $n$ does not vary between datasets. We assert that the linear decrease of clusters in their original locations will result in an increase in dissimilarity as clusters are moved. Figure 6.11(f) shows that when $\text{dataset}6.5_a$ is selected as the reference feature, only the GLS and NN-5 orders give this expected result.
for all datasets, that is, they show an increase in dissimilarity. When \( k = 2 \) (\textit{dataset6.5}_c) the DFS order records a peak in dissimilarity, calculating that \textit{dataset6.5}_d and \textit{dataset6.5}_e are more similar to the reference feature than \textit{dataset6.5}_c. A similar result can be seen for the BFS ordering when \( k = 3 \). This experiment is again repeated on a larger synthetic dataset with \( n = 21 \) clusters. Figure 6.12(a) shows the the initial \textit{dataset6.6}_a while Fig. 6.12(b) depicts the finial \textit{dataset6.4}_a (we do not show figures for \textit{dataset6.4}_{b−t}). For this larger dataset, Fig. 6.12(c) shows that all four region orderings identify the increasing dissimilarity as clusters are moved within the dataset.
FIGURE 6.11: Synthetic experiment with moving clusters: (a-e) Datasets dataset6.5a–e; (f) Increasing dissimilarity with respect to number of moved clusters.
To measure similarity between datasets we compare the spatial distribution, that is, how the density changes between regions. For these synthetic datasets, the GLS and NN-5 spatial ordering are robustly able to preserve the spatial neighbourhood information, as reflected in the fact that the DFS and BFS orders give unexpected results for part of this experiment. We use these synthetic datasets to evaluate and justify our approach as there is no baseline measure that can be used to measure similarity, we rely on the linear decrease in the number of clusters to confirm that these datasets show increasing dissimilarity.

In Chapter 8 we demonstrate how Density Tracing can be used for crime analysis and also compare its computational performance against other approaches for discovering cross-patterning discussed in this thesis.
6.2 Graph Mining

The second technique we propose to capture cross-distribution relationships models areal aggregated datasets as graphs which store the geospatial distribution of crime within given regions. These graphs can be used to discover datasets that show similar geospatial distribution characteristics. By utilising a graph representation we are able to efficiently and effectively discover cross-distribution patterns.

A summary of the technique is as follows; we first model the complex areal aggregated datasets as weighted directed multigraphs and then prune the graphs to leave only the most influential distribution changes. We investigate two techniques to achieve this: using a global minimum edge weight, and using the Minimum Spanning Tree (MST) of each graph. Once the graphs have been pruned, cross-distribution patterns can then be discovered by finding graphs that exhibit similar geospatial distribution changes. This is achieved by extracting each vertices degree and then computing a correlation coefficient such as Pearson’s $r$ or Jaccard’s Index $J$. The resulting patterns depict datasets that show a similar distribution in specific spatial regions.

6.2.1 Graph Based Representation

Given $n$ crime, census and feature datasets or layers, we construct $n$ weighted directed multigraphs. The vertices of the graph represent regions in the base map, and there is an edge between two vertices if the regions share a border to capture spatial neighbourhood.

Each directed edge $e$ has a weight defined as $w(e) = influence(e) \ast changeInDensity(e)$ where $influence(e)$ is the length of the shared border between source and target vertex divided by the total border length of the source vertex and $changeInDensity(e)$ is the change in density between the source and target vertex. We use a weighted directed multigraph as the edge weights are asymmetric and thus each region will have two edges for each neighbouring region.

We are analysing areal aggregated spatial datasets recorded for a particular study region, and thus all datasets share a common base map. To start with, each graph will have the same number
of vertices and edges. Figure 6.13 depicts the base map that is used for these experiments, the 216 urban suburbs of Brisbane, the capital city of Queensland, Australia. Each graph for this base map starts with 216 vertices and 1224 edges.

### 6.2.2 Working Principle

The working principle of the graph mining framework is explained with the example synthetic datasets shown in Fig. 6.14. To generate the datasets we start with 4 clusters in disjoint regions of the feature space (dataset6.7a) and for each subsequent dataset the $k^{th}$ cluster is moved to a random location within the space (dataset6.7b–e). The number of clusters does not vary between datasets and we assert that the linear decrease of clusters in their original locations will result in the datasets becoming less similar to the original dataset6.7a as clusters are moved.

The synthetic datasets are produced using a MATLAB program which generates random point data drawn from a mixture of multivariate guassians. We convert this point set to an areal aggregated dataset by assigning points to regions of the given base map. For the figures in this section we display the point data overlaid onto the base map instead of the density aggregates.
to aid readability. We generate ten sets of *dataset6.7_\text{a–e}* and take the average for results shown in this section.

The framework works as follows: Initially the density values are normalised into the range [0, 1] using the *min-max normalisation* [77] technique which retains the original distribution of scores except for a scaling factor. For each dataset a weighted directed multigraph is
then generated that stores density values for each region, and also that regions influence on neighbouring regions.

Each graph is pruned by extracting the most influential edges by either using a global minimum edge weight or by finding the MST of the graph using Kruskal’s algorithm [49]. The MST approach finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimised. The minimum edge weight approach attempts to achieve a similar tree using a user supplied parameter. This shortest path represents the path of most similar density change within the graph, while respecting the influence of neighbours. A small edge weight represents similar density change.

The set of pruned graphs can then be compared to discover datasets that show similar cross-distribution. For each pruned graph we list each vertices degree (edges coming into or out from the vertex) and then either use Pearson’s correlation coefficient or Jaccard’s Index to find correlation between the degree lists. This quantifies the correlation between two datasets cross-distribution. When we refer to a dataset’s correlation further in this chapter, we are referring specifically to the correlation of cross-distribution. Pearson’s correlation coefficient $r$ for variables $X$ and $Y$ gives a value between +1 and -1 inclusive. A value of 1 indicates a perfect positive correlation between $X$ and $Y$ while -1 indicates a perfect negative correlation. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J_{X,Y} = \frac{|X \cap Y|}{|X \cup Y|}.$$

We allow the user to restrict the resulting cross-distribution patterns to the top-$k$ and/or bottom-$k$ correlated datasets.

Figure 6.15 depicts the MSTs generated from the synthetic dataset in Fig. 6.14. Each MST includes every vertex where the total weight of all the edges in the tree is minimised. As compared with Fig. 6.13 we can see that a number of edges have been removed to form the MST of each dataset. The expected correlation results from the synthetic dataset dataset6.7
FIGURE 6.15: MSTs of synthetic dataset: (a-e) MST of Datasets \textit{dataset6.7}_{a–e} from Fig. 6.14
are that the correlation between datasets will decrease as clusters are moved. Figure 6.16 shows the correlations from the ten sets of \textit{dataset6.7} using a minimum edge weight of 0.025 and Pearson’s $r$. The average result from the ten sets is shown with a thick red line. This approach achieves the expected result but has the drawback of requiring a user supplied argument. We experimentally determined that for this base map a minimum edge weight of 0.025 was suitable, however this value is highly dependent on the particular base map used. Figure 6.17 shows
FIGURE 6.18: Correlations from synthetic dataset dataset6.7 using MST and Jaccard’s $J$

FIGURE 6.19: Synthetic experiment with moving clusters: (a) Initial dataset dataset6.8a; (b) Final dataset dataset6.8u; (c) Correlations using MST.
the correlations using MST and Pearson’s $r$. This approach also shows the expected results,
and has the advantage of being fully automatic given any base map. Figure 6.18 shows that
the MST approach with Jaccard’s $J$ index also returns the expected result. We repeat this
experiment using MST to further compare Pearson’s $r$ and Jaccard’s $J$ with a dataset of $n = 21$
clusters. Figure 6.19(a) shows the the initial dataset $6.8_a$ while Fig. 6.19(b) depicts the finial
dataset $6.8_u$ (we do not show figures for dataset $6.8_{b-t}$). Figure 6.19(c) shows that for these
synthetic datasets only Pearson’s $r$ give the expected result for all datasets, that is, a decrease
in the correlation measure as clusters are moved. For the patterns $a − n$ and $a − r$ an increase
is correlation is recorded by Jaccard’s J index.

6.2.3 Algorithm

Algorithm 6.2.4 Graph Mining for cross-distribution patterns

**Input:** A set $D = \{d_1, d_2, \ldots, d_n\}$ of spatial areal aggregated datasets, a common base map
$B = \{b_1, b_2, \ldots, b_l\}$, the chosen pruning method `pruneMethod`, the chosen correlation mea-
ure `corrMeasure` and the desired number of outputs $k$;

**Output:** Pairs of datasets in the top-$k$ and bottom-$k$ as ranked by the correlation measure `corrMeasure`;

1: `list < graph > listG`
2: `LoadDatasets(D);`
3:
4: `NormalizeDensity(D);`
5: **for all** $d_i \in D$ **do**
6: `upperLimit ← 1`
7: `lowerLimit ← 0`
8: `Hi ← d_i.max()`
9: `Lo ← d_i.min()`
10: `fact ← (upperLimit − lowerLimit)/(Hi − Lo)`
11: **for all** $x \in d_i$ **do**
6.2 Graph Mining

12: \[ d_i[x] \leftarrow (d_i[x] - Lo) \times fact + lowerLimit \]
13: \textbf{end for}
14: \textbf{end for}
15:
16: GenerateGraph:
17: \textit{graphG}
18: \textbf{for} each region \( r \) in \( B \) \textbf{do}
19: \( cp \leftarrow \text{centerPointOfBoundingBox}(r) \)
20: \textbf{if} \( cp \) is not within \( \text{BoundingBox}(r) \) \textbf{then}
21: \textbf{point1}: Move \( cp \) in the X direction until it enters into \( r \)
22: \textbf{point2}: Move \( cp \) in the X direction until it exits \( r \)
23: Centroid \( cp \) is calculated to be halfway between \textbf{point1} and \textbf{point2} on the same X axis
24: \textbf{if} More than one pair of polygon outlines cross the X-axis \textbf{then}
25: Compare each pair of outlines to see which pair creates the widest length along the X-axis
26: Calculate centroid \( cp \) as halfway between the points where this pair of outlines crosses the X-axis
27: \textbf{end if}
28: \textbf{end if}
29: \( G.add(newVertex(cp)) \)
30: \textbf{end for}
31: \textbf{for} each region \( r \) in \( B \) \textbf{do}
32: \textbf{for} Each neighbouring region \( n \) of \( r \) \textbf{do}
33: \( influence \leftarrow \text{sharedBorder}(r, n)/\text{border}(r) \)
34: \( G.add(newEdge(vertexOf(r), vertexOf(n), influence)) \)
35: \textbf{end for}
36: \textbf{end for}
37: \textbf{for all} \( d_i \in D \) \textbf{do}
38: \textit{graph localG} \leftarrow \textit{G}
39: \textbf{for all} \( edge \in \textit{localG} \) \textbf{do}
6.2 Graph Mining

40: \[ \text{densityOfSource} \leftarrow d[\text{edge.sourceVertex}] \]
41: \[ \text{densityOfTarget} \leftarrow d[\text{edge.targetVertex}] \]
42: \[ \text{changeInDensity} \leftarrow \text{densityOfSource} - \text{densityOfTarget} \]
43: \[ \text{edge.weight} \leftarrow \text{edge.weigh} \ast \text{changeInDensity} \]
44: \[ \text{end for} \]
45: \[ \text{listG.add(localG)} \]
46: \[ \text{end for} \]
47: 

48: PruneGraphs:
49: \[ \text{for all graph } G \in \text{listG do} \]
50: \[ \text{if pruneMethod = MinEdgeWeight then} \]
51: \[ \text{Require: User supplied minimum edge weight, default: } minEW \leftarrow 0.025 \]
52: \[ \text{graph } \text{originalG} \leftarrow G \]
53: \[ \text{for all } \text{edge} \in \text{originalG do} \]
54: \[ \text{if } \text{edge.weight} < minEW \text{ then} \]
55: \[ G.\text{remove} \leftarrow \text{edge} \]
56: \[ \text{end if} \]
57: \[ \text{end for} \]
58: \[ \text{else if pruneMethod = MST then} \]
59: \[ \text{list } < \text{edge} > T \]
60: \[ \text{priorityQueue } < \text{double, edge} > \text{ PQ} \]
61: \[ \text{for all } \text{edge} \in G \text{ do} \]
62: \[ \text{PQ.insert}(\text{edge.weight}, \text{edge}) \]
63: \[ \text{end for} \]
64: \[ \text{nodePartitionP(G)} \]
65: \[ \text{while !PQ.empty do} \]
66: \[ e \leftarrow \text{PQ.inf}(\text{PQ.findMin}) \]
67: \[ \text{PQ.delMin} \]
68: \[ \text{nodeu } \leftarrow \text{G.source(e)} \]
69: \[ \text{nodev } \leftarrow \text{G.target(e)} \]
6.2 Graph Mining

70: \textbf{if } \neg П.\textit{sameBlock}(u, v) \textbf{ then}
71: \hspace{1em} T.append(e)
72: \hspace{1em} П.unionBlocks(u, v)
73: \hspace{1em} \textbf{end if}
74: \hspace{1em} \textbf{end while}
75: \textbf{Delete all edges in } G \textbf{ that are not in } T
76: \textbf{end if}
77: \textbf{end for}
78: 
79: \textbf{CompareGraph:}
80: list < listint > nodeDegreeList
81: \textbf{for all graph } G \in \textit{listG} \textbf{ do}
82: \hspace{1em} \text{int nodeDegree}
83: \hspace{1em} \textbf{for all nodes } n \in \textit{listG} \textbf{ do}
84: \hspace{2em} \textbf{for all edges } e \in \textit{listG} \textbf{ do}
85: \hspace{3em} \textbf{if } G.source(e) = n \text{\textit{OR}}.target(e) = n \textbf{ then}
86: \hspace{4em} \text{nodeDegree } \gets \text{nodeDegree } + 1
87: \hspace{3em} \textbf{end if}
88: \hspace{2em} \textbf{end for}
89: \hspace{1em} \textit{nodeDegreeList}.append(nodeDegree)
90: \hspace{1em} \textbf{end for}
91: \textbf{end for}
92: \textbf{if } \textit{corrMeasure} = \textit{Pearson} \textbf{ then}
93: \hspace{1em} \text{Calculate mean for each nodeDegree list}
94: \hspace{1em} \textbf{while } \neg \textit{nodeDegreeList}.empty() \textbf{ do}
95: \hspace{2em} x \gets \textit{nodeDegreeList}.pop()
96: \hspace{2em} xMean \gets \textit{nodeDegreeMeanList}.pop()
97: \hspace{2em} \textbf{for all listy } \in \textit{nodeDegreeList} \textbf{ do}
98: \hspace{3em} yMean \gets \textit{nodeDegreeMeanList}.getItem(y)
99: \hspace{3em} \text{Calculate Pearson's } r \gets \textit{Pearsons}(x, y, xMean, yMean)
6.2.4 Validation with Synthetic Datasets

We evaluate the graph mining approach with synthetic datasets that show increasing dissimilarity. The datasets shown in Fig. 6.20 are a modification of the synthetic experiment detailed in Section 6.2.2. We start with $n = 4$ clusters in dataset6.9a, and for each subsequent dataset $k$ clusters ($0 < k < n$) are removed. We assert that the linear decrease in the number of clusters results in these datasets showing increasing dissimilarity, that is, as clusters are removed, dataset6.9b–d become less similar to the original dataset6.9a. Ten sets are generated to verify the robustness of the approach.

The expected result is that the correlation between dataset6.9a and the other datasets decreases with the number of clusters removed. Figure 6.21-6.23 shows that all three approaches return...
**FIGURE 6.20:** Synthetic dataset removing clusters: (a-d) Datasets \textit{dataset6.9}_{a-d}

**FIGURE 6.21:** Correlations from synthetic dataset \textit{dataset6.9} using minimum edge weight and Pearson’s \( r \)
6.2 GRAPH MINING

**Figure 6.22:** Correlations from synthetic dataset dataset6.9 using MST and Pearson’s $r$

**Figure 6.23:** Correlations from synthetic dataset dataset6.9 using MST and Jaccard’s $J$

the expected correlative patterns. We again repeat this experiment on a larger synthetic dataset to verify the robustness of this approach. Figure 6.24(a) shows the initial dataset6.10$_a$ with $n = 21$ clusters, and for each subsequent dataset $k$ clusters are removed (we do not show figures for dataset6.10$_{b-u}$). Figure 6.24(b) shows that for these synthetic datasets Pearson’s $r$ is most able to model the expected result, that is as clusters are removed from datasets dataset6.10$_{b-u}$ they become less correlated to dataset6.10$_a$. Due to the drawbacks outlined in Section 6.2.2
for the minimum edge weight approach, we recommend the MST approach with Pearson’s $r$ as the suggested setting for this approach.

In Chapter 8 we show how the graph mining approach can be used for crime analysis and also compare the computation aspect against other approaches for discovering cross-patterning discussed in this thesis.
6.3 Summary

The previous chapters have discussed two ways in which cross-patterning can be modelled, cross-association and cross-varying. Cross-association models the relationship across multiple spatial layers for each region, that is, it does not consider the influence of neighbouring regions in its calculation. Cross-varying models the relationship across multiple spatial layers while taking into consideration the effect of all local neighbouring regions.

In this chapter we have introduced a third way in which cross-patterning can be modelled, cross-distribution. Cross-distribution uses the spatial distribution of a dataset to model patterns over multiple layers and can consider the spatial distribution of both one local neighbour (i.e. one neighbour that shares a border with the region) as well as the global distribution (modelling the whole study region).

We present two new approaches that can discover these cross-distribution relationships over multiple layers or datasets. In the first part of this chapter we introduce the first of these approaches. Density Tracing models the spatial distribution of an areal aggregated dataset as a density trace and then computes the similarity between these traces to discover those that have a cross-distribution relationship. To model the dataset as a density trace we must first define a spatial ordering of regions. The density traces can then be constructed by projecting them on the Cartesian plane following the defined spatial ordering of the regions. We investigate four popular locational ordering methods to achieve this and show that in general the GLS technique is the most suited method to capture neighbours. We then demonstrate how to calculate the distance (dissimilarity) between two traces using the RWLIP algorithm. This dissimilarity score can then be used to rank spatial layers on the strength of their cross-distribution relationships.

One strength of using Density Tracing is that we can include clustering results into the calculation of trace similarity. The RWLIP distance measure that we introduce is able to use assigned region weights to assign those regions more importance during the distance calculation. For example, we may have knowledge (gained from either a domain expert or the use of a clustering algorithm) that 4 regions surrounding the CBD of a city show a hot spot for a particular
dataset. We can then use this information when calculating cross-distribution relationships. We demonstrate an example of including clustering results into Density Tracing in Chapter 8.

The second approach we introduce to discover cross-distribution relationships is Graph Mining. This technique models the spatial distribution of areal aggregated spatial datasets as weighted directed multigraphs. These graphs store the density values for each region, and also each region’s influence on neighbouring regions. We prune each of these graphs by extracting the most influential edges. A small edge weight represents similar density change (while taking into consideration the how much two regions share a border) and thus we can calculate the shortest path which represents the path of most similar density change within the graph. We investigate two approaches to calculate this shortest path: using a global minimum edge weight, and using the MST of each graph. We show that both approaches are suitable however we recommend MST as it is argument free. These pruned graphs are then compared by extracting each vertices degree to compute a correlation coefficient such as Pearson’s $r$ or Jaccard’s Index $J$ so that cross-distribution patterns can be discovered.

In Chapter 8 we show how these two approaches can be used to discover cross-distribution relationships for crime analysis and we also compare the computation aspect against each other and also the other techniques presented in this thesis for discovering cross-patterning.
The use of visualisations can improve our ability to understand and interpret data that would not be easily understood in its plain textual or numeric form. Using a visualisation tool can make interpretation much easier and faster for users. It is also able to capture and convey more complex relationships and information that would be difficult to convey in text or other forms.

A variety of factors can contribute to the formulation of crime, so to provide a greater level of insight into the complex nature of criminal behaviour a wide variety of datasets should be studied. As the number of datasets under analysis increases, so does the possible number of discovered patterns. The problem can become one of information overload; how can the user find interesting cross-patterning relationships amongst other patterns. The definition of interesting is often user specific, so a visualisation environment should allow the user to dynamically choose and refine patterns that are of interest to them.

The visualisation environment developed in this chapter serves two purposes. They are to:

- Allow for the combination of results from the cross-association, cross-varying and cross-distribution algorithms to highlight those patterns that may overlap with each other, indicating a more interesting cross-patterning; and,
- Allow the end user to easily access and interpret the cross-patterning relationships extracted from the datasets.

The visualisation environment we develop is a graph based system that allows the user to dynamically view and interact with the result set. Each vertex (feature or dataset) is labelled and its size is a depiction of the cross-patterning strength, where the size of the circle represents
the largest strength of its relationships. This strength could be the support and confidence of cross-association, the $L$ index correlation from our cross-varying approach or the dissimilarly of Density Tracing. The edges between vertices in the visualisation depict the patterns. The strength is depicted by the edge thickness and colour.

In the next section we first describe how the visualisation environment can be used to interact with a single result set (such as the results from discovering cross-distribution using the Density Tracing approach). This idea is then extended to show how multiple result sets can be visualised to highlight overlapping patterns that may suggest stronger cross-patterning relationships.

### 7.1 Single Result Set Visualisation

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B</td>
<td>10</td>
</tr>
<tr>
<td>A-C</td>
<td>5</td>
</tr>
<tr>
<td>A-D</td>
<td>1</td>
</tr>
<tr>
<td>A-G</td>
<td>8</td>
</tr>
<tr>
<td>B-D</td>
<td>5</td>
</tr>
<tr>
<td>B-E</td>
<td>5</td>
</tr>
<tr>
<td>D-E</td>
<td>10</td>
</tr>
<tr>
<td>E-F</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 7.1:** Synthetic result set 1.

We describe the visualisation environment using a synthetic result set that consists of seven datasets labelled A-G with cross-patterning described in Table 7.1. This synthetic result set simulates the output of our cross-patterning algorithms described in previous chapters. Figure 7.1 shows the initial random visualisation presented to the user. Each vertex represents a dataset that has a cross-patterning relationship in the result set. The size of each vertex is a depiction of the cross-patterning strength, where the size of the circle represents the largest strength of its relationships. For example, within the result set (Table 7.1) we can see that datasets A, B, D, E and F all have patterns with a strength of 10, hence these vertices are shown larger than those for datasets C and G. The edges between vertices in Figure 7.1 depict the patterns where the strength is depicted by the edge thickness and colour. Patterns are grouped into percentage ranges based their strength and this is used to assign the colour and weight to
7.1 SINGLE RESULT SET VISUALISATION

FIGURE 7.1: Visualisation of synthetic result set 1 - random layout.

FIGURE 7.2: Visualisation of synthetic result set 1 - circular layout.
edges. The percentage range used is automatically determined from the range of the result set. For example, from the result set shown in Table 7.1, the patterns A-B, D-E and E-F share the same edge colour (pink) and thickness (thick) as they fall within the same percentage range (in this case they also share the same pattern strength). The pattern A-G is in the middle range of strength values and is assigned another edge colour (blue) and thickness (medium). The remaining patterns are grouped into the low range of values and are assigned another edge colour (green) and thickness (thin). The user can right click on any feature to get detailed information for the pattern.

![Graph Visualization](image)

**FIGURE 7.3:** Visualisation of synthetic result set 1 - 2D spring layout.

The initial visualisation layout may not allow the user to easily visualise the cross-patterning information. In this case the user can either manually move or delete vertices that are not of interest, or they can select one of the built in layout functions. Figure 7.2 depicts the circular layout where each vertex is arranged to form a circle. This circular layout makes it easy to find particular features and to quickly identify those features with a large number of connections, however with a large number of datasets this layout is often undesirable as it can be hard to see individual patterns (such as those presented in Section 8.1.4).
7.1 Single Result Set Visualisation

Figures 7.3 and 7.4 show the 2D and 3D spring layout respectively. The spring layout simulates the graph as a system of mass particles where the vertices are the mass particles and the edges are springs between the particles [60]. The algorithms try to minimise the energy of this physical system so that a state of equilibration is reached. That is, the spring layout is a force directed layout that models the input graph as a system of forces and tries to find the minimum energy configuration of the system. The 3D layout screen capture is shown for demonstration, in the actual visualisation environment the user can spin, zoom and rotate the 3D display to better interact with the result set. Within the 3D layout the user can select vertices or edges to see specific information. We utilise the Library of Efficient Data types and Algorithms (LEDA) to implement these layouts.
7.2 Cross-Patterning Visualisation

We extend the visualisation described in the previous section to allow for the visualisation of multiple result sets so as to highlight overlapping patterns that may suggest stronger cross-patterning relationships. Table 7.2 and Figures 7.5 and 7.6 detail the second synthetic result set that we wish to visualise together with the result set from Table 7.1.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B</td>
<td>5</td>
</tr>
<tr>
<td>B-C</td>
<td>3</td>
</tr>
<tr>
<td>B-E</td>
<td>10</td>
</tr>
<tr>
<td>B-H</td>
<td>7</td>
</tr>
<tr>
<td>F-G</td>
<td>9</td>
</tr>
<tr>
<td>F-I</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 7.2:** Synthetic result set 2.

**Figure 7.5:** Visualisation of synthetic result set 2 - random layout.

Figure 7.7 shows the initial visualisation of the multiple result sets. The vertex and edge sizing are the same as for the previous visualisation described in Section 7.1, except that edges that are common across multiple result sets are given more weight and are always shown in gray. A
**Figure 7.6:** Visualisation of synthetic result set 2 - 2D spring layout.

**Figure 7.7:** Visualisation of multiple result sets - random layout.
7.2 Cross-Patterning Visualisation

**Figure 7.8:** Visualisation of multiple result sets - circular layout.

**Figure 7.9:** Visualisation of multiple result sets - 2D spring layout.
pattern is determined to be overlapping if the source vertex and target vertex of an edge match a pattern from a different result set. In this case we can easily see that the patterns involving A-B and B-E appear in both the result sets. Figure 7.8 shows the circular layout option, while Figures 7.9 and 7.10 depicts the 2D and 3D spring layouts.

**Figure 7.10**: Visualisation of multiple result sets - 3D spring layout.

### 7.3 Summary

The visualisation environment developed in this chapter can help end users to discover *interesting* cross-patterning relationships that may be hidden amongst other patterns. The user can dynamically interact with the result sets to highlight patterns or datasets that are of interest to them. The visualisation environment can also be used in an exploratory sense to highlight previously unknown groupings of cross-patterning relationships.

The result sets used in the previous two sections are designed to highlight how our visualisation technique works. They are not typical of real work cross-patterning results where many hundreds of relationships may be discovered among many spatial layers. Table 7.3 depicts a
synthetic result set comprised on 26 datasets (the letters A-Z) and 159 cross-patterning relationships. Given the data in this raw form, it is hard for the end user to fully comprehend the relationships between datasets. Figure 7.11 shows the visualisation of this data using the circle layout. With the large number of datasets and patterns, we can see that the relationships overlap each other around the edge of the circle making this layout not particularly useful for this dataset. The 2D spring layout shown in Fig. 7.12 provides a more useful visualisation.
For example, it is clear from this view that while dataset C has a number of cross-patterning relationships, they are of lower strength than those of involving dataset V.

**Figure 7.11:** Visualisation of synthetic result set 3 - circular layout.

The visualisation technique can display several degrees of relationships, and multiple relationships on the one screen. It can also visually display the strength of the relationships through varying mechanisms such as the thickness of an edge or line between two graphically represented features. It is then very quick to see the complex network of relationships between multiple entities and notice different trends. Many of these would be very difficult to identify in their plain text form.

Crime data analysis is the first step in finding interesting facts about criminal behaviour. Visualisation is the next step, allowing domain experts to interpret a large amount of this valuable mined information which may be otherwise too difficult to absorb. We present a case study of crime data analysis that highlights the usefulness of this visualisation in the following chapter.
FIGURE 7.12: Visualisation of synthetic result set 3 - 2D spring layout.
CHAPTER 8

Case Study: Crime Data Analysis

This chapter provides cross-patterning results from the crime data analysis of real crime data provided by the QPS. We analyse this data along with socio-economic, socio-demographic and geospatial features as described in Chapter 3. The aim of this case study is to provide better insight into why certain crimes may happen in certain locations.

8.1 Cross-Patterning Relationships

We first separately analyse the crime data for cross-association, cross-varying and cross-distribution patterns using the respective algorithms outlined in Chapters 4 - 6. We then show how the visualisation environment can be used to combine these results so that interesting cross-patterning relationships are highlighted.

8.1.1 Cross-Association Patterns

One drawback of mining cross-association patterns is that typically a lot of uninteresting frequent patterns are discovered. To alleviate this we perform a top down analysis by exploiting the hierarchical nature of the crime data as shown in the crime tree in Fig. 3.3. We examine different levels of the hierarchy in our experiments starting with the top-level (crimes abbreviated as: TOAPE, TOAPR and TOO), then the first-level and then the leaves of the crime tree. A description of the crime abbreviations from the dataset used in this case study can be found in Table 3.2.
As discussed in Section 4.1.2, the default method of ArcGIS, natural breaks, has been used as the categorisation method in our ACMAM algorithm for this analysis. The number of categories play a critical role, typically, 4-6 categories are recommended [25]. Too few categories may cause loss of detail while too many categories can cause confusion and excess uninteresting patterns. In this analysis, we experiment with 5 groupings for quantitative associations and 2 groupings for boolean associations. However, users can explore a given dataset with different categorisation schemes with different number of classes in my framework. All the experiments are carried out using minimum support = 0.05% and minimum confidence = 0.7%. We select rules with a lift value greater than 1 which indicates that the antecedent and the consequent appear more often together than expected, meaning that the occurrence of the antecedent has a positive effect on the occurrence of the consequent.

We visualise the cross-association patterns discovered by the ACMAM approach in two ways. First we visualise a subset of the discovered association rules using the 3D representation suggested by [92] where the rows of the matrix floor represent items and the columns represent item associations. The green and red blocks of each column (rule) represent the antecedent and the consequent. Identities of the items are shown along the right side of the matrix. In the 3D representation, the blue and the cyan represent the support and the confidence. The second visualisation is the technique presented in Chapter 7.

8.1.1 Boolean ACMAM

This analysis focuses on crime hot spots and densely featured areas, and reveals positive boolean multivariate cross-associations. Figure 8.1 and Fig. 8.2 show the boolean cross-associations that were discovered with ACMAM. Rules 1 - 7 are from the analysis of top-level crime types, rules 8 - 15 from first-level and rules 16 - 21 from leaf-level crime types. Figures 8.3- 8.5 depict the boolean cross-associations using the visualisation described in Chapter 7.

**Top-level Association.** This analysis considers the three top-level crime types with the geospatial features. Interestingly, Schools and Highways appear frequently with TOAPR. In particular, Schools or Highways can be found where TOAPR takes place with around 80% confidence.
FIGURE 8.1: Boolean ACMAM with support and confidence values.

FIGURE 8.2: Boolean ACMAM with support and lift values.
8.1 Cross-Patterning Relationships

**Figure 8.3:** Visualisation of Boolean ACMAM - Random layout.

**Figure 8.4:** Visualisation of Boolean ACMAM - Circle layout.
Highways seem to be frequently found with Schools, Railways, and Localities. This is an example of an uninteresting frequent cross-association pattern that often produces an overwhelming number of results. In addition, this top-level analysis reveals some asymmetric associations among geospatial features. Schools, Localities or Reserves seem to imply the existence of Highways with more than 74% of confidence. In addition, the constrained rule mining \((\ast \Rightarrow \text{TOAPR})\) reveals two combinations of features imply TOAPR with 60% of confidence level. Some other interesting frequent patterns and multivariate cross-associations are shown in Figure 8.1 and Fig. 8.2.

**First-level Association.** Cross-association patterns revealed in this analysis demonstrate which sub-categories of TOAPR are particularly associated with Schools and Highways. The analysis reveals TOUE exhibits high cross-association with Highways, Schools and Railways. Also, it reveals that even though TOOO does not show high cross-association with Highways, one of its sub-categories, TARO frequently exists with Highways with 14% of support. Multivariate cross-associations with high confidence reveals OTPD and MOVT are “possible lures” of TOUE.
Leaf-level Association. This analysis reveals cross-associations with more detailed (low-level) crime types and concretes frequent patterns found in the first-level association analysis. Leaf-level cross-association pattern mining reveals that UEWID is the main contributor to the high cross-association of TOUE with Highways/Schools. In addition, this analysis reveals cross-association of UEWIS with Highways. UEWID, UEWIS, or MOVT as consequent in the constrained mining (* ⇒ Leaf-level crimes) reveals several “possible lures” that attract these crime types. In particular, Schools attract UEWID and MOVT. This information could be used by residents nearby schools to better protect their properties.

8.1.1.2 Quantitative ACMAM

Quantitative ACMAM reveals cross-associations with discretised intervals. As discussed, natural breaks with 5 classes (very dense, dense, medium, sparse, very sparse) has been used in this experiment. Figure 8.6 and Fig. 8.7 show the quantitative cross-associations that were discovered with ACMAM. Rules 1 - 7 are from the analysis of top-level crime types, rules 8 - 13 from first-level and rules 14 - 18 from leaf-level crime types. Figures 8.8- 8.10 depict the quantitative associations using the visualisation described in Chapter 7.

Figure 8.6: Quantitative ACMAM with support and confidence.
8.1 Cross-Patterning Relationships

**Figure 8.7:** Quantitative ACMAM with support and lift.

**Figure 8.8:** Visualisation of Quantitative ACMAM - Random layout.
8.1 Cross-Patterning Relationships

**Top-level Association.** Even though Schools and TOAPR are frequently located, Quantitative ACMAM reveals Schools are not associated with any particular naturally broken TOAPR groups. However, Highways frequently co-exist with one of TOAPR groups (TOAPR(sparse)). Interestingly, Highways co-exist with sparsely populated top-level crime types. The cross-association of Highways and densely populated top-level crime types is not highly supported. Some interesting frequent patterns and multivariate quantitative cross-associations are shown in Figure 8.6 and Fig. 8.7.

**First-level Association.** Some minor crime types, not classified into major crime types but others, are found to be frequent in this analysis, in particular, MIOF(sparse), OTTH(sparse) and OTPD(sparse). ACMAM reveals these sparsely populated minor crime types imply Highways with more than 74% confidence.

**Leaf-level Association.** Similarly, this association analysis concretes some high-level cross-associations and detects new low-level crime cross-associations. Highways seem to co-exist with OTST(sparse) that confirms the cross-association of OTTH(sparse) and Highways. On
the other hand, the cross-association of Highways and RAAR(sparse) has been detected in this association test.

The number of frequent itemsets generated and the number of association rules discovered are dependent on both the minimum support and minimum confidence parameters. Figure 8.11 shows how varying these two parameters effects the number of frequent itemsets and association rules as applied to boolean ACMAM using leaf crime data. It can be seen that with even small changes to these parameters that the number of generated rules can become unworkable. Figure 8.12 shows a similar effect when using quantitative ACMAM.

8.1.2 Cross-Varying Patterns

To discover cross-varying patterns we utilise Lee’s $L$ index with the shared border length weights matrix as described in Chapter 5. These datasets are mined to extract the top-$k$ and bottom-$k$ spatial cross-varying patterns involving salient features or census classifications (for this experiment $k = 5$). We focus on the 39 crime types that form the leaves of the crime tree.
8.1 CROSS-PATTERNING RELATIONSHIPS

(Figure 3.3) and restrict the discovered patterns to those involving at least one type of crime. From the extracted patterns shown in Table 8.1 we can see that there is a strong pattern 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 pattern between *Arson* and *Persons aged 85-94*. This information can then be used to further investigate the cause of these specific cross-varying patterns and also allow for targeted crime prevention efforts.

Figure 8.13 shows the visualisation that depicts the top-50 cross-varying patterns using the technique detailed in Chapter 7. Each node (feature) is labelled 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 visualisation depict the top-$k$ cross-varying patterns. The correlation strength is depicted by the edge thickness and colour.

Figure 8.13 clearly shows that the top-50 cross-varying patterns form two clusters, one revolving around *Overseas Visitors, Other Dwelling - Improvised home* and the second cluster around...
### TABLE 8.1: Cross-varying patterns from the Brisbane crime dataset.

<table>
<thead>
<tr>
<th>top-$k$</th>
<th>$L_{X,Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlawful Entry With Intent - UNOCCUPIED</td>
<td>0.603</td>
</tr>
<tr>
<td>Unlawful Entry With Intent - AGED_25_29</td>
<td>0.580</td>
</tr>
<tr>
<td>Unlawful Entry With Intent - UNEMPLOYED</td>
<td>0.564</td>
</tr>
<tr>
<td>Other Stealing - OVERSEAS_VISITOR</td>
<td>0.551</td>
</tr>
<tr>
<td>Other Theft - OVERSEAS_VISITOR</td>
<td>0.535</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>bottom-$k$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arson - AGED_85_89</td>
<td>-0.154</td>
</tr>
<tr>
<td>Arson - AGED_90_94</td>
<td>-0.152</td>
</tr>
<tr>
<td>Liquor (excl. Drunkenness) - SEPARATE_HOUSE</td>
<td>-0.139</td>
</tr>
<tr>
<td>Arson - FLAT_UNIT</td>
<td>-0.139</td>
</tr>
<tr>
<td>Rape and Attempted Rape - reserve</td>
<td>-0.135</td>
</tr>
</tbody>
</table>

**Figure 8.13:** Visualisation of top-50 cross-varying patterns.

Unemployed, Age 25-29, Unoccupied Dwelling, Unlawful Entry With Intent, Unlawful Entry Without Violence. The visualisation environment is dynamic so that the user can move and delete any patterns that are not of interest. Figure 8.14 shows the visualisation of the bottom-50 cross-varying patterns. In this case, the visualisation easily allows the user to see that Arson shows a negative cross-varying relationship with People aged 80-100, locations with Flat or Units and locations of Universities and Colleges.
8.1 Cross-Patterning Relationships

**Figure 8.14:** Visualisation of bottom-50 cross-varying patterns.

**Figure 8.15:** Visualisation of selected top-\(k\) crime cross-varying patterns.
Users can also select specific crime features to visualise from the top-\(k\)/bottom-\(k\) results. Figure 8.15 shows the cross-varying 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 visualisation, duplicate geospatial and census features are not removed. As part of a crime analysis system these cross-varying patterns and other cross-patterning relationships can be used to further investigate and explain the behaviour of crime within the study region.

### 8.1.3 Cross-Distribution Patterns

#### 8.1.3.1 Density Tracing

To discover cross-distribution patterns we first use the Density Tracing approach detailed in Section 6.1. We utilise GLS as the region ordering due to its low edge overlap and low edge variance.

![Figure 8.16: Visualisation of Density Tracing top-50 cross-distribution patterns.](image-url)
Figure 8.16 shows the initial visualisation of the top-50 cross-distribution patterns from this experiment. The initial layout of patterns in the visualisation may make discovering interesting patterns unclear, so we apply the 2D spring embedder layout as described in Chapter 7. Figure 8.17 depicts the resulting visualisation. From this image it is clear that *Other Dwelling - Improvised home* is strongly cross-distributed with a number of datasets.

![Visualisation of Density Tracing top-50 cross-distribution patterns.](image)

**Figure 8.17:** Visualisation of Density Tracing top-50 cross-distribution patterns.

We can dynamically highlight datasets of interest in the visualisation by selecting *Other Dwelling - Improvised home* and *Overseas Visitors*. Figure 8.18 shows the datasets that show cross-distribution with these specific datasets.

There is a point where the number of discovered patterns can become too much for the end user to handle - so called ‘information overload’. As can be seen from Fig. 8.19 a low minimum similarity will not reveal any patterns, whereas a high limit may introduce unwanted ‘noise’. We offset this drawback by allowing the user to retrieve the \( k \) Most Similar and/or \( k \) Least Similar results. It is noted that even with a high minimum similarity the number of patterns
8.1 CROSS-PATTERNING RELATIONSHIPS

**Figure 8.18:** Visualisation of Density Tracing selected cross-distribution patterns.

**Figure 8.19:** Lowering minimum similarity to reduce cross-distribution patterns.
Figure 8.20: Comparison between GLS, DFS and BFS for kidnapping.

Figure 8.21: Comparison between weighted and non-weighted regions for reserves.

generated by our approach is much smaller than the number of patterns returned by ARM based techniques.

Figure 8.20 shows a comparison between the four region ordering approaches GLS, DFS, BFS and NN-5. The four techniques show similar results with these datasets and study region, demonstrating the minimal role of different orderings for this particular experiment.
8.1 CROSS-PATTERNING RELATIONSHIPS

In the next experiment we introduce clustering results into the framework. We select the reference feature reserves and assign weights to specific regions. These regions can be found using a clustering technique or may be a user defined influence on the study area. We discover two clusters in the reserves dataset; suburb Coorparoo is the densest cluster and is assigned a region weight of 0.1 while suburb Tennyson a weight of 0.5. Figure 8.21 compares the dissimilarity scores of reserves when cluster weights are used and when they are not. The region weights have lowered the dissimilarity of a number of crime types, with the most affected being Serious Assault (SEAS), Other Stealing (OTST) and Trespassing (TRAV).

8.1.3.2 Graph Mining

<table>
<thead>
<tr>
<th>top-k</th>
<th>Pearson’s $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquor (excl. Drunkenness) - IMPROVISED_HOME</td>
<td>0.784</td>
</tr>
<tr>
<td>Receiving Stolen Property - IMPROVISED_HOME</td>
<td>0.780</td>
</tr>
<tr>
<td>Unlawful Entry With Intent - SAME_ADDRESS_5Y</td>
<td>0.761</td>
</tr>
<tr>
<td>Unlawful Entry With Intent - AGED_30_34</td>
<td>0.760</td>
</tr>
<tr>
<td>Possess Suspected Stolen - IMPROVISED_HOME</td>
<td>0.759</td>
</tr>
</tbody>
</table>

**TABLE 8.2:** Correlations from the Brisbane crime dataset.

We utilise the Graph Mining approach detailed in Section 6.2 to mine the datasets to extract the top-$k$ cross-distribution 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 patterns shown in Table 8.2 it can be seen that there is a strong cross-distribution pattern between the Other Dwelling - Improvised home and the crimes Liquor (excl. Drunkenness) and Receiving Stolen Property. This information can then be used to further investigate the cause of these specific patterns and also allow for targeted crime prevention efforts. For example, police may increase foot patrols in areas that have a high concentration of improvised homes, or, the city council may choose to install security cameras and increase lighting in those areas.

Figure 8.22 shows the visualisation that was developed to help the user extract interesting patterns. For this example we use the same datasets as described above and extract the top-50 results. Each vertex (feature) is labelled and its size is a depiction of the correlation strength,
8.1 Cross-Patterning Relationships

**Figure 8.22:** Visualisation of Graph Mining top-50 cross-distribution patterns.

**Figure 8.23:** Visualisation of Graph Mining bottom-50 cross-distribution patterns.
where the size of the circle represents the largest $r$ score of its relationships. The edges between vertices in our visualisation depict the top-$k$ cross-distribution patterns. The correlation strength is depicted by the edge thickness and colour.

Figure 8.22 shows that the top-50 correlative patterns broadly form two clusters, one revolving around *Other Dwelling - Improvised home* and the second cluster around *Unlawful Entry With Intent, Unlawful Entry Without Violence*. The visualisation environment is dynamic so that the user can move and delete any patterns that are not of interest. Figure 8.23 shows the visualisation of the bottom-50 associative patterns. In this case, the visualisation easily allows the user to see that there is weak pattern between *Manslaughter, Universities and Colleges, Flat or Units*.

### 8.1.4 Cross-Patterning Visualisation

The visualisation environment described in Chapter 7 can combine cross-association, cross-varying and cross-distribution patterns so that interesting cross-patterning relationships are highlighted. Cross-patterning relationships that have a combination of cross-association, cross-varying and/or cross-distribution patterns may be of particular interest to the end user. These relationships contain a combination of spatial properties.

Figure 8.24 depicts the combination of all cross-association, cross-varying and cross-distribution patterns using the initial random layout. Overlapping patterns are highlighted in grey, however this random layout does not visualise these patterns in a useful manner. With such a large number of results, the most suited visualisations are the zoomed out 2D and 3D layout, Figures 8.25 and 8.26 respectively.

Figure 8.27 shows only the overlapping cross-patterning relationships using the circle layout. That is, only the relationships that exist in more than one of the cross-association, cross-varying and cross-distribution result sets are shown. The 2D layout shown by Figure 8.28 shows that these overlapping cross-patterning relationships form three clusters; one cluster centred around *Overseas Visitors* and another around *Other Dwelling - Improvised home*. The third cluster is loosely formed around *People aged 25-34, Unemployed People, People with Different Address*.
8.1 Cross-Patterning Relationships

**Figure 8.24:** Visualisation of all cross-patterning - Random layout.

**Figure 8.25:** Visualisation of all cross-patterning - 2D spring layout.
8.1 Cross-Patterning Relationships

**Figure 8.26:** Visualisation of all cross-patterning - 3D spring layout.

**Figure 8.27:** Visualisation of overlapping cross-patterning - Circle layout.
These results suggest that to enhance crime prevention, these specific groups should be targeted.

8.2 Computational Comparison

In this section we examine the computational aspect of each algorithm for discovering cross-patterning in areal aggregated spatial datasets. Figure 8.29 compares the computation time required for each approach as applied to the experiments described in this chapter. The techniques for discovering cross-association, cross-varying and the Density Tracing for cross-distribution show a similar computational performance with an average of 28 seconds of required computation. The Graph Mining technique to discover cross-distribution patterns is able to run in under 2 seconds.

We further compare the computation time required for Density Tracing and Graph Mining for discovering cross-distribution patterns in Fig. 8.30. We limit the number of datasets (crimes, census and features being compared) to show that in all cases the graph based approach is more
efficient. For the experiment described above (108 datasets) the Graph Mining approach averaged 1509 msec compared with 29090 msec for the Density Tracing algorithm. This is largely due to the efficient graph structure that Graph Mining employs to model the areal aggregated datasets.
8.3 Summary

Intelligent crime data analysis facilitates an improved understanding of the dynamics of criminal activities. Crime analysis requires a combination of heterogeneous data, such as socio-economic and socio-demographic factors, geospatial features and crime datasets so that useful patterns can be discovered.

The focus of this thesis is on the development of a framework for the discovery and visualisation of interesting patterns across many spatial layers of areal aggregated data. This chapter has studied how this framework can be used for crime analysis by applying the algorithms to crime, socio-economic, socio-demographic and geospatial feature data provided by the QPS, ABS and other agencies.

Our techniques are able to discover interesting cross-patterning relationships within these datasets. Coupled with our visualisation technique, we are able to show how certain factors influence the cross-patterning among other datasets. For example Figure 8.28 shows that Overseas Visitors and Other Dwelling - Improvised home depict a strong cross-patterning with a number of crime datasets. Police could use such a pattern to target locations with a high number of overseas visitors or areas with a high concentration of improvised homes.

As with many approaches to discover patterns and relationships in large datasets there are often a number of ‘obvious’ patterns discovered. For example, one cross-varying relationship we discovered is that People aged 80-100 has a negative relationship with Arson. Such a pattern may be of no interest or already known to the user. Our visualisation environment is useful in such situations as the user can dynamically move and delete any of these uninteresting patterns. As a decision support tool our framework allows users to interpret, understand and make informed decisions.

This chapter focused on how the techniques developed in this thesis can be used to enable intelligent crime analysis. However this is not the only application that can benefit from the techniques we have developed. The cross-patterning techniques we have presented can be used
with any spatial areal aggregated dataset where the requirement is to find patterns across many spatial layers.
The contributions we have made to the discovery of cross-patterning from areal aggregated spatial datasets are significant because they provide tools that knowledge seekers can use. This has been shown through our application of these methods to Queensland crime, socio-economic, socio-demographic and geospatial data, where we have been able to quantify and analyse cross-patterning relationships using the developed techniques and visualisation environment.

In this chapter we highlight the contributions that this thesis makes and then present possible future avenues for research.

9.1 Summary of Contributions

The work conducted in this thesis addresses the four main areas we outlined at the start of this project. They were to:

- Identify and define cross-patterning relationships across multiple areal aggregated spatial datasets;
- Develop new techniques to discover and model these cross-patterning relationships;
- Develop a visualisation environment to enable users to easily interpret the discovered relationships; and,
- Use the developed tools for the analysis of crime datasets along with socio-economic, socio-demographic and geospatial features.
The first contribution of this thesis is to identify and define cross-patterning relationships that can occur across multiple areal aggregated spatial datasets. Cross-patterning allows relationships between different spatial layers across locations to be modelled and quantified.

Discovering cross-patterning in areal aggregated spatial datasets allows for a greater understanding of the various factors that can influence datasets. These discovered cross-patterning relationships can be of use to a number of people and agencies such as regional planners, politicians, police and residents. For example, we may discover that particular suburbs with a high population of overseas visitors show a cross-patterning with high incidents of fraud in surrounding suburbs. Once information such as this is discovered it can be used as the basis for decisions about crime prevention and planning. While these cross-patterning relationships may not be casual factors the information is still of value to police and policy makers.

We identify and define three ways in which these cross-patterning relationships can be modelled. They are, cross-association patterns that model the relationship between multiple datasets for each region while ignoring the effect of neighbouring regions, cross-varying patterns that take into consideration the effect of all local neighbouring regions and cross-distribution patterns that consider one local neighbouring region and the global distribution of the dataset. The most suited type of cross-patterning is dependent on the application and dataset. The goal of this thesis is to provide a generic framework that can discover all three types of cross-patterning from areal aggregated spatial datasets.

The second contribution is to develop techniques to discover and model these three types of cross-patterning relationships. In Chapter 4 we propose a framework for automating the detection of cross-associations based on a given areal base map. We investigate a series of geospatial preprocessing steps involving data conversion and classification so that traditional ARM can be applied to areal aggregated spatial datasets. Our ACMAM framework is able to discover multivariate cross-associations.

In Chapter 5 we introduce the concept of bivariate spatial association and show how this can be used to discover cross-varying relationships. We compare and contrast two techniques that are able to quantify this bivariate spatial association, Lee’s $L$ index and Wartenberg’s Cross-MC.
9.1 Summary of Contributions

We demonstrate that Lee’s $L$ index is more robustly able to capture bivariate spatial association and suggest it as a suitable candidate for cross-varying discovery. We then demonstrate that for areal aggregated spatial datasets recorded over regions of irregular size and shape (such as suburbs) Lee’s $L$ index cannot reliably capture cross-varying. We present a modification of the weight matrix used when calculating the spatial lag vector (which in turn is used to calculate Lee’s $L$ index) that is able to overcome this limitation.

In Chapter 6 two new approaches are developed to model and quantify cross-distribution relationships. The first of these approaches is Density Tracing. Density Tracing models the spatial distribution of an areal aggregated dataset as a density trace and then computes the similarity between these traces to discover those that have a cross-distribution relationship. The density traces are constructed by projecting them on the Cartesian plane following a defined spatial ordering of the regions. We investigate four popular locational ordering methods to calculate this spatial ordering and suggest GLS as the most suitable. We then demonstrate how to calculate the distance (dissimilarity) between two traces using the RWLIP algorithm. This dissimilarity score can then be used to rank spatial layers on the strength of their cross-distribution relationships.

The second approach we introduce to discover cross-distribution relationships is Graph Mining. This technique models the spatial distribution of areal aggregated spatial datasets as weighted directed multigraphs. These graphs store the density values for each region, and also each regions influence on neighbouring regions. We then prune these graphs by extracting the most influential edges (that is, edges that represent similar density change among neighbours). We investigate two approaches to achieve this pruning and suggest MST as it is argument free. These pruned graphs are then compared by extracting each vertices degree to compute a correlation coefficient such as Pearson’s $r$ or Jaccard’s Index $J$ so that cross-distribution patterns can be discovered.

The goal of this thesis is to provide a generic framework that can discover cross-patterning from areal aggregated spatial datasets. The approaches presented to discover cross-association, cross-varying and cross-distribution can be used individually or together to discover cross-patterning relationships.
The third contribution of this thesis is to develop a visualisation environment that enables users to easily interpret discovered cross-patterning relationships from our algorithms. The main purpose of the visualisation environment is to allow the end user to easily access and interpret the cross-patterning information extracted from the datasets. It also allows for the combination of results from the cross-association, cross-varying and cross-distribution algorithms, in order to identify those patterns that overlap with each other. These overlapping patterns may indicate a stronger cross-patterning, and are highlighted to the user. The overall goal of the visualisation environment is to enable users to easily interpret the cross-patterning relationships contained within the datasets.

The algorithms developed for discovering cross-patterning relationships are designed so they can be applied to areal aggregated spatial datasets from a wide variety of disciplines. In this study we focus on crime analysis and the fourth contribution is to analyse crime datasets using the developed tools to discover interesting cross-patterning relationships across areal aggregated crime, socio-economic, socio-demographic and geospatial features. Our techniques were able to discover interesting cross-patterning relationships within these datasets.

By performing each of these tasks we have satisfied the aims of our research and contributed advancements in crime analysis and data mining of areal aggregated spatial datasets.

## 9.2 Future Work

There are a number of areas where future work could be undertaken. In this thesis we have focused on cross-patterning relationships that happen in one time period. Future work to discover spatio-temporal cross-patterning may prove useful. In particular, extending the Density Tracing approach to model spatio-temporal cross-distribution may be beneficial. Density traces are constructed by projecting them on the Cartesian plane following a defined spatial ordering of the regions. It may be possible to extend the concept of these traces to include a third axis (time) and then modify the RWLIP algorithm to calculate the similarity of these.
Another possible extension of the Density Tracing approach is to further integrate it with clustering algorithms. Density Tracing is able to incorporate clusters by assigning regions weights (where clustered areas would have larger weight). At this stage the weights of the regions must be specified by the user (gained from either a domain expert or the use of a clustering algorithm). It may be beneficial to integrate this use of clustering algorithms with Density Tracing to provide a more user friendly approach.

The research presented in this thesis is applicable to a wide range of other applications that may lead to economic, social and environmental advantages. We have presented a case study using our techniques for crime analysis however the techniques could also be applicable to geospatial datasets in a variety of other fields including disaster management, epidemiology, business intelligence, geology, environmental monitoring, marketing and e-commerce. We wish to undertake further case studies to investigate the potential use of our framework in these other disciplines.

In this thesis, we have achieved the goals that we set out to achieve, through the development of techniques for the discovery of cross-patterning relationships in large areal aggregated spatial datasets. We have shown that the techniques are able to perform on real crime data to find interesting patterns. Given such a wide variety of future research challenges, this work has provided a good foundation in an area that has not yet been widely investigated.
References


REFERENCES


[38] Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)


REFERENCES

Canberra, Australia (1998)


