Ubiquitous Intelligence for Smart Cities: A Public Safety Approach

(Omowunmi Elizabeth, **ISAFIADE** (ofalola@cs.uct.ac.za))

This thesis is submitted in fulfilment of the academic requirements

for the degree of

Doctor of Philosophy

In the Department of Computer Science, Faculty of Science

University of Cape Town

March 29, 2017

CERTIFICATION

As the candidate's supervisors, we have approved this thesis for submission.

Supervisor: A/Prof. Sonia Berman and A/Prof. Antoine Bagula

Signature: _____

Date: _____

DECLARATION

I declare that this thesis is my own work. Where collaborations with other researchers has occurred, or materials generated by other researchers is included, the parties and/or materials are indicated in the acknowledgements or are explicitly referenced as necessary.

This thesis is being submitted for the degree of Doctor of Philosophy in Computer Science at the University of Cape Town, South Africa. This work has not been submitted to any other university or institution for any other degree or examination.

Signed by candidate Signature removed

17 November 2016

Signature

Date

Omowunmi Elizabeth ISAFIADE

DEDICATION

To the Almighty God, the master planner, through whom ALL things are possible, His grace has seen me through this.

To my late father, who believed so much in education and made a lot of sacrifices to see me to this.

To my sweet mum, whose invaluable support and selfless-sacrifices have fuelled this exploration and encouraged this endeavour.

To my sweetheart and lovely angels, whom have always been a source of inspiration and encouragement.

To my families at all levels, who are constantly in support of my dreams and aspirations.

Abstract

Citizen-centered safety enhancement is an integral component of public safety and a top priority for decision makers in a smart city development. However, public safety agencies are constantly faced with the challenge of deterring crime. While most smart city initiatives have placed emphasis on the use of modern technology for fighting crime, this may not be sufficient to achieve a sustainable safe and smart city in a resource constrained environment, such as in Africa. In particular, crime series which is a set of crimes considered to have been committed by the same offender is currently less explored in developing nations and has great potential in helping to fight against crime and promoting safety in smart cities. This research focuses on detecting the situation of crime through data mining approaches that can be used to promote citizens' safety, and assist security agencies in knowledge-driven decision support, such as crime series identification. While much research has been conducted on crime hotspots, not enough has been done in the area of identifying crime series.

This thesis presents a novel crime clustering model, CriClust, for crime series pattern (CSP) detection and mapping to derive useful knowledge from a crime dataset, drawing on sound scientific and mathematical principles, as well as assumptions from theories of environmental criminology. The analysis is augmented using a dual-threshold model, and pattern prevalence information is encoded in similarity graphs. Clusters are identified by finding highly-connected subgraphs using adaptive graph size and Monte-Carlo heuristics in the Karger-Stein mincut algorithm. We introduce two new interest measures: (i) Proportion Difference Evaluation (PDE), which reveals the propagation effect of a series and dominant series; and (ii) Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series.

Our findings on experimental quasi-real data set, generated based on expert knowledge recommendation, reveal

that identifying CSP and statistically interpretable patterns could contribute significantly to strengthening public safety service delivery in a smart city development. Evaluation was conducted to investigate: (i) the reliability of the model in identifying all inherent series in a crime dataset; (ii) the scalability of the model with varying crime records volume; and (iii) unique features of the model compared to competing baseline algorithms and related research. It was found that Monte Carlo technique and adaptive graph size mechanism for crime similarity clustering yield substantial improvement. The study also found that proportion estimation (PDE) and PSE of series clusters can provide valuable insight into crime deterrence strategies. Furthermore, visual enhancement of clusters using graphical approaches to organising information and presenting a unified viable view promotes a prompt identification of important areas demanding attention. Our model particularly attempts to preserve desirable and robust statistical properties.

This research presents considerable empirical evidence that the proposed crime cluster (CriClust) model is promising and can assist in deriving useful crime pattern knowledge, contributing knowledge services for public safety authorities and intelligence gathering organisations in developing nations, thereby promoting a sustainable "safe and smart" city.

Acknowledgements

My profound gratitude goes to the Almighty God, the greatest time keeper, the master planner, who blessed me with life and good health that aid the completion of this thesis. God, I give you all the glory.

This research would not have been successful without the invaluable support of my supervisors, Professor Sonia Berman of the University of Cape Town (UCT) and Professor Antoine Bagula of the University of Western Cape (UWC). Thank you for the constructive criticism during each phase of this research work and the sacrificial time in proofreading the content therein. Your patience, encouragement and contributions have fuelled this exploration in innumerable ways. To the authorities at the University of Cape Town, the African Institute for Mathematical Sciences & German Academic Exchange Service (AIMS-DAAD), and the L'Oŕeal-UNESCO for Women in Science Fellowship, I am grateful for the financial opportunities given to me during my research. I wish to acknowledge and thank the Google Anita Borg scholarship, the Anita Borg Institute (Grace Hopper Celebration (GHC) of Women in Computing), and the Heidelberg Laureate Forum (HLF) authorities, for providing me a unique opportunity that inspires excellence. Meeting and interacting with young bright minds, senior colleagues and scientific leaders of computing from all over the world was extremely inspiring and encouraging for young female scientists like myself.

Special thanks to the South African Police Service (SAPS) for the useful information supplied on crime attributes and crime datasets. In particular, I would like to thank the office of the Western Cape Provincial Commissioner, DWO Leggett, WO Lotz, and LT. Col. McLean at the Table-View station, the FCS department - Milnerton station, and Col. Lemmer Scholtz at the Crime Intelligence unit, Bishop Lavis, for their support and selfless response during experimental phases of this research.

Much love to my family, especially my sweet mum, whose unwavering support has been my strength, I am blessed to have you. With a sincere heart, I appreciate my best-half, the larger part of me, Adeniyi Isafiade, and my precious angels, Oluwatamilore and Oluwatobiloba; you have been a great source of joy and encouragement, thank you for your unwavering love and support. On a resounding note, I love you all.

Contents

CERTIFICATION i

DECLARATION ii

DEDICATION iii

Abstract iv

List of Abbreviations Used xx

List of Figures xxi

List of Tables xxi

1 INTRODUCTION 1

1.1 Urbanisation and the Smart City 1

1.2 Trend of Urbanisation and Crime in South Africa 5

1.3 Research Aims 9

1.4 Research Scope and Limitation 11

1.5 Research Rationale . . . 12
1.6 Research Questions . . . 12
1.7 CSP Mining Problem Definition (CriClust Approach) . . . 13
1.8 Contributions and Outline . . . 15
1.9 Declaration of Recent Publications . . . 16
1.10 Thesis Outline . . . 17
1.11 Chapter Summary . . . 18

2 BACKGROUND AND LITERATURE REVIEW 20

2.1 Background Study and Related Research . . . 21
2.2 Overview of Crime Data Mining . . . 22
2.2.1 Classical Sequential Approach to Data Mining . . . 26
2.3 Current Implementation to Crime Data Mining . . . 28
2.3.1 Existing Strategies in Crime Data Mining . . . 28
2.3.2 Existing Applications and Techniques for Crime Data Mining . . . 30
2.4 Comparison of Crime Data Mining Applications and Techniques . . . 37
2.5 Crime Series Research . . . 40
2.6 Highly Connected Sub-graphs Concept . . . 45
2.7 Research Goal Revisited . . . 45
2.8 Summary of Research Gaps and Opportunities . . . 46
2.9 Chapter Summary . . . 47

3 MODEL FORMULATION AND DESIGN METHODOLOGY **49**

3.1 Research Design: System Model for Crime Analysis 50

3.1.1 Research Focus and Approach 52

3.2 CriClust in Crime Series Pattern Detection 56

3.2.1 Problem Definition and Model Formulation 57

3.2.2 Cluster Identification Using Highly Connected Sub-graphs 69

3.2.3 CriClust Problem Specification 71

3.2.4 Adaptive Graph-Size-Based Contraction Operations in Crime Series Detection 73

3.3 Reasoning: Statistical Interest Measures on Identified Potential Series 76

3.3.1 CSP Proportion Difference Evaluation 78

3.3.2 CS Pattern Space Enumeration 79

3.4 Evaluation Metrics 80

3.5 CriClust Algorithm 81

3.6 Chapter Summary 83

4 RESULTS, DISCUSSIONS AND EXPERIMENTAL EVALUATION **84**

4.1 Sample Rape Data For Experiment 84

4.2 Empirical Observations on Experimental Data 86

4.2.1 Background Level Cluster Information: Graduated Colour Map of Crime Information . . . 86

4.3 Empirical Analysis of Series Information Across Locations 90

4.3.1 Identified Series Information Across Locations 90

4.3.2 PDEs of Identified Series Information Across Locations 91

4.3.3 CriClust PSE Identified Across Locations 96

4.4 Quantitative Evaluation of CriClust Model 101

4.4.1 Assessment of Peculiar Features in CriClust Patterns 106

4.4.2 Performance of CriClust on Controlled Experimental Data 106

4.5 CriClust Scalability Trend on Cluster Identification 110

4.5.1 CriClust Analysis Class Model 110

4.6 Systematic Comparison of other Series Detection Models with Our Proposed Approach 113

4.7 Benchmarking: Baseline Comparison with Common Clustering Techniques 113

4.8 Chapter Summary 118

5 CONCLUSION, RECOMMENDATION AND FUTURE WORK 120

5.1 Research Summary and Conclusion 121

5.2 Synthesis of Empirical Findings 122

5.2.1 How can crime mining and machine learning support and promote the identification of crime series patterns (CSP) to provide actionable insight from crime data, and what heuristics can be used to augment such analysis? 122

5.2.2 What level of confidence can be expressed in such a model and how does the technique scale-up with increasing numbers of crime records? 123

5.2.3 Contribution to Smart City Development in Developing Nations 125

5.3 Recommendation 126

5.4 Limitations of Research 127

5.5 Opportunities for Future Research 128

5.5.1 Possible Expansion of the CriClust System . . . 128

5.5.2 Consideration for Streaming Data . . . 129

5.5.3 Consideration for Real Crime Data . . . 129

5.5.4 Long Term Qualitative Consideration and Evaluation . . . 130

References **130**

A Implementation Visualisations - CriClust System **144**

A.1 CriClust System Login Interface . . . 144

A.2 Process Selection Interface for Data Processing Features . . . 145

A.3 Cluster Processing Interface . . . 145

A.4 Interface for Graduated Colour Map of Cluster Information . . . 146

A.5 Cluster Information Interface . . . 146

B CriClust Sample Back-end Information **149**

B.1 GPS Coordinates of Some Locations . . . 152

List of Figures

1.1 Key components of a connected city: smart city of the future. 3

1.2 A depiction of the relevance of crime analytics in tackling crime. 4

1.3 Urban population growth rate (annual %) in South Africa. 5

1.4 Crime taxonomy and trend in South Africa [91]. 6

1.5 Distribution of households perception of changes in violent crime level across three-year intervals. 7

1.6 Distribution of households perception of changes in violent crime level across provinces (2008-2013). 8

1.7 Annual trend in reported sexual crimes in South Africa [1]. 9

1.8 A depiction of serial predator in related crime scenarios in a city. 10

2.1 A model of routine activity theory (RAT). 22

2.2 Citizens opinions on top priority for dealing with crime. 24

2.3 A review framework of data mining approaches to crime situation recognition. 25

2.4 A taxonomy of crime data mining tasks and techniques. 26

2.5 General iterative steps in data mining. 27

2.6 A depiction of a complex Bayesian network of suspect modelling. 43

2.7 Research Topics and Issues in Crime Mining in Developing Nations 47

3.1 Components of a situation management system 50

3.2 A depiction of citizen-centred safety promotion in a smart city development. 51

3.3 Essential components of data Mining for knowledge delivery. 52

3.4 A depiction of the research phases in CriClust system. 57

3.5 Depiction of the 2-D geometric projection for day attributes. 62

3.6 Depiction of the 2-D geometric projection for time attribute. 64

3.7 Depiction of a crime similarity graph for clustering. 68

3.8 Identifying sufficiently connected nodes in a crime similarity graph (red edges are min-cut). 68

3.9 Process of crime series detection with CriClust model. 70

3.10 A depiction of an adaptive process in promoting highly connected sub-graphs. 76

3.11 A depiction of serial predator colony. 77

3.12 A depiction of the propagation effect of two crime series. 77

4.1 Sample rape database for experiment 85

4.2 graduated colour map of identified locations of crime 87

4.3 The locations of crimes with corresponding crime densities: graduated colour map of cluster information on 1000 records 88

4.4 The locations of crimes with corresponding crime densities: graduated colour map of cluster information on 1500 records 88

4.5 The locations of crimes: graduated colour map of cluster information on 5500 records 89

4.6 Crime trend across locations with varying record size 90

4.7 The locations of crimes series. 91

4.8 Visualisation of the propagation effect and proportion difference of corresponding series at Mowbray. 92

4.9 Visualisation of the propagation effect and proportion difference of corresponding series at Wynberg. 93

4.10 Trend of series observed across locations with varying data size 96

4.11 Series cluster-1 identified for Mowbray location. 97

4.12 Series cluster-2 identified for Mowbray location. 97

4.13 Series cluster-1 identified for Cape-Town-Central location. 98

4.14 Series cluster-2 identified for Cape-Town-Central location. 99

4.15 Corresponding PSE for cluster-1 identified for Cape-Town-Central location 100

4.16 Corresponding PSE for Cluster-2 identified for Cape-Town-Central location 100

4.17 A depiction of the PDE and PSE of series at Grassy-Park location 101

4.18 An assessment of the amplified success rate (SR) with Monte Carlo and adaptive graph size approaches 103

4.19 Average Reciprocal Rank on CriClust Model 104

4.20 Time cost comparison of KS and AGS on training data 105

4.21 Precision and recall measure on CriClust model 105

4.22 Precision score for peculiar features based on significance threshold 106

4.23 Node-value indicator for number of series identified per location 107

4.24 CriClust performance on controlled experimental crime dataset: The case of 2-series clusters (showing PDE) identified per location 108

4.25 CriClust performance on controlled experimental crime dataset: The case of 3-series clusters identified per location 108

4.26 Scalability assessment of CriClust with increasing data size 110

4.27 Analysis class model showing interaction between objects in CriClust 112

4.28 Time cost comparison for common techniques compared 116

A.1 CriClust system login interface for authentication, access and privacy control: By considering factors and concerns relating to the sensitivity and peculiarity of a crime mining system, the use of the system requires approved authentication by a public safety personnel or the user thereof. . 144

A.2 CriClust selection interface for data processing features: Upon successful login, the functionalities of the system use this interface for flexible feature selection, based on the functional environment and the interesting facts required by public safety officers. 145

A.3 CriClust processing interface: on selecting the "Process Data" feature, the system systematically accesses the database and process the required cluster. The completion of the process activates or enables the "cluster" tab on the left pane of the system to be able to view the map. 145

A.4 Identified locations of crime: The system offers a high level view of information on the level of crime at a particular location. Each colour code is representative of the crime intensity at the highlighted location as captioned in the map (ranging from "more" crimes to "least" crimes). This is synonymous to hotspot-related identification. 146

A.5 The system reveals crimes per suburb: this example shows crime density in 8 locations based on information in 5500 records. 147

A.6 An instance of PDE and raw data of the dominant series identified at Steenberg location, where crimes were committed by a fat Indian male who kidnaps victims on Saturdays, usually around noon. 147

A.7 CriClust performance on controlled experimental crime dataset where n=4: That is 4-series were generated per location showing these were correctly detected as 4 clusters. 148

A.8 Database technology for connecting to CriClust: the system relies on a MySQL database (XAMPP for linux, PHP Version 5.5.6) technology to run its features. 148

B.1 Back-end interface: Connection to the database. The CriClust system depends on the MySQL database structure using the Java Database Connectivity (JDBC) API and XAMP for linux technology. 150

B.2 Back-end interface: Deploying CriClust. When the CriClust system is launched, it establishes connection with the database and once connection to the database is successful, it deploys the application for further process feature selection and cluster processing. 150

B.3 Back-end interface: Deploying CriClust with 4000 records. The log information is steadily documented on the GlassFish server as the process executes. 151

B.4 Back-end interface: Information on cluster processing. The system progressively and systematically identifies and groups clusters based on the similarity condition and the full pipeline for the CriClust model. 151

B.5 GPS coordinates of locations across Western Cape, South Africa. 152

List of Tables

1.1 CSP Detection Problem Statement 14

2.1 Description of existing crime pattern types for tactical and strategic analysis. 23

2.2 Support and confidence quantity of an association rule. 31

2.3 Comparing features of current crime data mining applications. 38

2.4 Comparing features of current crime data mining techniques. 39

3.1 Time of day categorisation and mapping. 54

3.2 RapeDataDB: Depiction of crime data information generation. 55

3.3 Description of the different categories of important features considered. 56

3.4 A simplified analogy to show that transitivity may not hold. 59

3.5 Computing distance measure for location attribute. 61

3.6 Angle and corresponding distance (x, y) values for day attribute. 63

3.7 A depiction of the 2-D components for determining prevalence characteristics. 65

3.8 Sample day attribute characteristics measure. 66

4.1 Locations revealing that series may not necessarily happen at hotspots ("most crimes") locations. 94

4.2 Trend of PDEs across locations with increasing crime record. 95

4.3 A depiction of the characterising (peculiar) features emerging for each series (S_i). 102

4.4 Performance of CriClust on controlled experimental data. 109

4.5 Comparative evaluation of existing crime series detection systems with CriClust System 114

4.6 Sample Cluster: K-Means 117

4.7 Sample cluster: Hierarchical clustering approach. 117

Glossary

List of Abbreviations Used

ANN: Artificial Neural Network

AGS: Adaptive Graph Size

BBN: Bayesian Belief Network

BFS: Breadth First Search

CDCI: Crime Detection and Criminal Identification

CriClust: Crime Cluster (model produced in this work)

CS: Crime Series

CSP: Crime Series Pattern

CSR: Crime Situation Recognition

EDA: Exploratory Data Analysis

GIS: Geographic Information System

HCS: Highly Connected Sub-graphs

HOT: Hotspot Optimization Tool

HSA: Hot Spot Analysis

IACA: International Association of Crime Analysts

min-cut: Minimum Cut

MO: Modus Operandi

NCPS: National Crime Prevention Strategy

PDE: Proportion Difference Evaluation

PSE: Pattern Space Enumeration

Pp: Predictive Policing

SA: South Africa

SART: Sexual Assault Response Team

S/N: Serial Number

STCA: Spatio-Temporal Crime Analysis

SVM: Support Vector Machine

TAR: Temporal Association Rule

2-D: Two-Dimensional

Chapter 1

INTRODUCTION

This chapter presents a general introduction and motivation for this study. The objectives and goals of this research are also stated. The motivation for this research is two-fold: (i) the challenge of combating crime faced by security and public safety agencies, especially in resource constrained environments such as in developing nations of Africa, where police are short-staffed and crime is rapidly increasing; and (ii) the potential of crime data mining in promoting a sustainable "safe and smart" city of the future, by providing useful knowledge that can be derived from the under-utilised plethora of information archived by the public safety authorities, which can further assist in presenting viable and efficient means of pro-actively tackling crime. As the term "smart city" is gaining prominence across different domains, there is some variety in what different researchers or smart city developers understand under this label or notion. Briefly discussed is the notion of smart city characteristics and their applicability in this domain of interest. Finally, an overview of the thesis structure is given.

1.1 Urbanisation and the Smart City

A smart city is one where intelligent technological approaches are used to tackle problems arising from the global urbanization phenomenon [24, 35]. This research investigates the use of data mining and machine learning for detecting crime series in a (rape) crime database and presenting this in an actionable way, in order to assist public

safety improvement in resource constrained settings. Combating crime is one of the top priorities for decision makers and public safety agencies in most cities and in smart city development. Public safety and security agencies often seek to identify smart ways or effective strategies to deter crime [90] not only in South Africa but across the world. It is important to recognise that crime data mining complements a number of citizen-centred safety related objectives in a smart city development [14, 22]. Indeed data mining has great potential in public safety application by extracting meaningful information (smart statistics) from large data sets and putting this information to effective use in criminal science and investigation [21, 41, 68]. A report of the United Nations' State of the World Population in 2014 [98] indicated that over 50% of the world's population already lives in cities. Furthermore, trends suggest that over 60% would be living in the cities by the year 2030, as a result of the continued rural-urban migration [94].

Crime can be anticipated to increase as the population increases [9]. This is particularly challenging in developing nations where resources for overcoming crime issues is relatively limited [44]. Hence, deterring crime becomes an important consideration for realising a sustainable "safe and smart" city in any nation. Consequent upon rapid urbanisation, it is reported that new kinds of problems are being generated in the cities as a result of the complex and enormous congregations of the populace [26]. These problems manifest themselves as environmental (e.g. pollution), health, governance, technical, social, physical or economic problems, among others, and particularly citizen's security (crime prevention and control) [97].

These problems have already placed continuous and tremendous strain upon infrastructure, residential and commercial properties, and social communities alike, and have resulted in enormous increase in related efforts to improve living conditions of the populace. Owing to these diverse challenges, it is widely recognized that current overcrowded cities need to be re-designed and re-modelled into cities of the future termed "smart city" [35]. Figure 1.1 highlights the key components that make up a smart city. The goal of a smart city is to transform existing cities into better and more intelligent ones [24], so as to achieve interdependent, interconnected, intelligent digital environments. However, the smart city concept has been open to a variety of interpretations [35], depending on the goals set out by a smart city's designer or planner. It is important to recognise that the notion of smart cities involve quite a number of topics within the Information and Communication Technology (ICT) domain, but extend to other areas in a trans-disciplinary approach. The ITUs Telecommunication Standardization

Sector (ITUT) [33], however, define smart city as one that strategically utilizes many smart factors such as ICT to increase the city's sustainable growth and strengthen city functions, while promoting citizens happiness and wellness. Despite the variety of definitions, they seem to all share two key characteristics that form the core of smart city development:

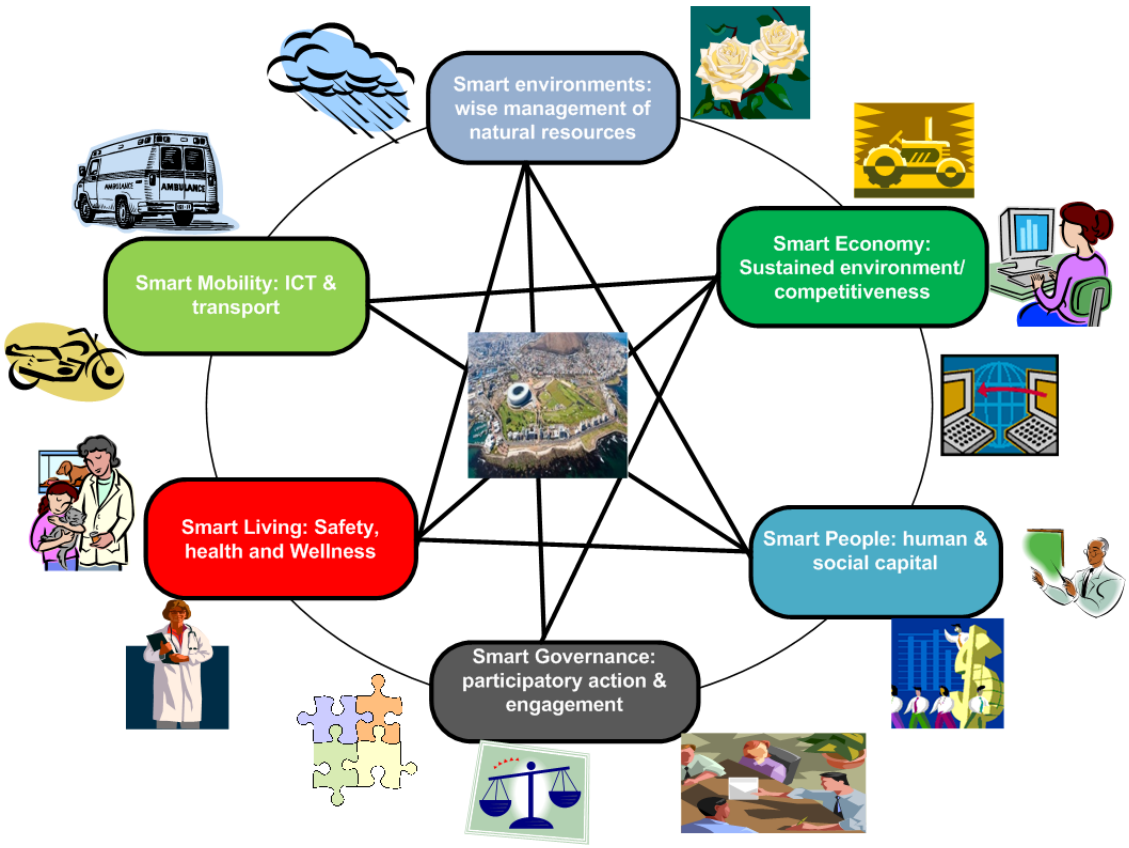


Figure 1.1: Key components of a connected city: smart city of the future.

- Increasing or promoting citizens' quality of life, and
- Improving the quality and efficiency of the services rendered by businesses, governing entities and decision makers.

Therefore, a smart-sustainable city is undoubtedly currently a top priority for decision makers and current effort is seeking ways to leverage Information and Communication Technology (ICT) in order to achieve a safe and better city life [59]. Rapid urbanisation may be problematic particularly for developing nations in terms of crime

outbreak, if not well managed. Hence the need for smart city development coincides with the need for developing effective crime prevention and control strategies.

While most smart city initiatives place much emphasis on the use of modern technology or equipment (e.g. explosive detectors, smart devices [1], armed weapons etc.) to fight crime, it is important to note that this may not be sufficient. Hence, deterring crime through the adoption of "smart statistics" [68] becomes an important strategy for realising a sustainable "safe and smart" city. This can be achieved through an effective use of crime data, known as crime data mining or analytics as shown in Figure 1.2.

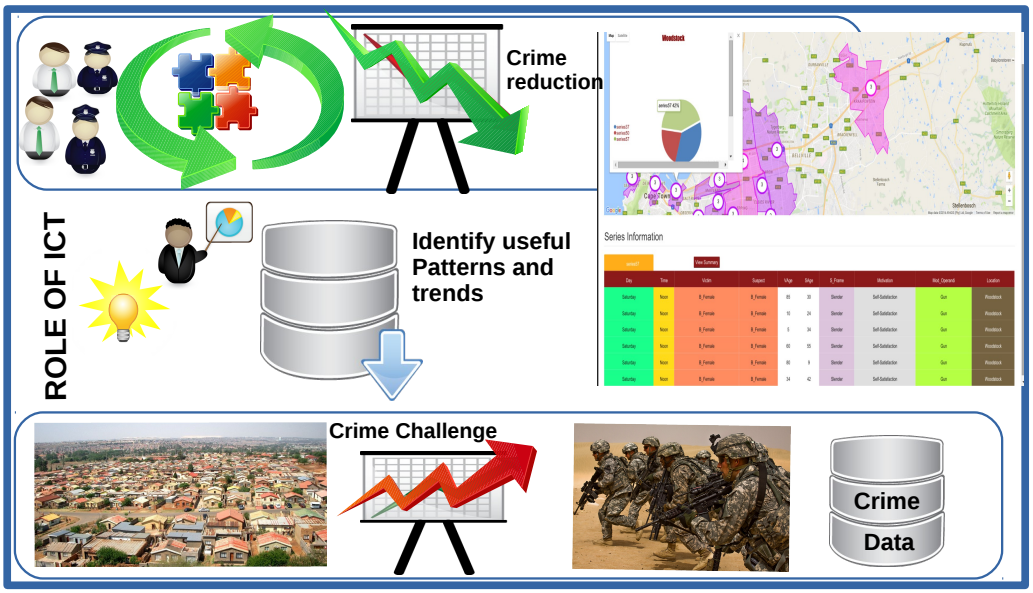


Figure 1.2: A depiction of the relevance of crime analytics in tackling crime.

Data mining or analytics is recognised as a means by which data could be transformed into smart statistics to gain insight for knowledge support [15, 31]. Moreover, security and effective policing have been identified as one of the best ways in which data analytics could have a great effect on "Smart Cities" [29]. Furthermore, it is widely recognised that being "smart" also involves the use of predictive geo-analytics to forecast the likelihood of crime events and focus policing resources on high-risk areas [34]. This can enhance operational planning with deep insights from crime data. Hence, the domain of smarter public safety is receiving quite significant attention

[1] http://streetsmartwomen.com/about/

lately.

1.2 Trend of Urbanisation and Crime in South Africa

The most recent survey by the South African Institute of Race Relations (SAIRR) indicated that two-thirds of South Africa's population now reside in urban areas [94]. The report confirms that the percentage of people who reside in urban communities increased from 52% in 1990 to 62% in 2011. Figure 1.3 presents the trend of rapid urbanisation in South Africa, between the years 2002 and 2010.

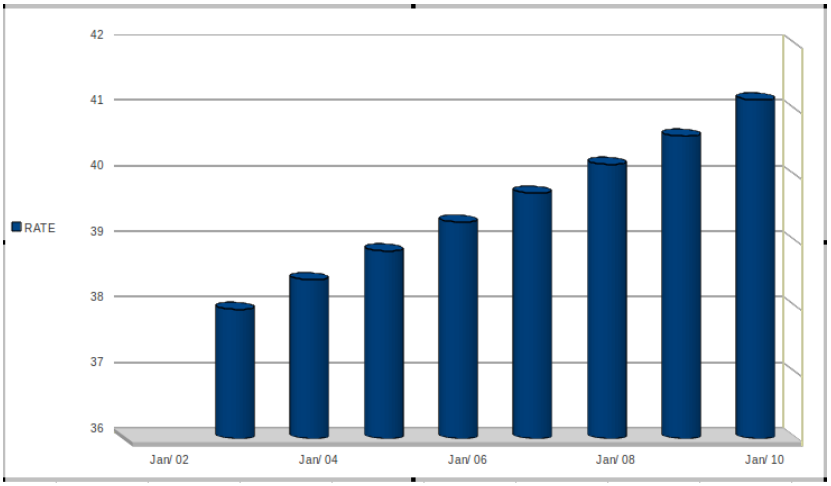
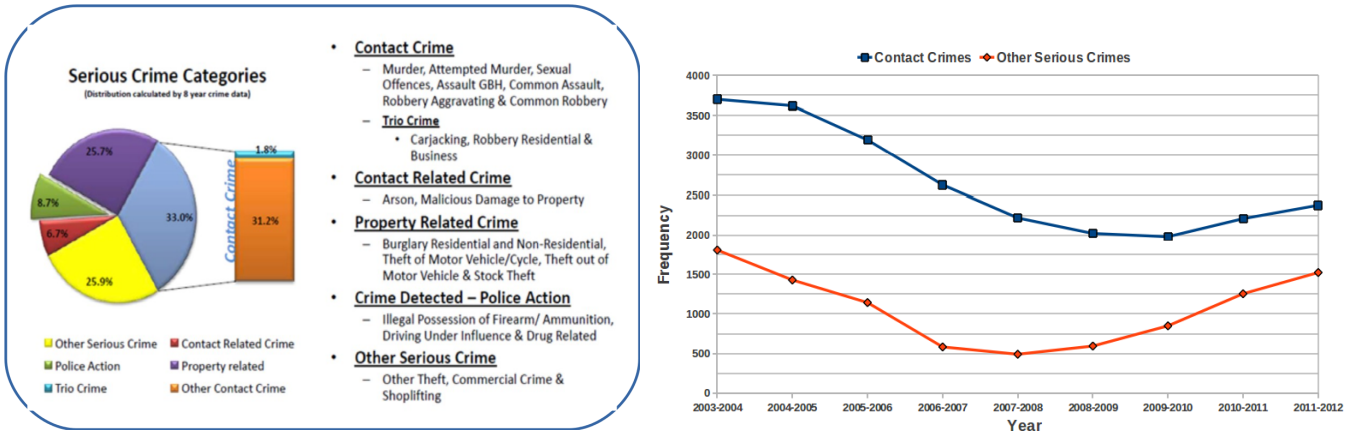


Figure 1.3: Urban population growth rate (annual %) in South Africa.

It is worth noting that the challenges related to public safety are magnified in developing or emerging countries such as in Africa, where crime rates are relatively high and resources for preventing them are relatively low [44]. This makes it crucial to devise smart means of tackling such challenges in resource constraint settings.

A number of smart city initiatives have been launched in different parts of the world [52, 59, 92], which focus on general enhancement of service delivery such as smart industry. However, it is widely recognised that a safe and secured environment is critical for a sustainable smart city. Figures 1.4(a) and 1.4(b) [91] present serious crime categories that have been identified and the trend of serious crime in one of the suburbs of the Western Cape in South Africa respectively. In South Africa crime prevention and control is a major concern [90]. However, crime mining shows promise to assist in crime reduction [15, 44].



(a) Distribution of serious crime categories in South Africa.

(b) Annual crime frequency at a suburb in Western Cape.

Figure 1.4: Crime taxonomy and trend in South Africa [91].

Crime data mining innovation is driven forward by two major factors [14]: (i) increase and richness of crime information archived by security agencies; and (ii) dynamics of offenders' behaviour. Building upon criminology studies, crime mining adds refinements and features that reflect the hidden knowledge in crime data, using data mining techniques, for example hotspot identification, leading to better interventions and policing strategies. However, the official crime statistics reported by the South African Police Services (SAPS) does not provide a sufficient understanding of the true crime situation [110]. The summary statistics usually reported [1] are at best only able to provide a rough indication of deterioration or improvement within different suburbs in South Africa or between police districts [2].

Containing crime remains a global issue [74]. In recent times, there has been a significant increase in the crime rate in South Africa (SA) [1] and this has been a major motivation for this research work. Figure 1.5 presents the distribution of households perception of changes in violent crime level across three-year intervals, while Figure 1.6 presents perception by province between the years 2010-2013 [1]. According to the distribution, 41.3% of all the households across provinces in South Africa indicated that violent crime increased significantly between the years 2010-2013. Therefore, the government's vision is to invent new strategies for improved public safety outcome [90].

[2] https://www.issafrica.org/crimehub/

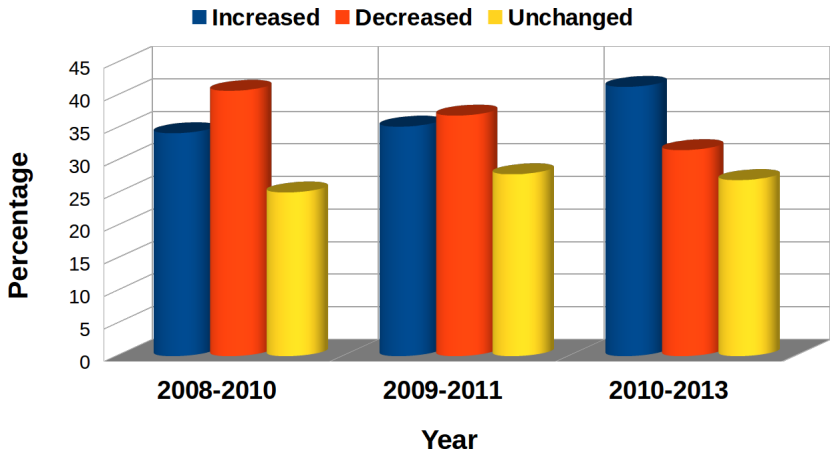


Figure 1.5: Distribution of households perception of changes in violent crime level across three-year intervals.

In light of this, security of citizens has inevitably become one of the priority areas in the Vision 2030 [3] of the South African Police Service. Security agencies are also under increased pressure in the face of ever-changing crime attributes and dynamics of offender behaviour. Tackling crime effectively is more challenging in resource constrained settings such as in developing nations, where crime intelligence experts and security personnel are limited and not enough technological solutions are in place to meet up with daily operational safety needs. However, the use of "smart statistics" [68], which can be derived through crime data mining approaches could assist in making optimal use of these limited resources.

The proliferation of perpetrators, the dynamic mode of offender's behaviour and the relentless use of traditional means by the security and law enforcement agencies in developing nations make crime prediction, prevention and control cumbersome. In spite of high crime reporting to public safety authorities in many developing nations in particular, the adoption of advanced, automated and computerised approaches of analysis to facilitate crime investigation and prediction remains conservative. Security agencies (particularly in developing nations) need to adopt more reliable and promising crime mining solutions to realise better tactical and strategic ways of dealing with crime. One such way is through the identification of a crime series pattern (CSP).

Crime series refer to a set of crimes considered to be committed by the same offender or set of offenders [78]. This work serves to assist in identifying crime series in a rape data. The motivation for considering rape crime is the fact that despite the heightened sensitivity and understanding about sexual assault and violence, South

[3] http://2030vision.co.za/publication/company/south-african-police-service-saps/

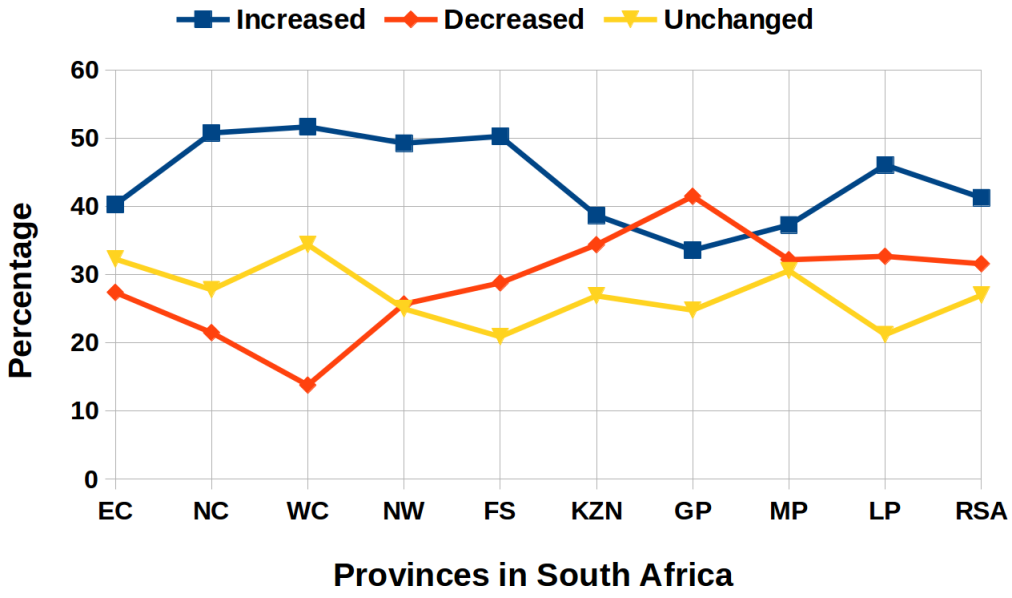


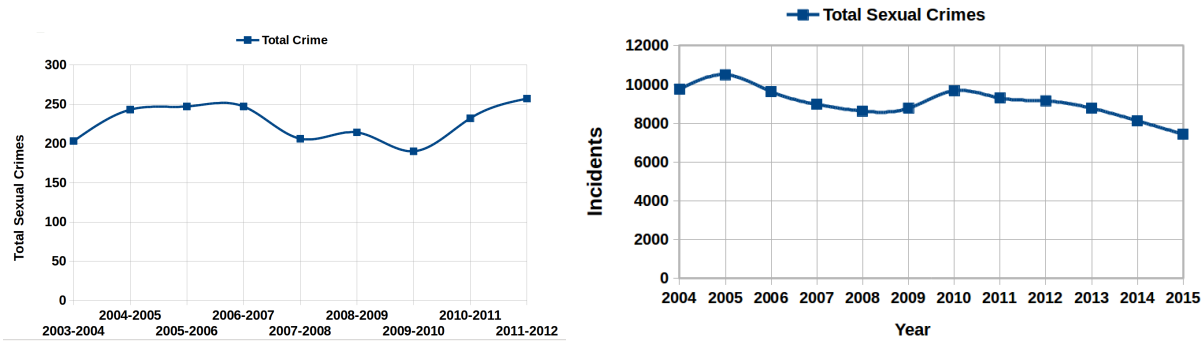
Figure 1.6: Distribution of households perception of changes in violent crime level across provinces (2008-2013).

African communities happen to be a place where rape, assault and murder of people (and particularly women and children) is of great concern [4]. Figure 1.7 presents the rate of sexual crimes in the Western Cape province of SA and one of its suburbs [1], which suggests that there is no significant reduction in sexual crimes committed and the sex crime rate is high. Moreover, there is evidence that sexual crime has generally been a neglected area of research in most parts of the world [108]. In addition, sexual crimes fall within the category of contact crimes, which is the most feared crime in SA. The University of Cape Town (UCT) has also expressed a deep concern about rape as some of its students have been raped lately and more may likely be, if proper checks are not put in place [77]. The reported rape incidents have some similarities in terms of the modus operandi involved in the attacks and are considered to have been perpetrated by the same offender. The assailant in the reported incidents also tends to be violent, resulting in several injuries for the victims. This experience resulted in the prompt formation of UCT's Sexual Assault Response Team (SART) [5]. It is widely recognised that an incident related to sexual assault brings trauma to its victims and leaves the victims in a very distressing situation. Such incidents also re-ignite bad memories for those who had been or close to those who are victims in the past. It has a crippling effect on individuals but also on the society at large. Consequently, the national crime prevention strategy

[4] http://rapecrisis.org.za/

[5] http://www.uct.ac.za/dailynews/?id=9679

(NCPS) [75] is also currently focusing on sexual assault and rape crimes, among others. While there exists some governmental and non-governmental awareness and education programmes (such as NCPS and SART) aimed at promoting awareness of crimes against children and gender crimes, this seems to be insufficient. The onus lies on the security and public safety authorities to gather the most intelligence out of the archived crime records, to better assist with a much more proactive response to such violations.



(a) Trend in reported sexual crimes at Gugulethu in Western Cape.

(b) Trend in reported sexual crimes in Western Cape.

Figure 1.7: Annual trend in reported sexual crimes in South Africa [1].

1.3 Research Aims

- The aim of the research was to identify potential crime series, a set of crimes thought to be committed by the same offender(s), by developing a crime mining solution, called CriClust (Crime Cluster), using a dual threshold scheme and highly connected sub-graphs approach, in order to assist the police in identifying recurring patterns timeously and subsequently channelling their resources accordingly. Furthermore, CriClust aims to preserve useful statistical properties, present actionable solution for public safety improvement in developing nations, and evaluate the scalability of the model in crime series identification.

Public safety agencies usually have a plethora of under-utilised crime incident reports at their disposal. If efficiently analysed, this could reveal some previously unknown useful insights evidence that can assist in apprehension of suspects, better attribution of past crimes and improved policing strategies. Furthermore, this can assist in the areas of determining criminal trends and knowledge-driven decision support. Research revealed [30, 57] that many

crimes are perpetuated by the same offender or criminal (called serial or repeat offender). This means that it is highly probable that a particular offender or group of offenders could be responsible for a large proportion of the crimes committed in a particular locality at a particular period of time. Therefore identifying such an offender or group of offenders using "smart statistics", which can be derived through data mining process, is a smart way to deal with crime deterrence and reduction. Furthermore, location-based approaches, where security agencies target certain times and locations, result in safer communities and crime reduction [109, 110], whereas random preventative patrol by the police does not significantly help in deterring crime [6]. Hence, it becomes crucial to adopt more tactical means of deploying resources for deterring crime, particularly in resource constrained settings. If patterns are identified timeously the police can prevent further recurrence. However, without proper means of identifying such patterns police can spend weeks, months or even years sifting through the crime records to discover patterns. Thus, crime series identification as depicted in Figure 1.8 is critical for suspect prioritisation in the public safety domain.

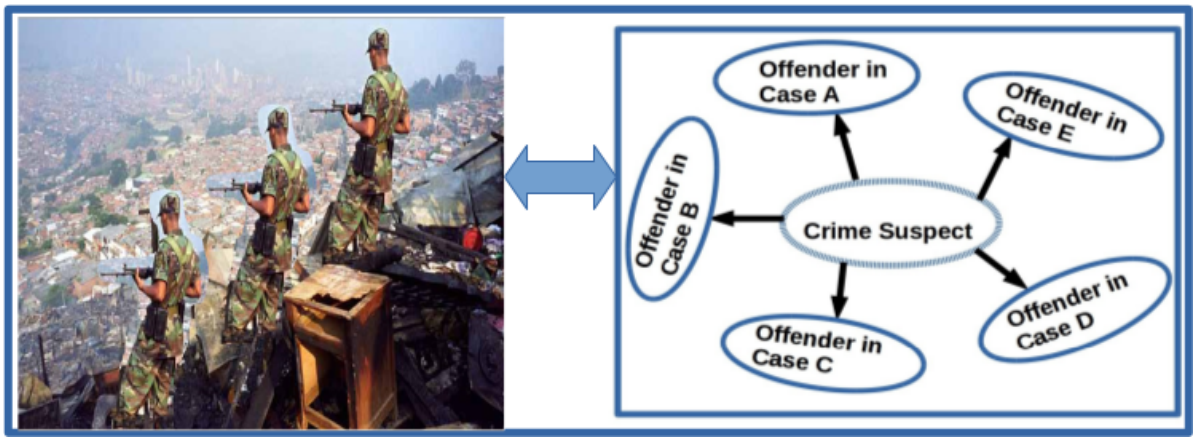


Figure 1.8: A depiction of serial predator in related crime scenarios in a city.

Several automated tools such as Predpol [86] exist for crime analysis, but are mostly only able to estimate general background crime levels (for example, hotspots which are areas or locations with high density of crimes). Moreover, while identifying crime (series) patterns remains critical for public safety improvement, our findings reveal that its analysis in developing nations has largely remained a manual task, which is tedious, resource-consuming and error-prone [73]. For example, the "Analyst's Notebook [7]" software is mostly only deployed at provincial police

[6] http://renegadenoble.com/weblog/evaluating-the-effectiveness-of-random-preventive-patrol/

[7] https://www.theknowledgeacademy.com/za/courses/ibm-training/ibm-i2-analyst-s-notebook-premium-module/

levels in SA because it is expensive and requires high technical expertise [8]. This means that data from local stations are sent to higher levels for analysis and may experience a (long) waiting period in pattern identification. Thus, the basic Excel tool is sometimes used at the local levels to filter attributes and identify possible patterns. This is typical in resource constrained settings.

1.4 Research Scope and Limitation

Trends in police recorded crimes have raised a debate as to whether the crime reports presented by the police actually reflect the true state of crime [9]. Further research has in fact proven that more often than not, a large percentage of all crimes go unreported or unrecorded [38][87], which is referred to as "dark figure of crime". However, this is not the focus of this research, which is rather concerned with how crime intelligence in developing nations can be promoted and further strengthened through the extraction of useful knowledge from archived crime data. Carefully analysing instances of crime information would help to understand the type of offenders' pattern or modus operandi under execution. The international association of crime analysts (IACA) has identified concise and promising means by which these crime data can be explored for tactical analysis [78]. These include the analysis of crime sprees, series, hotspots and hotplaces. This research has focused on series identification as it is currently not well studied, particularly in developing nations, and has great potential for assisting crime intelligence aimed at deterring crime. For experimental purposes, this research considered a rape database as an application scenario for identifying crime series, however the idea considered in this research can be extended to other forms of crime. The research further considered evaluating the scalability of the model. In this context, implementation on a single machine was appropriate; however parallelisation on a cluster would be beneficial but this is beyond the scope of this work.

[8] http://www.issafrica.co.za/iss_i2analysts_notebook.html

1.5 Research Rationale

Public safety agencies in developing nations need to be empowered and assisted with knowledge support for crime deterrence. Our findings reveal that crime experts particularly at the local police stations still make use of basic tools such as Microsoft Excel in filtering related crime incidents to detect patterns, which is cumbersome and time consuming. Crime mining complements a number of citizen-centred safety related objectives, such as strategic and tactical crime investigative analysis [45]. Furthermore, crime mining has the capability to deliver timeous evidence that can assist the public safety authorities in formulating security policies and allocating security resource(s) more efficiently. Such evidences can further set the trajectory in diverse ways [42], namely:

- Identification of normal and crime prone locations.
- Better attribution of past crimes.
- Identification of prime suspects.
- Apprehension of suspects.
- Better understanding of serial crime.
- Determination of mitigation priorities.
- Deriving a set of integrated and sustainable crime interventions.

1.6 Research Questions

This research draws on established theoretic crime mining model and framework [15] and assumptions from different theories of environmental criminology (crime pattern theory and routine activity theory) [18]. The focus here is based on "find-related-crime-patterns" query and possibly hotspots, among others.

1. *How can crime mining and machine learning support and promote the identification of crime series patterns (CSP) to provide actionable insight from crime data, and what heuristics can be used to augment such*

analysis?

In this work, the machine learning technique chosen for investigation was the use of clustering, and specifically the highly-connected subgraph approach to clustering. This is augmented using a dual threshold scheme with geometric projection.

2. *What level of confidence can be expressed in such a model and how does the technique scale-up with increasing numbers of crime records?*

   The model needs to be not only accurate, but in today's information-rich society, it should be scalable. Based on rape crime information that the experiment is conducted upon, this study investigated scalability in terms of computation time. Furthermore, in order to establish a level of confidence in the model, a generated crime series data was introduced and the predicted cluster output was checked in order to determine whether the output correlates with the initial input and what the error percentage was, if any. Furthermore, the overall viability of the model was investigated by conducting a comparative analysis with competing baseline algorithms and a high level comparison with existing related research in the domain of interest was done.

1.7 CSP Mining Problem Definition (CriClust Approach)

This study builds on the insight of crime mining and situation recognition and combines machine learning techniques for crime data analysis. This should provide new, deeper insights into crime situation recognition for improved public safety outcomes in developing nations, in order to promote a sustainable "safe and smart" city of the future. Crime series pattern discovery can be conceptualised as a type of subspace clustering [85], that is a clustering problem with cluster-specific attributes selection [107]. The proposition in this study is that most crime patterns exhibit at least a k minimum principal set that characterise the modus operandi (MO) of the offender(s) behaviour. This minimum principal set induces a similarity graph of crime objects and has the capability to reveal specific and general crime trends.

While there might exist slight variations in an offender's behaviour because past learning, current targets, situational and geographical attributes influence each crime outcome, a significant number of studies have revealed

that criminals act consistently in close proximity of space and time [70, 111]. The literature reveals that there is consistency in some behavioural variables such as site selection [7]. Hence, one could leverage these behavioural characteristics to extract statistically meaningful information from crime information [112].

Let C be a set of crime items or objects, where each crime object, say $i \in C$, is defined by a set of attributes A. Our interest lies in crime objects that exhibit a coherent pattern on a subset of attributes of A. The problem addressed in this research is peculiar and different from a "universal clustering problem" because there was a need to generate a relevant concept description of crime similarity in addressing the crime mining problem. Table 1.1 summarises the CSP detection problem.

Table 1.1: CSP Detection Problem Statement

Given	A spatial framework and a set of spatial-features with embedded instances. A similarity spatial neighbour relation, S_f. A significance threshold, $\mathcal{S}$. A prevalence threshold, $\mathcal{P}$.
Find	All CSPs with interestingness measure $\vdash (\mathcal{S}, \mathcal{P})$.
Objective	Completeness (all CSPs are detected). Correctness (CSPs are highly correlated).
Constraints	Statistically interpretable pattern. Minimum computational cost.

To identify crime series in a (rape) crime database, a hybrid model called CriClust, which combines similarity concepts, geometric projection, and graph connectivity (highly connected subgraphs), was adopted. CriClust is augmented with a dual threshold scheme. The detail of the model is described in chapter 3. Firstly, a crime

similarity function was derived which is used to connect crime instances that share related attribute information, based on the dual threshold scheme. The similar objects are then modelled into a graph structure, which is then partitioned into highly connected sub-graphs of related crimes. A Monte Carlo approach and adaptive graph-size contraction heuristics are employed to amplify the algorithm success. In addition, two new interest measures were considered to augment the analysis of prevalent pattern:(i) Proportion Difference Evaluation (PDE), which reveals the propagation effect of a series and identifies a dominant series; and (ii) Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series. Heat maps (Google map application programming interface) are then used to enhance visualisation of locations where series activities are prevalent. The research shows promise with the generated patterns, which was substantiated with the optimistic reaction and input received from experts in this domain.

1.8 Contributions and Outline

In addressing the research questions, the following are the anticipated contributions from this research:

- Conceptualisation of a crime series mining approach to promoting a "sustainable safe and smart" city of the future in a resource constrained setting.
- The development of a crime mining model, CriClust, augmented with a dual-threshold scheme, which applies established theoretical concepts from clustering (highly connected sub-graph and similarity ranking) to derive useful evidence to security agencies.
- Visual display of clusters and patterns for prompt identification of important areas demanding attention.
- The empirical analysis of the model, in order to ascertain what level of confidence can be expressed in such a model in terms of correlation level of results obtained, and performance measure to determine its usefulness in realistic high volume crime analysis.
- The comparison of features of the CriClust model with existing models, in order to clearly outline some of their similarities and differences.

To the best of the researcher's knowledge, a dual-threshold scheme with geometric projection, and highly connected sub-graphs with adaptive graph size heuristic have never been hybridized for crime series mining purpose. It was anticipated that making the highlighted contributions, police in resource constrained settings such as in developing nations, in particular, can be empowered to be more proactive in tackling crime and will be able to focus their resources where it matters most. Moreover, in our evaluation CriClust led to the identification of up to three series at some of the locations investigated in a dataset of 5500 rape crime records across 40 locations (suburbs) in Western Cape.

1.9 Declaration of Recent Publications

Some ideas and figures in this thesis have appeared in the following recent articles published from the research work.

Refereed Book Chapter

- Isafiade O.; Bagula A.; Berman S. (2016) On the Advancement of using Data Mining for Crime Situation Recognition. *In Data Mining Trends and Applications in Criminal Science and Investigations.*.pp 1-31. IGI global, USA.
- Isafiade O.; Bagula A.; Berman S. (2016): On the Use of Bayesian Networks in Crime Suspect Modelling and Legal Decision Support. *In Data Mining Trends and Applications in Criminal Science and Investigations.*. pp 143-168. IGI global, USA.

Refereed Conference Publications

- Isafiade O., Bagula A, Berman S.: A Revised Frequent Pattern Model for Crime Situation Recognition Based on Floor-Ceil Quartile Function. In *proceedings of the Information Technology and Quantitative Management (ITQM 2015),* Procedia Computer Science, pp 251-260, Elsevier, Science Direct, July 2015, Rio-De Janeiro, Brazil.

- Isafiade, O. and Bagula A. CitiSafe: Adaptive Spatial Pattern Knowledge Using Fp-growth Algorithm for Crime Situation Recognition. In *proceedings of the IEEE International Symposium on Ubiquitous Intelligence and Autonomic Systems (IEEE-UIAS)*, PP 551-556, December 2013. Italy.

- Isafiade Omowunmi and Bagula A. Efficient Frequent Pattern Knowledge for Crime Situation Recognition in Developing Countries. In *Fourth Annual Symposium on Computing for Development (ACM Dev4)*. December 6-7, 2013. Cape Town, South Africa.

- Isafiade Omowunmi and Bagula A. Towards Citizen-Centred Safety Enhancement Framework: A Smart City Application. In *Proceedings of the SAICSIT-IoT 2013 workshop*, 7th October, 2013. East-London, South Africa.

Poster Presentation

- Isafiade Omowunmi. Ubiquitous Intelligence for Smart Cities: A Public Safety Approach. *In Poster session at 3rd Heidelberg Laureate Forum (HLF)*, August 2015, Heidelberg, Germany.

1.10 Thesis Outline

- *Chapter 2: Research Background and Related Research*

  This chapter provides a general background to crime mining and highlights different research areas that have been explored for crime mining purposes. Ways of integrating crime data sources for tactical analysis, as highlighted by IACA are presented. A critical review of existing research in this domain of interest and their current and potential limitations where applicable is presented. An overview of existing data mining techniques is presented using a high-level tabular representation. Features of existing applications and techniques, such as model selection, exploratory basis and algorithm advancement are compared. Previous research that relates to cluster and crime series identification are also explored. The chapter concludes with a summary of gaps and opportunities that have been identified in previous research.

- *Chapter 3: Model Formulation and Design Methodology*

The formalism of the model that supports crime series identification is presented in this chapter. A dual-threshold scheme (significance and prevalence thresholds) is used to augment the process analysis. How the crime series identification procedure was implemented is summarised. The mathematical concepts and properties of the techniques adopted are also presented in order to demonstrate the underlying principles of these models as employed in current research. Furthermore, the model formulation and the system overview followed by a simple analogy and algorithmic procedure of the implementation strategy for the model is presented. The chapter concludes with a brief description of the evaluation mechanism for the model.

- *Chapter 4: Quantitative Results and Experimental Evaluation*

  In this chapter, the experimental results obtained from the quasi-real crime data are presented. How the results were generated is summarised, followed by selected graphical views of generated series at certain locations. The reliability of the model is further confirmed through the derived results from the controlled experimental data, which reveals the level of correlation and rate of clusters generated. The usefulness of results obtained by introducing the two interest measures (PDE and PSE) are revealed. Furthermore, a benchmarking experiment using common clustering techniques is presented. This chapter ends by comparing common series detection and data mining models with our proposed CriClust model.

- *Chapter 5: Conclusion and Future Research*

  This chapter commences by restating the research problem and the research questions. A synthesis of how the research questions were addressed is presented. The chapter concludes the research and provides a motivation for future potential research in crime data mining and crime series identification, by presenting open research issues.

1.11 Chapter Summary

The discussion in this chapter presents significant facts about the status of crime in developing nations, using South Africa as a case study. It further offers a general overview of the study, namely the motivation for the study and needs for proactive and efficient crime analysis methods within the society as an intelligent crime situation recognition and management approach. As the term “smart city” is gaining prominence across different domains,

there is some variety in what different researchers or smart city developers understand under this label. However, the goal of a smart city is largely focused on deriving ways to improving the general well-being of citizens as well as services rendered to them. While most smart city initiative places much emphasis on the use of modern technology as a necessity for fighting organised crime or other forms of crime, it is important to stress that it may not be sufficient. From a strategic and tactical solution perspective, crime information archived by public safety authorities could be mined for useful knowledge, which can assist in fighting crime proactively. This can further assist in achieving a sustainable "safe and smart" city.

However, while much research has been conducted on crime mining, for example, in the area of hotspots and spatio-temporal related research, there is a paucity of research in crime series identification particularly in developing nations, despite its critical importance for public safety improvement in a smart city development. Moreover there is room to develop more heuristics and hybrid models for effectively analysing crime information in identifying related criminal offences, referred to as crime series. Hence this research focuses on crime series data analytics, for improved public safety outcomes. An extension in this work is the benchmarking phase, where existing techniques are evaluated.

Chapter 2

BACKGROUND AND LITERATURE REVIEW

This chapter re-establishes the challenge of crime and provides a general background to crime mining as a promising way of deterring crime. The chapter provides an extensive review of related work in this domain of interest in order to gain an understanding of how previous findings have contributed to public safety improvement while emphasising the lack of adequate attention to these processes in identifying (potential) crime series for public safety improvement. Emerging research efforts in identifying crime series is presented. Different research areas that have been explored for crime mining purpose are highlighted and ways of integrating crime data sources for tactical analysis, as conceptualised by the international association of crime analysts (IACA), are presented. The mathematical background to the mutual information and highly connected sub-graph models is also documented in order to demonstrate the underlying principles of these models as employed in current research.

2.1 Background Study and Related Research

There is no doubt that crime costs our societies dearly in several ways, and therefore needs to be curbed or eliminated [1]. Crime is not entirely committed in a random manner and may not necessarily occur in a consistent manner either. Its execution is either carefully planned or opportunistic [115]. However, It is important to recognise that particular places or objects are potentially crime attractors and generators, as implied by the two long-established theories of environmental criminology associated with crime [18]. These theories are: (i) Crime Pattern Theory (CPT), and (ii) Routine Activity Theory (RAT).

- Crime attractors and generators refer to sites, properties, objects, and locations which offenders are very familiar with and offers several criminal opportunities. For example, due to the crime incidents that happen there i.e drug dealing locations, public transport terminals, a cul-de-sac area with narrow escapes, shopping malls, to mention a few.
- Crime detractors, on the other hand, are locations that discourage crime perpetrators and criminal activities. For example, a steady business activity or regular natural surveillance could have such a positive impact on a particular location.

The theories of environmental criminology points to implicit and explicit spatial and temporal attributes or factors that heighten the trend in crime events or patterns [18]. The RAT theory depicted in Figure 2.1, stipulates that a crime can only be committed when three specific criteria are involved. These criteria are that there must be a motivated offender, a suitable target or vulnerable victim (available at an appropriate location and time), as well as the absence of a capable guardian. RAT is a sub-field of crime opportunity theory that focuses on situations or circumstances of crimes. Similarly, CPT seeks to explain the reason crime occurs in certain locations with a high intensity. It stipulates that crime occurs when the activity space of a target or victim meets with the activity space of an offender. Activity space in this context refers to locations one finds themselves in everyday life. For example, school, work, home, shopping malls.

These two theories, RAT and CPT, are fundamental in understanding crime trends and criminal activities. The recognition of such fundamentals provides insights into a list of factors to consider when trying to come up with

intervention strategies using analytics. Crime data mining could be used to determine what the suspect's next move could potentially be and the information derived subsequently can be used for crime deterrence.

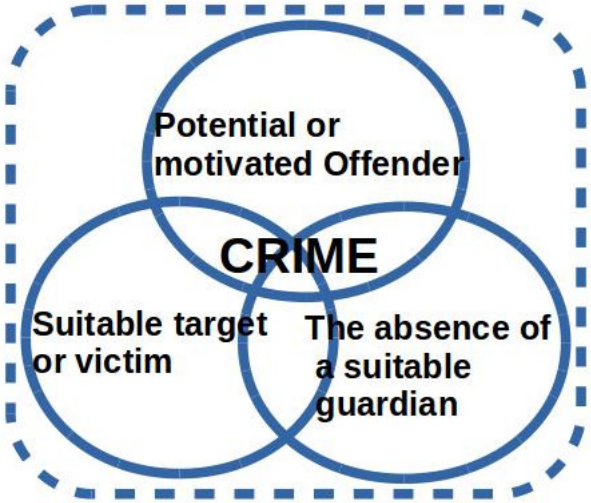


Figure 2.1: A model of routine activity theory (RAT).

2.2 Overview of Crime Data Mining

Crime data mining has become a well researched topic with the advancement in technology, and owing to the huge amount of crime information archived by public safety authorities [19, 44]. In the wake of resource constraints but better technology solutions, public safety agencies increasingly depend on viable targeted interventions, and they do more with less. Furthermore, the international association of crime analysts (IACA) [78] has presented unique ways of exploring crime patterns. Table 2.1 reveals the manner in which crime patterns are conceptualized according to IACA [78]. A proper understanding of these crime patterns is necessary if crime reduction and criminal apprehension targets are to be achieved through crime data mining. It is important to note that while these patterns are explained individually in Table 2.1, they sometimes interconnect.

According to a report [71] about citizens' opinion on what they consider as a top priority for dealing with crime, "prevention" takes the lead with the highest percentage of 43 %. Figure 2.2 presents the statistics on the PREPU (Prevention, Rehabilitation, Enforcement, Punishment, Unspecified) model of dealing with crime [71], revealing responders' opinion on what they consider as a top priority for crime deterrence. It makes more sense to dedicate

Table 2.1: Description of existing crime pattern types for tactical and strategic analysis.

Crime Pattern Types			
S/N	Pattern Classification	Description	Strategic Analysis
1	Hot Spot	A set of related crimes committed at locations within close vicinity of one another.	Involves analysis to include target location characteristics.
2	Hot product	A set of crimes committed by one or more criminals and targeted at a specific product or item.	Focuses analysis on types of stolen property or product.
3	Hot place	A set of related crimes perpetrated by one or more persons at the same location.	Include spatial analysis to include location types and properties.
4	Hot setting	A set of related crimes committed by one or more culprits that are essentially connected by the type of location where crimes took place.	Involves spatial analysis to include attributes or characteristics of the target environment.
5	Hot prey	A set of crimes committed by one or more persons on victims sharing related physical characteristics or behavior.	Involves querying data to identify common or repeated victim characteristics or behavior.
6	Spree	A kind of crime series associated with a short time frame, high frequency, and almost continuous criminal act.	Focuses on analysis to include temporal properties and offender behaviour.
7	Series	A set of related crimes considered to be committed by the same culprit or group of culprits.	Involves analysis to include offender behaviour or profiling characteristics.

resources to "prevention of crime" rather than consume resources on its aftermath. However, having a good knowledge of past or current trends, which crime data mining offers, is a smart way of dealing with crime and would serve as a guide to the best prevention strategies to consider or deploy.

Crime analysts and researchers have put much effort into making sense out of the "non-sense" in the available raw crime data documented by security agencies [2, 44, 101, 103, 106, 107]. Consequently, this has stirred crime analysts and researchers to respond to the six major questions associated with crime:

1. *Why* does crime occur (motives, loopholes, attractors)?
2. *When* does crime happen (temporal information- season, day, hour)?
3. *What* kind of crime is committed (contact crimes, contact-related crimes, property-related crimes)?
4. *Who* is affected by or perpetuating crime (offender behaviour and dynamics, victim characteristics)?
5. *Where* does crime happen (location features, hotspots)?
6. *How* is the crime committed (modus operandi, weapons used)?

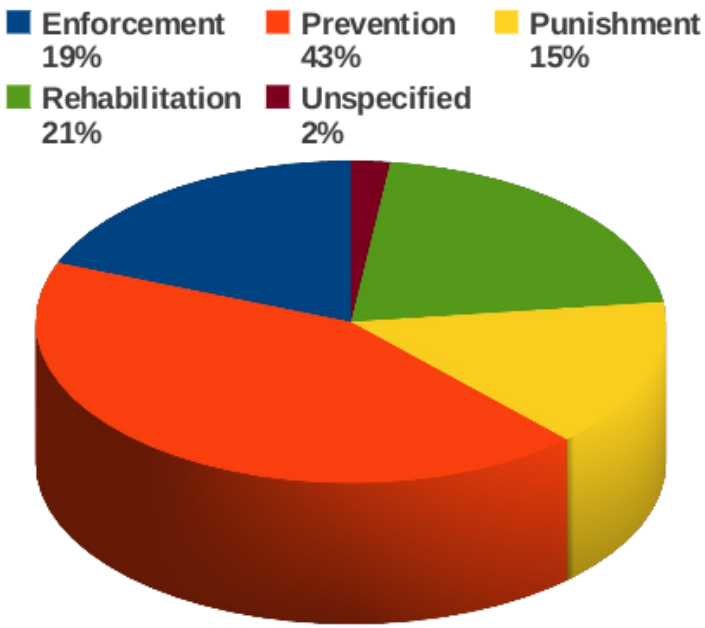


Figure 2.2: Citizens opinions on top priority for dealing with crime.

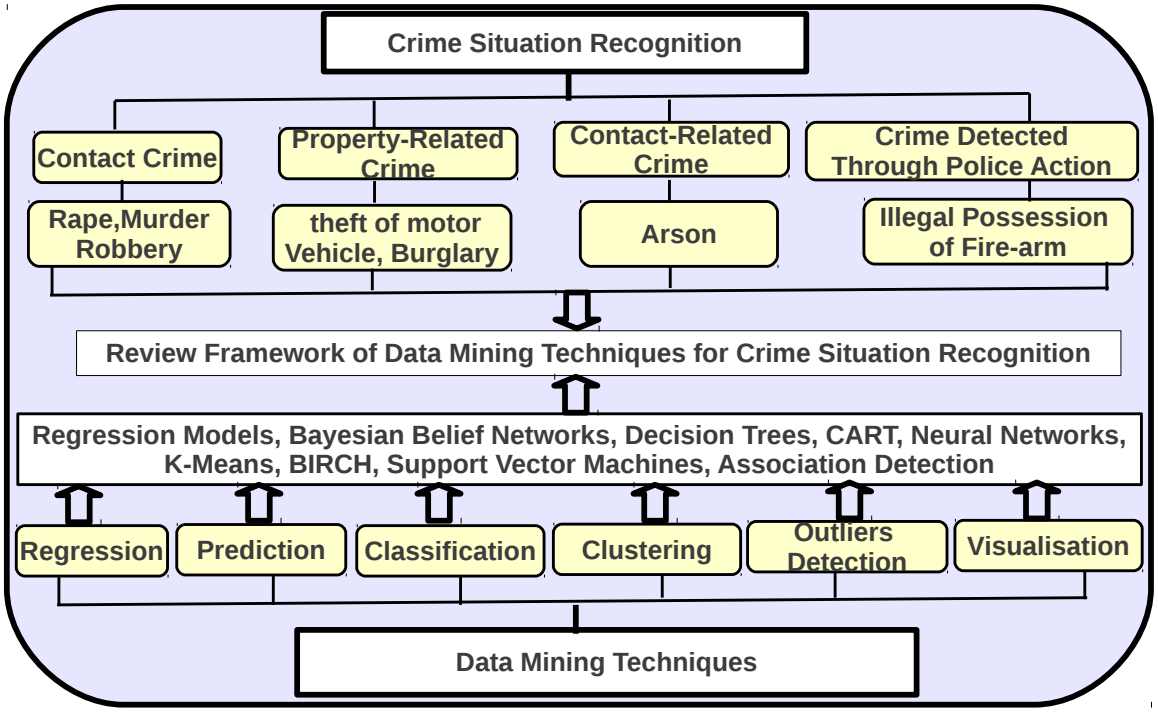


Figure 2.3: A review framework of data mining approaches to crime situation recognition.

Owing to the different crime related questions to be addressed a lot of research, discussed hereafter, has been developed to tackle each of these questions using different means and techniques, and drawing on different existing theories from criminal science [25]. A number of techniques and approaches have been considered for crime data mining. Figure 2.3 presents a general review framework of data mining approaches to crime situation recognition. Some research work focus only on one question per time, while some others address two or more of the six. Crime data mining can be achieved through two major techniques or approaches, these are descriptive and predictive approaches. The details of these approaches and sample techniques available for crime mining purposes are further summarized in Figure 2.4. More generally, the choice of approach or technique would depend on target analysis and goal of crime investigation.

Crime mapping and modeling activities by security agencies and criminology researchers, were not aided by technology until late 1970s. Prior to late 1980s, geographic information science (GIS) was not well researched [58]. However, a major transposition occurred when the great potential of spatial techniques to map and discover crime trends was realised [15]. In recent times crime mining has revolutionised into adopting hybrid and advanced

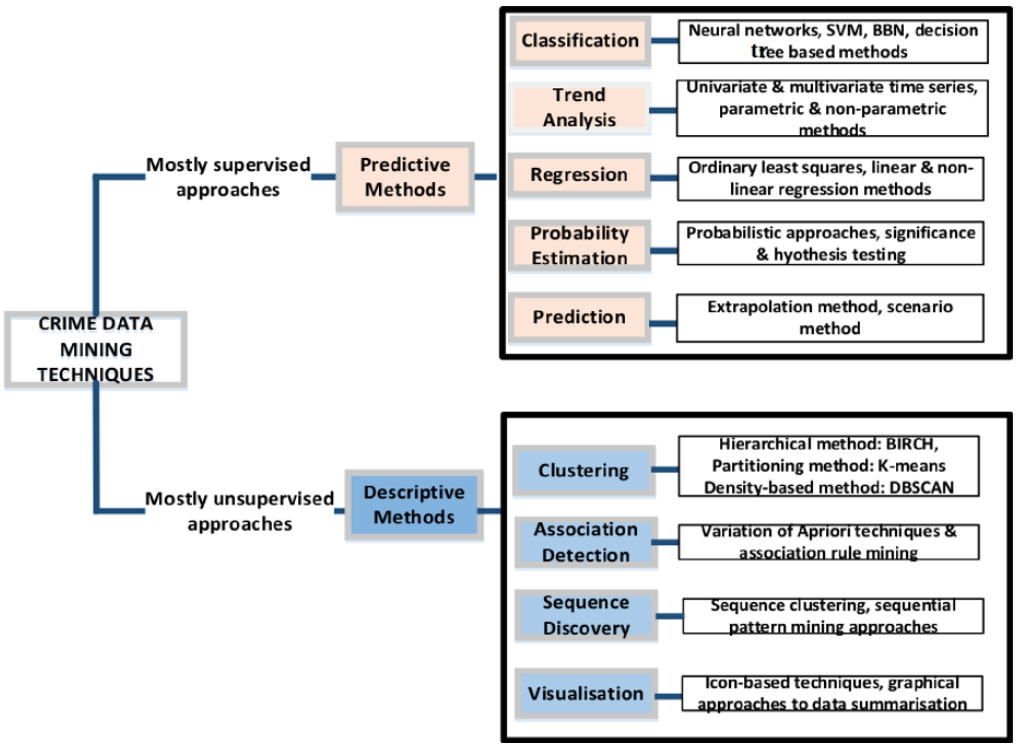


Figure 2.4: A taxonomy of crime data mining tasks and techniques.

crime mapping, modeling and mining techniques.

Considering varying goals, objectives and context, GIS has contributed immensely in addressing the "*where*" question. For example, GIS is a major contributor in predictive policing and hotspots crime analysis [45]. While, by extension, GIS tries to derive solutions to the other five questions, it however does not directly provide deeper insight pertaining to the "why, when, what, how and who" of crime events [8]. For example, it does not have the full capability to determine what set of crimes are committed by the same individual or group of offender(s) at certain locations. Thus research in these areas continue.

2.2.1 Classical Sequential Approach to Data Mining

Crime data mining is a powerful tool that gives predictive insight into probable crime events or activities. In order to achieve better quality results in data mining, data preprocessing activities are usually adopted [55]. The preprocessing step is essential for dealing with inconsistencies and missing data in the dataset. Figure 2.5 presents

the general iterative steps in data mining tasks. Established data mining methodologies, such as SEMMA (Sample, Explore, Modify, Model, Assess) and CRISP-DM (Cross-Industry Standard Process for Data Mining) [80], also adopt these general steps.

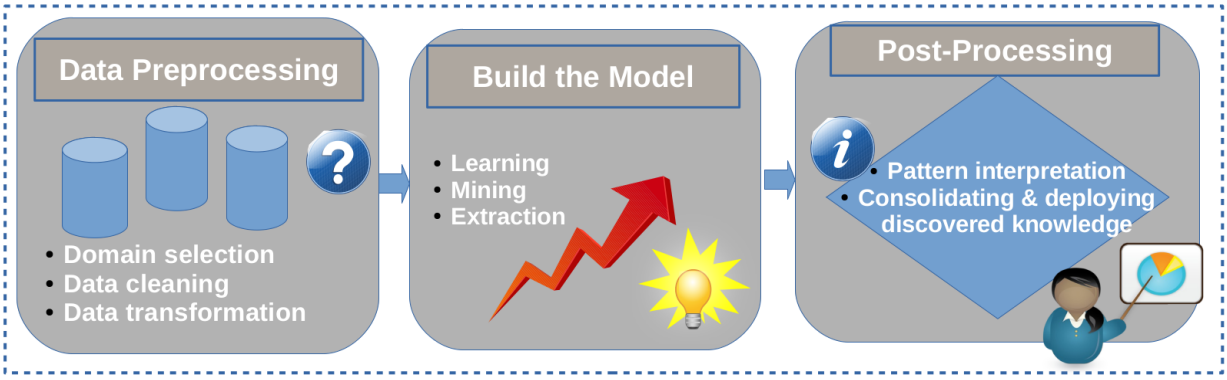


Figure 2.5: General iterative steps in data mining.

- Pre-processing Phase: The pre-processing steps or activities include domain selection, data cleaning and data transformation.
  - Domain selection: This involves the prioritisation and selection of one or more domains for which an analysis is to be initiated. For example, the selection of a contact-crime domain (e.g. Robbery, Rape). This is followed by the selection of data (attributes) that are relevant to the analysis.
  - Data cleaning: The data cleaning step involves smoothing noisy data, dealing with missing values, detecting and removing outliers and resolving inconsistencies.
  - Data transformation: This step involves the conversion of the data into an organised set, such as merging of data from multiple sources and putting the data in the correct form for a specific mining technique. Attribute selection, dimensionality reduction and data normalisation are also done at this stage.
- Build the Model: Pre-processed data from the previous step is analysed in this phase. Building the model entails the selection and utilization of one or more of the techniques highlighted in Figures 2.3 and 2.4 to achieve specific crime deterrence objectives. It has been noted that one of the major challenges in data

mining is the determination of the technique(s) best suited to a given problem [113]. The choice of method selection at this stage will depend on the type of problems to be addressed and expected results to be generated, for example, whether predictive or descriptive or both.

- Post-Processing Phase: The post-processing phase would involve consolidating the discovered knowledge and making it available, accessible and understandable for experts and non-experts in the public safety domain. The predictive or descriptive information obtained from the analysed crime data is summarized at this level. This presents understandable structures, patterns or trends to decision makers.

2.3 Current Implementation to Crime Data Mining

2.3.1 Existing Strategies in Crime Data Mining

Crime data mining emerged as a result of different forms of knowledge that could be derived, which could assist decision makers and public safety authorities in improving their efficiency [68]. Exploratory data analysis (EDA) is a strategy for analysing data sets, in order to summarize their major attributes or characteristics. While much research has been conducted on spatial crime data analysis, temporal crime data analysis, and spatio-temporal crime analysis [47, 54, 64, 72], other motivations for exploring crime data include identifying crime series or causal relationships among crime variables or attributes [106]. For instance, this can determine what crime types affect a particular gender group or what the age distribution for a crime type is, or identify a crime series and serial offenders. The three most common strategies for exploring crime data mining, discussed hereafter, exploit:

- Spatial knowledge
- Temporal knowledge
- Spatio-temporal knowledge

1. **Spatial Crime Data Analysis:** Spatial data analysis employs statistical analysis techniques that address certain problems relating to spatial characteristics of data, such as spatial heterogeneity and spatial dependency (auto-correlated patterns and structures) [41]. A relevant scenario is the identification of crime hot

spots or clusters of homogeneous crime. Ferreira et al. [45] reports the usage of hot spot analysis as being the most common approach to analysis in crime detection. The inherent geographical property of crime is a natural inspiration for spatial data analysis. A criminal must have come from a place and must necessarily commit or perpetrate a crime at a particular location. Thus, the emergence of sophisticated tools such as the Radio-Frequency Identification (RFID) and GIS combined with rigorous statistical approaches have paved the way for spatial data analysis [11].

GIS has the ability to integrate spatial data and information from numerous sources into one interface. This ability has triggered advanced spatial analyses that would not have been realisable or at least very difficult to achieve without the GIS. Wang *et al.* [104] proposed a spatial mapping tool, called Hotspot Optimization Tool (HOT), using a geospatial discriminative patterns notion. This can assist with decisions regarding the identification of safest path for road users. Possible or prevalent crime areas can also be determined using geographic profiling. In South Africa (SA), geographic profiling has been considered for determining possible or prevalent crime locations. In SA, social-economic factors including over 74 crime variables and 250 census variable were used to identify 20 different categories from census data [8]. The categories were then used to prioritize intervention in each police station. In addition to creating awareness for the individual stations, the categories also provide insight into the factors that cause these crimes.

2. **Temporal Crime Data Analysis:** Crime temporal analysis methods deal with prediction or summarisation of crime temporal attributes. For example, what crime type occurs at a particular time of the day and what crime categories are committed together at a particular time. It generally leads to time dependence or temporal variations analysis, with the goal of forecasting, which is usually the end result of temporal analysis. It can be used to predict when, where and how future incidents in a crime series, trend or pattern will occur.

   Temporal Association Rule (TAR), a form of incremental mining of temporal association rules, has gained much attention over the years. However, a major draw back of TAR is the requirement of multiple database re-scanning. In contrast to pioneering works on incremental mining, developed by Cheng et al. [16], which mostly focused on reducing the frequency of database rescanning, Vincent *et al.* [100] presented a time-series based analysis, *ITAR*, that uses a negative border method to address the problem of database re-scanning.

The negative border method is used to speed up the updating of association rules when new data is added to the database. This avoids the costly re-scanning in the previous research. Slightly different research on temporal analysis was the work of Ashby *et al.* [4], which detailed a comparison of methods for temporal analysis of indeterminate crimes.

3. **Spatio-Temporal Crime Analysis (STCA):** Spatio-temporal analysis scenarios or methods consider both space and time in their analytic processes, for example, the PredPol technology [86]. This type of analysis is based on coordinates of events, such as locations of crime activities, and the time frame in which such criminal acts occurred [80, 72]. STCA attempts to address the questions of "where" and "when" the crime occurs. The uniqueness of STCA is evident in cross- or auto-correlation between variables and parameters, where time and space variables carry much weight during crime analysis [103]. Effective production of distance measures or metrics for spatial and temporal conditions is a common feature of spatio-temporal analyses. STCA can give clues as to what specific paths are safe at what period of the day, bearing in mind that shortest paths are not always the safest paths.

2.3.2 Existing Applications and Techniques for Crime Data Mining

Over the years, considerable effort has been put into improving the performance of existing algorithms [42, 54, 64, 116] and inventing new algorithms for crime data mining and analysis [59, 84, 105]. Thus, numerous projects and applications have been developed in this domain of interest [41, 69, 70]. Conventional techniques of data mining such as association rule mining identify patterns in purely structured data sets. However, contemporary approaches to data mining often identify patterns in both structured and unstructured data sets [15]. Some of these techniques are discussed hereafter.

Apriori is a classic technique for learning association rules and one of the most influential techniques in exploring patterns of interest in a dataset [10, 118]. It uses a bottom-up approach for expanding frequent subsets (i.e. sets of items that frequently occur together), one item at a time; a step known as candidate generation. A simplified example of an association rule could be as follows:

$$\text{robbery in location } K \Rightarrow \text{murder in location } K$$

Table 2.2: Support and confidence quantity of an association rule.

Rule	Support (P $\Rightarrow$ Q) $= \frac{\lvert Q \cap P \rvert}{\lvert \text{tuples} \rvert}$
	Confidence (P $\Rightarrow$ Q) $= \frac{\lvert Q \cap P \rvert}{\lvert P \rvert}$

This could mean that a serial offender or group of offenders (gang) involved in robbery, in location K, ended up killing their victims. However, it is very important to recognise that association rules depend on the quality of the results generated from frequent crime item-set mining, for example, FP-Growth [36]. FP-Growth model is a frequent pattern growth model that can be used to identify crime items (incidents) that are frequently committed together [42].

Apriori has been noted for easy parallelisation and simple implementation [118]. However, a major drawback of this technique is that it requires repeated database scans, which is inefficient when the database cannot be memory resident. Consequently, several variants of the algorithm, such as AprioriTid and AprioriHybrid, emerged over the years to address its shortcomings [100]. AprioriHybrid combines Apriori and AprioriTid to amalgamate the strengths of both algorithms. The variants of association rule mining address inefficiencies through better pruning, refinement techniques, data structures, partitioning techniques and search strategies. Association rules present facts about the dependency or relationship between two or more items. It was initially motivated by market basket analysis, to identify items that are usually purchased together [118], but can be extended to other relevant fields such as crime mining [41], to identify crimes that have been committed together or at the same location. An association rule must satisfy two major factors:

1. Support (say α), which indicates the frequency of co-occurrence, must be sufficient.
2. Confidence (say γ), which indicates the correctness or applicability of the rule, must also be adequate.

The support and confidence measure is given in Table 2.2. Thus association rule is a two-step process; first find all the frequent itemsets, then generate strong association rules from these sets.

In addition to Apriori, other algorithms such as FP-growth and Eclat have been presented for generating association rules. The FP-growth algorithm, introduced by Han *et al.*[36] mines frequent patterns without generating

candidate itemsets. First it builds a compact data structure named FP-tree, which is later traversed using a bottom-up approach, to extract frequent itemsets. The CitiSafe application [41] employs the FP-growth algorithm for crime data mining. An improvement made to the algorithm was the *batch-merge paradigm* adopted, which simultaneously generates all the frequent pattern trees for a particular crime itemset, since the pattern trees are independent of one another, and then ultimately merge the results, as opposed to the conventional sequential approach adopted in the FP-growth algorithm. The modelling technique successfully identified crime patterns from a fairly large number of crimes, as opposed to the limitations of current manual analysis. CitiSafe provides an effective and efficient crime incident analysis tool for law enforcement agencies and public safety organisations.

The COPLINK application [15] is one of the early large-scale projects in crime data mining. It is a project on crime database integration that was launched with the Tucson Police department and funded by the National Science Foundation and National Institute of Justice. Khan *et al.* [54] presented four case studies to emphasize the importance and relevance of the COPLINK project. These are: automatic entity extraction from police narrative reports; using a neural network based entity extractor for deceptive criminal identity detection; identifying false witness in the database through criminal network analysis; identifying subgroups and key members in a criminal network and authorship analysis in cybercrime. In the area of predictive policing, PredPol has been developed in order to identify and prevent potential crime incidents [86]. PredPol only predicts "where" and "when" a crime incident is most likely to happen. As no personal information about individuals or demographics are utilized in making these predictions, PredPol does not predict a potential offender or victim type. Therefore, PredPol's crime prevention technology does not pose any personal privacy or profiling concerns and has proven useful in place-based crime deterrence in the USA.

Techniques such as clustering and fuzzy apriori approaches have been adopted in crime pattern discovery [10, 47]. A detailed study on the use of clustering techniques in crime data mining was presented by Krishnamurthy et al [55]. Clustering techniques sort data into specific classes based on certain features or functions, thereby maximising their intra-class similarity [22, 47]. K-means is a common unsupervised clustering algorithm, which partitions data items into a predefined number of clusters [96, 65]. The algorithm aims to minimise the objective function in Equation (2.1), a squared error function.

$$J = \sum_{i=1}^{k} \sum_{j=1}^{l} \|l_j^{(i)} - c_j\|^2. \tag{2.1}$$

Where $\|l_j^{(i)} - c_j\|^2$ is a chosen distance function between a data point $l_j^{(i)}$ and the cluster centre c_j, is a measure of the distance of the data points from their corresponding cluster centres. K-means has been identified with a number of limitations, namely:

- High sensitivity to initialisation (choice of K items as first set of centroids)
- Problems with clusters of varying sizes
- Problems with clusters of varying densities

However, it remains a baseline to most clustering algorithms, and research has shown that multiple runs can help solve the initial condition (centroid) problems [47]. Moreover, some other research proposed statistical algorithms that can be used to identify an appropriate value to use for "K" [82]. Clustering algorithms could also be computationally expensive. Hence, several improvements on clustering techniques are emerging. The K-means algorithm is a special case of the Expectation Maximisation (EM) algorithm [66]. This special case also points to the fact that almost, if not, all aspects of K-means have been improved in the following areas [116]:

- Initialization
- Efficiency
- Centroid and objective definitions
- Overall process enhancements
- Distance measures

A hybrid clustering model that combines the advantages of K-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was proposed by Malathi *et al.* [65]. DBSCAN uses the notion of density accessibilitty (closeness to the δ-neighbourhood) to define a cluster. Essentially, a point j is directly density-accessible from a point k if it is not farther away than a given distance δ. A major difference between DBSCAN

and K-means is that DBSCAN does not require an initial specification of k, the number of clusters in the data, as opposed to K-means. Instead, DBSCAN requires the minimum number of points, say ϵ, needed to form a dense region. Experimental results in the work of Malathi et al [65] indicated that DBSCAN outperforms the K-means algorithm. A similar algorithm to DBSCAN is OPTICS (Ordering Points To Identify the Clustering Structure), which is also used for identifying density-based clusters in spatial datasets [3].

Another notable contribution is the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm by Zhang *et al.* [117]. BIRCH implements a scalable clustering algorithm with respect to quality of clustering and number of objects. It utilises local information in which clustering decisions are finalised without multiple data scans, and outputs a dendrogram, which encodes the clustering information of the dataset in a compact manner.

A related effort to BIRCH, with a slight modification in distance measure, was proposed by Chiu *et al.* [17]. They proposed a probabilistic model which derives a distance measure, where the distance between two clusters is equivalent to the decrease in the log-likelihood function resulting from merging. Their algorithm further implemented a procedure which allows for automatic determination of the appropriate number of clusters and the assignment of cluster membership to noisy or incomplete data.

A strategy for crime detection and criminal identification (CDCI) was proposed in the work of Taya *et al.* [96]. CDCI implemented data mining techniques consisting of six modules, which involve data extraction and preprocessing, clustering using K-means and classification using K-Nearest Neighbour (KNN) algorithms. KNN is a machine learning technique based on a non-parametric approach, for regression and classification [65]. CDCI identifies similarities among different crimes in the feature space and organises them into predefined clusters, for classification and prediction. CDCI was tested using crime data, between years 2000-2012, from seven different Indian cities with high crime rates and the WEKA software was used for verification and validation of the results from K-means. Clusters were defined based on two crime attributes; the year a crime is committed and frequency of a crime. Visual aids to K-means were enhanced using Google Map Application Programming Interface, which provides an effective crime analysis visualisation tool for crime investigating agencies.

The information (sensor) fusion capabilities, extracting information from several data sources in parallel, that is

inherent in the Artificial Neural Network (ANN) technique has attracted the attention of several researchers. Recent research on crime mining [46] utilised the back propagation Neural-Network (BPNN) classifier and compared its performance with a data association algorithm in tracking criminals automatically. In the BPNN classifier algorithm, five input nodes, eight output nodes and two hidden layers are considered. While the eight output nodes correspond to different classes of crimes such as violent crime, burglary, property-related crime, the eight input nodes are identifying features such as eye, age, frame and complexion. The association algorithm on the other hand uses a priori knowledge of frequent items set to identify emerging patterns. It uses an iterative level-wise approach, where K-item-sets are used to explore $K+1$-item sets. The aim was to identify and extract correlations from the crime incident summaries. Empirical results from the research reveal that the back propagation network classifier outperforms the data association algorithm, in terms of identifying and tracking potential criminal and predicting future crimes based on the receiver operating characteristics.

A slightly different approach of applying ANN is seen in the work of Helbich *et al.* [39], but a similar notion has been presented by Keyvanpour *et al.* in [53], where important entities were extracted from police narrative reports, written in plain text, and a self-organising map (SOM) clustering technique used for crime analysis, and ultimately a crime matching procedure. In the former, Helbich *et al.* [39] applied a SOM to text mining and focused on unstructured crime narrative reports. SOM is a kind of artificial neural network trained to derive a low-dimensional discretised representation of the input space, using unsupervised learning. The novelty in their research is a multi-step methodology that projects the SOM clustering output from the unstructured crime data to geographical space. Point analysis was used to explore the spatial distribution of the clusters. Their research further demonstrates that collective surveillance by individuals, through e-mails, personal interviews, written statements and telephone recordings, are powerful avenues to explore in crime analysis and prediction.

Earlier research proposed the CATCH (Computer Aided Tracking and Characterisation of Homicides) and CATCHRAPE application software [48]. The CATCH software uses the artificial neural network (ANN) system for classification of criminals in rape and murder cases and provides a means of clustering related cases that can be attributed to the same offender. The assumption was that two different individuals with similar characteristics are capable of committing crimes with similar behaviour. SOM was used for clustering, where no initial training data was provided. SOM provided a convenient visual representation by organising the data vectors into similar data clusters

on a two-dimensional map.

Bayesian belief networks (BBN) have been employed in different ways to establish probabilistic models in criminal science and investigations [43]. Appropriate selection of a scoring function, a structure learning algorithm and a parameter-learning algorithm are critical for the BBN to be successful. Riesen *et al.* [89] proposed a Bayesian network classifier to predict victimisation. The K2 and Hill climbing algorithms were employed for prediction and optimisation respectively. In the structure learning algorithm, a local score based structure learning, tabu search, was utilised. A major trade-off in their approach was between an increase in memory usage and a faster learning time. They concluded that only the K2 algorithm successfully built a model with more than three parent nodes. A related effort [5] addressed the challenge of space complexity by modifying the greedy search K2 algorithm, through the inclusion of a-priori conditional independence relations amongst the nodes, and achieved a 15% improvement on the built model. Moreover, Neill et al. [76] proposed a model for space-time event detection using a multivariate Bayesian scan statistics approach for timely event detection and characterisation. A recent advancement on the use of Bayesian network [119] investigated how to model crime linkages using a Bayesian network. While their research presents a crime linkage model for identifying the interesting aspects of the reasoning underlying crime linkage, it has the limitation of not capturing all the problems or attributes, such as relevance of trace material and many other uncertainties that are important to consider, that play a role when linking crimes.

Support Vector Machines (SVM), a set of supervised learning methods for regression and classification, have also been employed in crime mining. A recent work on cybercrime detection and prevention models to reduce data damage by malicious code, proposed a hybrid model called *AdaSVM* [23], which combines special features from the Adaboost algorithm with SVM. AdaBoost [88] is a classification technique that combines predictions from several classifiers and generates a single and robust classifier, which is very often more effective than individual classifiers. A Facebook dataset was used to evaluate the performance of the model.

Several applications and tools have been developed for spatial data mining [11]. A common scenario for spatial data mining is the identification of crime hotspots. Existing hotspot mapping techniques include thematic mapping, point mapping and kernel density estimation. Hot Spot Analysis (HSA) and Hotspot Optimisation Tool (HOT) are examples of existing hotspot mapping applications or tools. Wang *et al* [104] claim that the HOT application

differs from other hotspot mapping tools, because HOT considers spatial crime and socio-economic related factors in addition to focusing on target crime density.

From the literature, it is clear that research in crime situation recognition, prediction and control is ongoing, as several algorithms and approaches have been implemented and even hybrid models have emerged. However, the daily increase in crime information still necessitates more effective and advanced techniques for crime analysis.

2.4 Comparison of Crime Data Mining Applications and Techniques

We compare algorithm selection, content selection, exploratory basis and result summarisation of existing crime data mining techniques and applications in Table 2.3. While the *exploratory basis* presents the problems addressed in each, the *model selection column* describes the major algorithm used in addressing the problem(s). *Algorithm advancement* points to any improvements that have been made on the chosen model and *content selection* describes the nature of data used for the experiment and/or implementation. The last column, *result summary*, gives the mode in which the results of the experiments are presented. While *qualitative* depicts analysis results presented with graphical views or visualisation, *quantitative* refers to results presented mostly in table format, specifying corresponding crime results or level of coherence in patterns derived.

Table 2.4 highlights the characteristics, strengths and weaknesses of some of the currently used techniques or models. It is evident that currently adopted techniques and models have requirements (e.g. appropriate parameter selection) and characteristics that determine their degree of relevance and efficiency.

Tables 2.3 and 2.4 give an insight into the open research issues for future consideration. While there has been some effort in improving many facets of existing approaches and techniques [54, 64, 101], it is clear that there is room to develop more hybrid models [23] and richer spatial and/or temporal approaches to enhance crime situation recognition and ultimately improve the productivity of public safety authorities.

Table 2.3: Comparing features of current crime data mining applications.

S/N	Features	Exploratory basis	Model selection/ Technique(s) adopted	Algorithm advancement	Content selection	Result summary
1	Crime-terrorpattern detection [54]	Detect crime spree by the same culprit	Clustering based (K-means)	N/A	real crime data	visualisation (qualitative)
2	CitiSafe Algorithm [41]	Spatial pattern & crime spree	Frequent-pattern based (FP-growth)	Batch-merge paradigm	real & Synthesised data	Qualitative & quantitative
3	Predicting victimisation [89]	Victimisation classification	Bayesian belief network (K2 & hill climbing)	N/A	real crime data (NCVS)	Quantitative
4	Identify change in crime patterns [64]	Predicting irregular crime patterns	Cluster-based prediction & KNN	distance metric & learning vector quantisation (E-KDD)	real per capita & crime data	Visualisation (qualitative)
5	Incremental mining for crime pattern [100]	Spatial &Temporal crime mining	(TAR) Temporal Association Rule	Incremental TAR (ITAR) with Negative border method	Synthesised & real data	Visualisation
6	Automatic tracking of criminals [46]	Tracking criminals & preventing future attacks	Back-propagation NN-classifier & Association algorithm	N/A	Synthesised data	Quantitative
7	Crime hotspot mapping [104]	Spatial crime mining	hotspot mapping approaches	Hotspot Optimization Tool (HOT) with socio-economic factors	real data set	Visualisation (qualitative) & quantitative
8	Cybercrime detection & prevention model [23]	Reduce data damage from malicious code	Support vector machine (SVM) & Adaboost	AdaSVM (hybrid model)	Facebook dataset	Qualitative & quantitative
9	Crime & criminal detection in India [96]	Crime detection & criminal identification	K-means clustering & KNN	N/A	unstructured crime data	Qualitative & quantitative
10	Criminal profiling [5]	Offender behaviour model for criminal profiling	Bayesian Network model	Modified K2 (K2′) to reduce the search space	real crime dataset	Quantitative

Table 2.4: Comparing features of current crime data mining techniques.

S/N	Features	Major characteristics	Strength	Weakness	Choice Implication
1	Association rules e.g Apriori	A strategy for exploring patterns of interest or association relationships in a dataset.	Has great capability for generating candidate itemsets. Easy parallelism and implementation	May require repeated database scans. Measure of confidence and support factors are critical.	Slow unless the required database is resident in memory
2	Clustering	involves organising or grouping objects into groups whose members are similar in some way.	Clustering presents a quick abstraction of the available data set and can be used for pattern discovery and prediction.	It is quite sensitive to initial conditions during the merge or split process.	Statistically isolates dataset into homogeneous subsets based on specified conditions or similarity
3	Decision Trees (DT)	It is a model with a tree-like classification structure, for prediction and decision support.	DT is easy to understand and can be used for data exploration and to predict the outcome for new samples	Some DT approaches may use some heuristics in the classification process; overfit is common	DT describes data, not decisions and may give the illusion of comprehending the target or dependent variable
4	Artificial Neural Network (ANN)	A biologically inspired system that predicts an output based on the set of inputs	Very suited to classification and estimations problems.	Too many input features may affect the results of the analysis.	Complex models may lead to faulty learning process.
5	Support Vector Machine (SVM)	A model that analyses dataset and constructs separating hyperplanes for classification problems.	Suited for regression & classification tasks. The kernel trick enables it to perform both linear & non-linear classification	Generalisation ability of the model depends on the location of the separating hyperplane	Appropriate selection of training samples is crucial for effective analysis.
6	Visualisation	Effective for graphical or iconic summarisation of a large dataset	Presents a high level abstraction of the available information for data exploration	It may be difficult to assimilate or understand for non-domain experts, especially when structured in a hierarchical mode	Can be combined with other mining techniques and may rely on quality analysis result or "complete" data for a good knowledg erepresentation

2.5 Crime Series Research

We note that while there exist a wealth of research [11, 41, 53, 69, 70] on spatial-temporal and hotspot detection (see Table 2.3), the same cannot be said about academic research efforts in the area of crime series identification, despite its critical importance for public safety enhancement. Research on crime series identification has not gained prominence in the data mining community, when compared to other data mining problems such as hot spots. In what follows, we present some key research in this domain.

Crime series analysis focuses on crimes thought to have been committed by the same individual or offenders, and may not necessarily happen at hotspot locations [78]. Hence, it is a unique kind of problem. Experience has shown that many crimes are due to repeat offenders [57], therefore crime series identification is critical for suspect prioritisation and crime deterrence in the public safety domain. However, our findings reveal that the crime intelligence unit in most of the developing nations (e.g., South Africa) do not currently have an automated means of identifying these similar attributes or incidents. The degree of similarity between the features or evidence in two or more crimes can further motivate a number of research questions: (i) How does the evidence transfer between previous and current crimes? (ii) To what extent can we attribute the crime observations or offender characteristics in the current crime to previous ones, for example in terms of Modus Operandi (MO) reported? Such questions are also at the intersection of efforts in predictive policing (Pp). The goal of conducting such an analysis lends itself well in the area of Pp, which relates to identifying potentially related offences, similar criminal attributes and potential criminal activity, in order to deploy actionable measures in deterring crime.

Predictive policing deals with the usage of analytical and predictive techniques in law enforcement to identify and curb potential criminal activity. Existing software technology such as PredPol is a good sample application in the area of Pp [86]. However, PredPol does not predict a potential offender, but helps in identifying and preventing potential crime incidents by predicting the next probable crime location. The majority of Pp software applications are capable of detecting general background levels of crime density (for example, hotspots), which is less challenging to predict as they primarily require mostly location and time information, and not suspect or victim information. Furthermore, it is important to recognise that general background information of crime may not be directly actionable, while identifying the exact M.O. of criminals is directly actionable. This is due to the

fact that while detailed crime information (such as suspect or victims characteristics) is required for crime series identification, the same is not necessarily required for estimating general background levels of crime. Predictive policing approaches are categorised into four distinct methods, these are:

- methods for predicting perpetrators' identities.
- methods for predicting offenders' characteristics.
- methods for predicting crimes.
- methods for predicting victims of crime.

While police typically carry out random patrols at certain locations in order to deter crime, this has proven to be insufficient. Research revealed that place-based patrols can significantly lead to crime reduction, provided it is well articulated and coordinated [41, 70]. Therefore, identifying areas where related crimes are perpetuated, using "smart statistics", could result in the apprehension of the prime suspects in those areas as well as better understanding of past crimes, among others. More generally, crime series detection can help in achieving crime reduction targets faced by security and public safety agencies, by revealing raw-data rows and exact attributes that characterise the M.O. of the perpetrators.

A fair amount of research efforts has gone into (potential) series identification. However, some of the research falls short in considering metrics that can guide actionable decision by public safety agencies [20, 61, 85, 105]. One of the goals in this study is to focus on statistically interpretable patterns that can be actionable for crime reduction. According to Porter M.D. [85], there exist three major types of approaches to series identification, they are:

1. pairwise linkage, which seeks to establish if a pair of crimes share a common criminal.
2. Reactive linkage, which starts off with a seed crime or set of crimes and tries to identify the other crimes that share a common offender with the seed crime.
3. Proactive linkage, which is crime series clustering approach, using viable mathematical and statistical approaches to link crime incidents.

Kamal et al. [20] were part of the early researchers that have considered investigating crime series analysis. Their research focuses on a classification system for serial criminal pattern detection, and implements a hybrid neural networks system, and a rule based heuristics system for the classification process. A Kohonen network is used to construct a clustering system for identifying and grouping potential serial crimes. While their work seeks to propose an automated design process capable of systematically identifying groups of crime records as potential patterns for serial offenders, certain limitations are observed in their approach, which might require some post-processing of identified patterns. For example, their approach feeds all attributes into one huge neural network and assumes that all attributes are of similar importance in deducing the resulting classifications. This assumption of similar importance for all attributes may not necessarily be adequate. Certain prominent attributes which may help in determining pattern specific MO of offenders should naturally carry more weight during analysis. For example, in order to capture prevalent patterns, prevalent information such as the location, time and day of crime should not be of similar importance with some other attributes during the analysis. Furthermore, their approach fails to classify crimes that occur shortly before midnight and those that occur shortly after midnight together. Again, this has the potential of omitting certain attributes of interest and important patterns. Our approach to address this limitation employs a 2D geometric projection of the time information such that the computed distance between crime incident occurrences is considered during analysis. This means that crimes objects that are close enough in time and space could be in the same cluster. CriClust uses a dual threshold scheme, that is the notion of prevalence and significance thresholds, to preserve certain measures of interest in identifying the final CSP. This allows certain prevalent attributes (such as location, time and day information) to be appropriately weighted and subsequently influence the analysis in the final CSP detection.

A related effort is the work of Lin et al.[61] who focus on linking criminal activities using an outlier-based association detection, whereas Zoete et al. [119] also attempted to link crime cases using a Bayesian network model approach. The limitation of a Bayesian approach is that identifying a serial offender may become really cumbersome in a very complex network of linked crimes [43], as a result of building connections (directed acyclic graph) of identified related attributes as depicted in Figure 2.6. Furthermore, Vlek et al. [102] noted that constructing a Bayesian Network (BN) for complex scenarios is sometimes cumbersome and thus proposed a tool for exploring crime scenarios in a series of distinct stages by building a BN in a step-by-step fashion. This step-wise notion exploits

unfolding crime scenarios with variations. The idea only focuses on crime scenarios of interest at a given time in an investigation and expands each scenario as evidence unfolds. Such an approach therefore results in a relatively more rigorous network of crime incident observations. The downside, however, is that inclusion or exclusion of exculpatory evidence could result in serious overestimation or underestimation of the strength of the remaining cases or scenarios. This can also be categorised as a reactive approach, which has the potential of producing incomplete evidential scenarios, depending on the choice of the initial and subsequently unfolding scenarios.

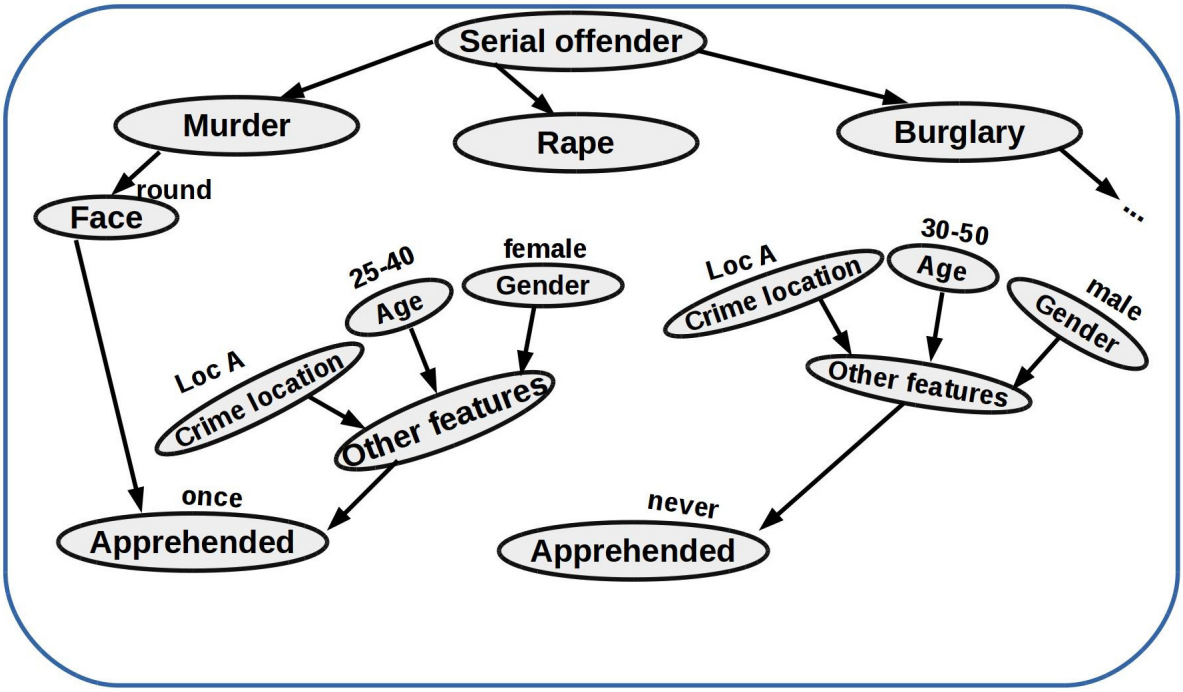


Figure 2.6: A depiction of a complex Bayesian network of suspect modelling.

This approach to crime linkage model as described in [119] simplifies crime scenarios and does not capture all the factors that play a role when linking crimes. While their research approach is useful for discovering interesting aspects of the reasoning underlying crime linkage, they however noted that their approach may not be recommended for actual casework.

Ellingwood et al. [28], on the other hand, investigated the role of similarity coefficients and behavioural themes in crime linkage. We note that behavioural profile, in conjunction with geographic profile, can provide a sound basis for crime series modelling and analysis [7], and consequently help public safety agencies expedite follow-up investigative goals. Hence, our research tries to leverage these behavioural themes and similarity concepts.

Quite recently, Wang et al [107] also contributed to the area of crime series detection through the identification of "rotten core sets" in crime data. In particular, they focused directly on discovering series in housebreakings (burglary). Their earlier attempt [105] was based on a reactive linkage method where an iterative approach is adopted for series identification. In this approach, crime series are identified sequentially, one crime at a time, starting from a seed of one or more crimes. This has some limitations in terms of scalability and efficiency, and has an important implication of producing results that basically depend on the choice of the "seed crime" identified at the onset. Thus, a wrong seed will be a bad choice for further mining. An advancement to this approach launched a consideration for "core sets", where they considered pattern-general and pattern-specific attributes to identify series-like patterns. A similarity graph is derived based on a single threshold, after which core sets are mined using some integer linear programming (optimisation) approach. The size of the core sets, which serve as seeds for further mining, is pre-defined. These sets are then used for deriving bigger (merged) patterns. The derived core sets are typically merged or expanded to identify further series-like patterns or clusters of interest. In their evaluation, they focus attention on a core set that is good, that is core set that do not overlap or cover more than one real pattern. They conduct experiment with 51 hand-labelled series-like patterns supplied by a crime intelligence unit, specifying the size of the core set during analysis. They learn the pattern general cohesion and cut-off threshold in the aforementioned 51 historical crime-series identified by crime analysts. In our approach, we, however, learn the similarity pattern from scratch without any hand-labelled series-like pattern (core sets). This means that we do not have access to any historical (pre-identified) series. Furthermore, we used an adaptive graph size contraction heuristic in Karger-Stein algorithm to derive useful clusters.

In this research, we present a Crime Clustering (CriClust) model based on a proactive linkage mechanism, which considers behavioural themes of offenders in pattern prevalence preservation and crime series identification. A dual-threshold scheme was used to derive similarity graph and highly connected sub-graph (HCS) concepts were adopted to distinguish crime series. Moreover, we introduce two new interest measures: (i) Proportion Difference Evaluation (PDE), which returns the propagation effect of a series and the dominant series at a location; and (ii) Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series. The Google map application programming interface (GMAPI) was used to enhance cluster visualisation. In what follows we provide a brief overview of the HCS as employed in the current research.

2.6 Highly Connected Sub-graphs Concept

The highly connected sub-graph (HCS) approach is a technique that uses graph connectivity for clustering purposes [37]. In the HCS technique, there is no need for prior knowledge about the number of clusters to be derived as opposed to techniques like K-Means [47, 82], which requires the specification of the number of clusters in advance. HCS is a graph where the longest distance between any two vertices (its "diameter") is at most two [37, 40]

There are two major phases involved in the technique: first, the data is represented in a similarity graph based on some similarity condition(s) and, second, sufficiently connected nodes (HCS) in the similarity graph are identified as clusters. The latter is achieved by repeatedly using a "minimum cut" to partition the similarity graph [13]. A cut partitions a graph G into two subgraphs by removing N number of edges. If no cut of G can have fewer than N edges, then this is a minimum cut (or "min-cut"). By using min-cuts in HCS, nodes having many interconnections among themselves are not partitioned. That is, remain together and emerge as a cluster. More formally, let K(G) be the minimum number of edges whose removal disconnect a graph G . A connected graph G with n vertices is highly connected if the following condition is satisfied:

$$K(G) > \frac{n}{2}$$

In order to achieve clustering using a graphical approach, it is necessary to find some way to quantify similarity between elements. However, it is important to recognise that what constitutes a similarity condition is application dependent and usually guided by the goal of the analysis. In this study, our notion of similarity condition hinges on a dual threshold mechanism. The dual threshold scheme helps to quantify the relative measure of information similarity content in a set of crime. Further details about the modelling approach and techniques considered in this research are described in the succeeding chapter.

2.7 Research Goal Revisited

This study is motivated by the persistent need to combat crime, particularly in resource constrained settings where police are short staffed. This necessitates adopting smart tactical techniques for dealing with crime [68]. Figure

2.7 presents a summary of open research topics and issues in data science and crime mining.

Identifying crime series, which are crimes considered to be committed by the same set of offenders, is recognised as one of the smart ways of dealing with crime. However, the literature shows a paucity of research in this area when compared to other areas of crime mining. Identifying crime series patterns is of practical importance because research has shown that many crimes that are committed at specific locations are as a result of repeat offenders [112]. This can be addressed through the use of machine learning and data mining techniques with appropriate heuristics, since they can reveal potentially useful information about crime activities. The idea is not to replace crime experts or analysts by crime analysis software, but to improve their productivity in responding to the safety needs of the community and achieve a more sustainable "safe and smart" city. Hence, this research serves to equip the public safety domain, in resource constraint environments, by providing a viable means by which useful information could be effectively derived from crime data.

2.8 Summary of Research Gaps and Opportunities

The background and related studies have identified some research gaps and opportunities that have motivated this study. These are summarised as follows:

i. There is need to develop models to support and promote crime intelligence in:

   a. Resource constrained environments such as in developing nations, where police are short-staffed and have limited resources in fighting crime.

   b. Crime series analysis to identify crimes thought to have been committed by the same offender(s) and present statistically interpretable patterns for actionable solution.

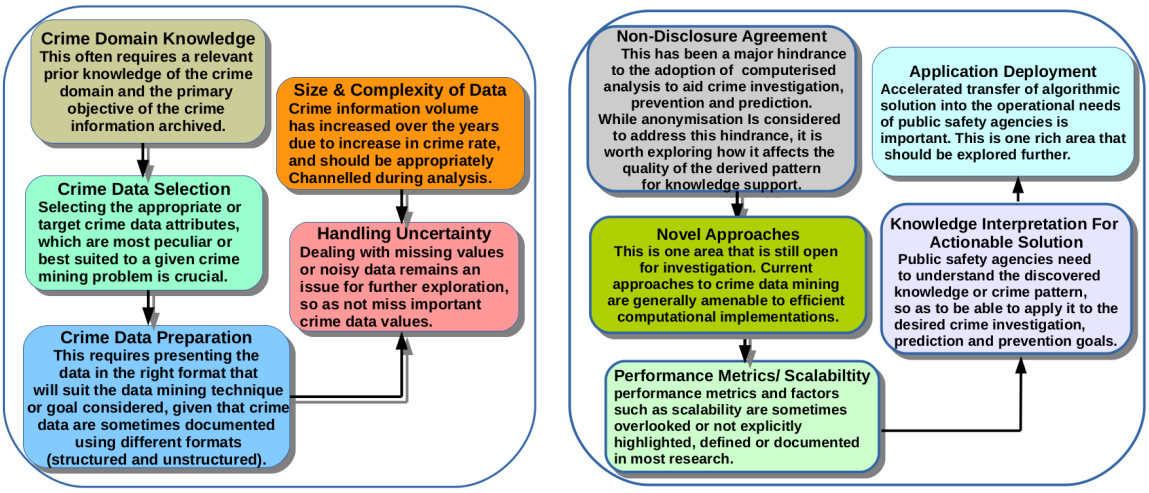
   c. Placing emphasis on how crime mining related research can promote citizen-centred safety in a smart city development of the future.

ii. Crime series mining is less explored in developing nations and has great potential in fighting against crime and promoting safety in a smart city development.

a. Most crime mining research have focused on hotspot detection related tasks and failed to address the specific need and scenarios of series identification.

b. Most successfully deployed crime mining strategies and applications have been in the developed nations, such as the USA.

c. Existing crime series related research are generally amenable to more advanced techniques and heuristics, such as mutual information, dual-threshold scheme and similarity concepts, for improved analysis.



(a) Identified topics in crime mining (b) Some key issues in crime mining and data science

Figure 2.7: Research Topics and Issues in Crime Mining in Developing Nations

2.9 Chapter Summary

The prominence given in today's world to the field of smart public safety and crime data analysis emphasises its importance in a smart city development. This chapter offers a review of the background study as well as related literature in the field of crime data mining, evaluates existing knowledge and specifically presented the gap that this study is intended to fill, namely crime series mining to obtain knowledge that is actionable in developing countries. Furthermore, an overview and critical review of existing data mining techniques are presented using a high-level tabular representation. Features of existing applications and techniques, such as model selection, exploratory basis and algorithm advancement were compared in this chapter. Previous research efforts that relate to crime series

identification were also explored, and a general background to highly connected sub-graph approach documented. The chapter concludes with a summary of gaps and opportunities that have been identified in previous research as well as key open research topics and issues in the domain of interest.

In a smart city development it is recognised that the use of modern technology is necessary to fight different kinds of crime, this may not be sufficient. Actionable knowledge, through smart statistics, can be derived from available datasets. There is paucity of research in crime series detection. For example, while there exists a wealth of research in hotspot related analysis, the same cannot be said about crime series detection. In addition, crime series analysis can provide more actionable information for deterring crime in resource constraint settings. By resource constrained settings we mean scenarios (such as in Africa) where: (i) police are short-staffed; (ii) crime intelligence experts are limited; and (iii) not enough technological solutions are put in place to meet up with the daily operational safety needs and citizens' security. For example, where the quantity of assets or population of citizens' to be protected significantly outnumbers the available public safety or security personnel, it becomes very difficult to achieve and sustain a safe society or community. However, an effective use of data mining approaches can ameliorate these problems or challenges. The next chapter presents the detail of the CriClust model proposed in this work.

Chapter 3

MODEL FORMULATION AND DESIGN METHODOLOGY

This chapter re-establishes the research rationale and presents the model, properties and algorithm specification for CriClust. CriClust focuses on Crime Series Pattern (CSP) detection using a dual threshold scheme, whereby pattern-prevalent information is encoded using a similarity graph. A highly connected sub-graph approach to crime analysis, adopting the Karger-Stein implementation, is also documented in order to demonstrate its procedure and applicability to current research. The two new interest measures in CSP detection, proportion difference evaluation (PDE) and pattern space enumeration (PSE), are also presented. We recognise that what constitutes a (rape) crime sometimes varies by tradition and legal jurisdiction, and that significant domain knowledge is generally required to achieve a viable solution in data mining applications, as such we have worked closely with crime intelligence personnel in South Africa for this purpose. We had a number of formal meetings with the police and crime experts to clearly identify what analysis they consider useful, as well as essential attributes focused in this research. An expository procedure used for data generation and acquisition is discussed. This research is driven by two main objectives:

- to promote a "sustainable safe and smart" city of the future in a resource constrained setting, typically in

a developing nation where police are short-staffed and crime rate is high;

- to leverage data mining techniques and behavioural characteristics of offenders as a means of extracting statistically meaningful information from crime data, in order to provide actionable knowledge to public safety agencies.

It is worth noting that the goal is not to substitute security personnel or analysts by crime analysis software, but to improve their service delivery and pro-activeness by providing a viable means by which actionable information could be derived from crime data.

3.1 Research Design: System Model for Crime Analysis

A detailed framework of situation management is presented by Jakobson *et al.* in [32]. Moreover, we have summarised the major components of situation management in a crime domain as presented in Figure 3.1.

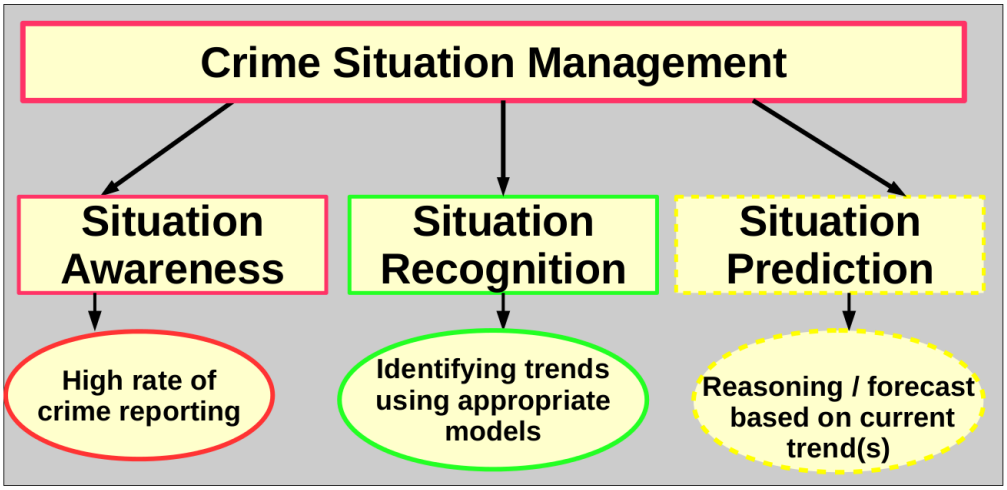


Figure 3.1: Components of a situation management system

Figure 3.2 presents an overview of a target system for public safety enhancement in a smart city development, by revealing different layers as interrelated in the real world. In particular, it reveals how decision makers can benefit from the research and how crime situation recognition, focused in the research, aligns with the smart city concept, goals and objectives. The bottom layer is considered the "sensing" layer which comprises of citizens and victims

of crime. Crime victims report crime to appropriate channels (e.g security agencies), online or in-person. At the middle layer, these crime reports or data are archived by public safety and security agencies for analysis purposes and crime pattern identification, which is crime situation recognition. Such identified patterns can reveal crucial information about the situation of crime, which can be used for suspect prioritisation and crime control. At the topmost layer, derived knowledge of crime trends or patterns are reviewed by security agencies (policy makers, city planners and designers) and used for different safety and situation management decisions, mitigation priorities, as well as timely response or proactive call to e-government and e-safety, among others.

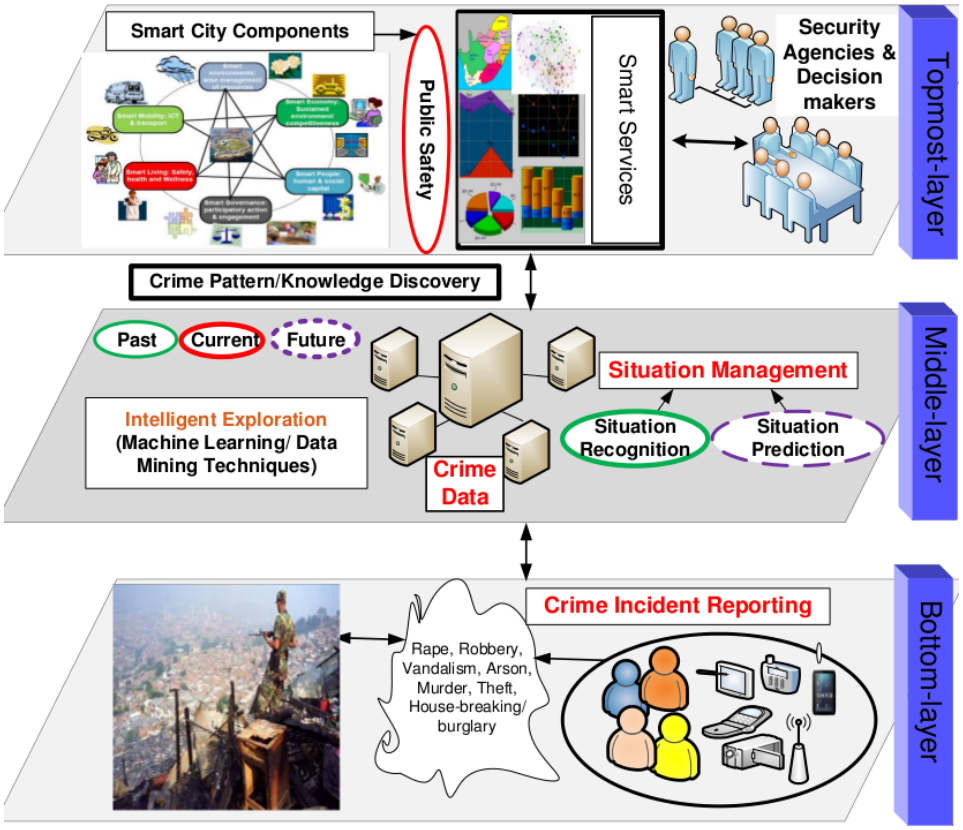


Figure 3.2: A depiction of citizen-centred safety promotion in a smart city development.

This research only focuses on the middle layer of the overall framework, that is identification of crime pattern (potential crime series) using machine learning and data mining techniques, in order to assist security and public safety agencies in achieving their crime reduction targets.

3.1.1 Research Focus and Approach

While Figure 3.2 describes a broader research overview in crime mining and public safety enhancement, the focus of this research is how intelligent exploration using data mining and machine learning techniques can be applied to crime data for useful knowledge discovery. We note that while much research has been conducted in the areas of spatial analysis and hotspot detection [41, 45], there is less exploration in the area of crime series detection, particularly in developing nations with high public safety resource constraints. Hence the motivation and focus in this research.

Essentially, data mining is achieved through two basic components, (i) data query (raw data extraction), and (ii) knowledge query (deducing insight), as presented in Figure 3.3. While data query involves the extraction of data field(s) of interest (attributes), knowledge query on the other hand involves the identification of target tasks (e.g. clustering, classification), and the application of the appropriate model for pattern identification and knowledge support. Figure 3.3 also agrees with existing data mining standards, such as SEMMA (Sample, Explore, Modify, Model, Assess) and CRISP-DM (Cross-Industry Standard Process for Data Mining) [80]. The domain of interest in this research is the (rape) crime domain and the target task involves the identification of crime clusters (CriClust) for knowledge discovery to support and promote crime reduction targets.

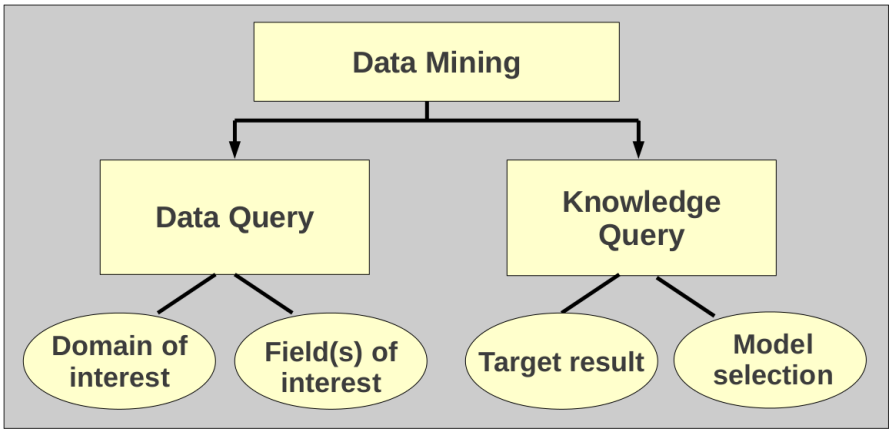


Figure 3.3: Essential components of data Mining for knowledge delivery.

Sample Data for Exploration: Rape Database

While this research explores a rape database for identifying crime series, the idea considered in this study can be extended to other forms of crime. The motivation for considering a rape database is the fact that despite the increased sensitivity and understanding about sexual assault and violence, South African communities turn out to be a place where rape is of great concern [77] [1]. Typically, repeat sexual offenders use key characteristics that minimise their possibility of being a suspect, and carefully plan every attack, such as zeroing in on a victim, scouting a location, and planning escape routes. The compulsive recurrent nature of rape attack could be investigated using past data records. Evidence from such analysis would: (i) guide crime deterrence strategies; and (ii) create some awareness and guide policies that could help victims become more empowered by keeping out of sight of such serial rapists, being aware of the fact that offenders tend to establish a "comfort zone" and keep to strategies that assure them of both access to potential victims and an easy escape route [2]. The comfort zones could vary from "home-invasion" rapist to "outdoor" rapist.

It is crucial to note that understanding the crime data information and focus of the crime analysts is key to resolving a crime clustering problem. The parameter and attributes specification in this research were defined and prescribed by the Western Cape crime intelligence unit in South Africa, and attributes were double-checked by the police to make sure they align and fit with police description of data occurrences.

Police usually keep data on "who", "what", "where" and "when" of crime, among others. The "who" factor basically involves victim and suspect characteristics such as, age-group, race (ethnic group), gender, frame, height, to mention a few. In some instances, they might know or presume who the suspect is from fingerprint or semen data, in which case this might be recorded in the form of their unique ID number. An issue of important consideration in this context is the categorisation issue. The ideal categorisation needs to be carefully considered for each attribute (such as the modus operandi, motivation of suspect). For example, for ease of analysis we categorise the time of day into four major categories, which are: (i) morning; (ii) noon (iii) evening; and (iv) night. This is done in accordance with the current police categorisation, as depicted in Table 3.1. However, many reported crimes record morning, noon, evening or night depending on the approximate period the victim can

[1] http://rapecrisis.org.za/

[2] http://www.uct.ac.za/dailynews/?id=9679

Table 3.1: Time of day categorisation and mapping.

S/N	Time of day	Details
1	Morning	00:00 to 08:59
2	Noon	09:00 to 14:59
6	Evening	15:00 to 20:59
8	Night	21:00 to 23:59

recall. For example, if the crime lasted longer than a specified period or straddled between any two periods, so the four "time of day" values were used. The categories for each attribute need to be consistent and well-defined, typically this may vary by the nature of a crime committed. Note that the terms attribute and feature are used interchangeably in this research.

As access to real crime data with the police has been difficult, and particularly in developing nations, this research considered a quasi-real dataset for experimental purposes. However, the data information and attributes considered are: (i) guided by inputs and recommendations from the police; and (ii) confirmed as representative of what the police keep in realistic scenarios of crime.

In generating the (quasi-real) crime data, we first assume that the necessary pre-processing of raw data (structured and unstructured) has been completed. That is all inconsistencies have been checked (see preprocessing in section 2.2.1). We then use a pseudo-random algorithm based on a Gaussian distribution to populate the database, based on ground-truth provided by the police. This is based on the "java.util.Random" package. First, a database to store the crime data information is established; for example, say RapeDataDB (final String DATABASE_URL = "jdbc:derby://localhost:1527/RapeDataDB;create=true;";), then an array of possible realistic values $(val1, val2, val3, \cdots)$, as prescribed by crime experts, is defined for each attribute or feature. As an example for the "suspect" and "victim" categories, they use the following:

$$\{\text{“I\_Male”}, \text{“I\_Female”}, \text{“W\_Male”}, \text{“W\_Female”}, \text{“B\_Male”}, \text{“B\_Female”}, \text{“C\_Male”}, \text{“C\_Female”}\};$$

The prefix on gender information (e.g., I-male, B-female) represents the different racial population categories in SA. Similarly for incident day, we have:

String incidentDay[ ] = {"Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"};

[3]. Thus, we loop through the database (say, RapeDataDB) based on the pseudo random number generated and specified record size, say n = 10000, and then automatically insert corresponding values for each attribute using the structured query language (SQL) as follows:

for (int i = 0; i < 10000; i++) {

String sql = "insert into RapeDataDB values(' " + incidentDay[rand.nextInt(7)] + " ', ' "+

suspect[rand.nextInt(8)] + " ', ' "+ ⋯ + " ')";

statement.executeUpdate(sql); }

Note that the arguments rand.nextInt(7) and rand.nextInt(8) takes on the values 7 and 8 respectively, which correspond to the attribute domain, i.e possible number of days (incidentDay), and defined number of ethnic group-gender-values for suspect information in the array (suspect[ ]). Thus, the value to be inserted at each iteration would depend on the value returned by the pseudo method, and the corresponding value in the array defined for each attribute. Randomisation approach is crucial in the data generation process in order to ensure that particular feature characteristics among all the possible crimes are well distributed.

Table 3.2 presents a sample data generated using the pseudo-algorithm, while Table 3.3 presents a description of some features and subjects considered in this research. These features are considered in order to assist the analysis in determining the prevailing modus operandi (MO) of a series, with respect to other attributes. Other features relate to the offending process, such as method for getting a victim, and whether the suspect was perceived to be on drugs, or disguised in any form (e.g., masked), among others.

[3]http://www.statssa.gov.za/publications/P0302/P03022010.pdf

Table 3.2: RapeDataDB: Depiction of crime data information generation.

Day	Time	Location	Victim	Suspect	⋯
Tuesday	Morning	Woodstock	I_Female	B_Male	⋯
Friday	Night	Mowbray	B_Female	W_Male	⋯
⋮	⋮	⋮	⋮	⋮	⋮

Table 3.3: Description of the different categories of important features considered.

Features	Categories
Victim and Suspect information	Indian-male (I-male)
	Indian-female (I-female)
	Black-male (B-male)
	Black-female (B-female)
	White-male (W-male)
	White-female (W-female)
	Coloured-male (C-male)
	Coloured-female (C-female)
Method of victim capture	Lured
	Kidnapped
	Weapon
	Deceit
Incident location/day/time	Time and location information (as detailed in section 3.1.1)
Substance_abuse_suspected	traces of substance (drug) abuse: (yes, no, unsure)
Suspect_disguised (Masked)	(yes, no, unsure)

3.2 CriClust in Crime Series Pattern Detection

To achieve our objective of deriving potential crime series pattern in a crime dataset, we propose a hybrid model, called CriClust (CrimeCluster), which combines techniques from information theory, geometric projection and graph clustering. Figure 3.4 outlines the main phases involved in our research. It is summarised in four steps or phases. In the first phase, we identify the crime items (attributes) of interest to be analysed using the expert knowledge defined metrics or characteristics. The second phase involves learning a similarity graph using a dual threshold scheme, which is based on a similarity function. Having encoded pattern prevalence information using a graph similarity structure, clusters are identified in the third phase using a highly connected sub-graph approach

that borrows from the Karger-Stein method, and finally regarding reasoning, in the fourth phase, the derived cluster is statistically interpreted in a manner that will assist actionable solution. This involves using two interest measures which are: (i) proportion difference evaluation (PDE) and; (ii) pattern space enumeration (PSE). Further details of the methodology as well as computational approaches to the Crime Series Pattern (CSP) discovery task are presented hereafter.

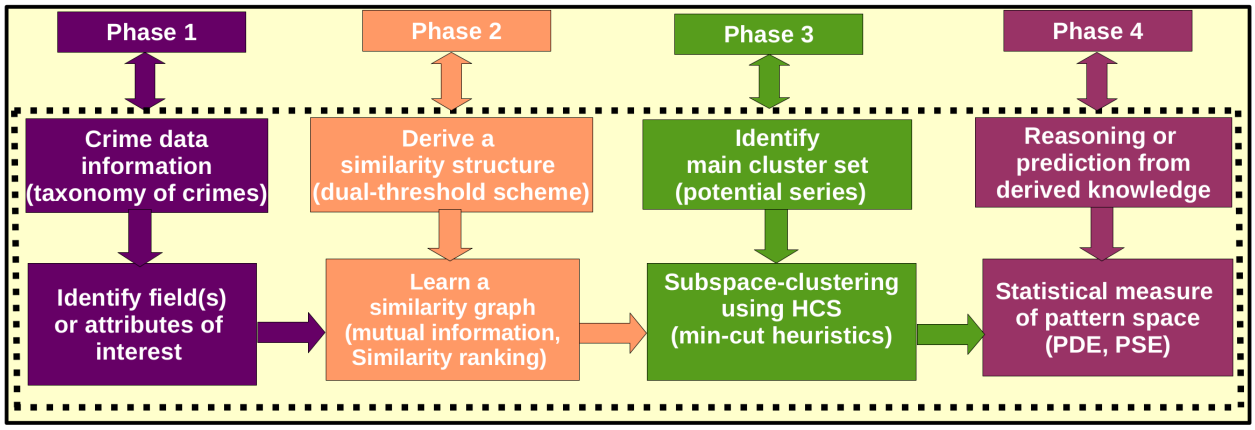


Figure 3.4: A depiction of the research phases in CriClust system.

3.2.1 Problem Definition and Model Formulation

Let C be a set of crime items or objects, where each crime object, say $C_i \in C$, is defined by a set of attributes $\mathcal{A}(C_i)$, with cardinality F. Our interest lies in crime objects that exhibit a coherent pattern on a subset of attributes of $\mathcal{A}$. The problem addressed in this research is peculiar and different from a "universal clustering problem" because it necessitates generating a relevant concept description for crime similarity. This requires understanding the different characteristics of a data set and prioritising features that will promote the goal of the analysis. The measure used in this work identifies similarity attribute between crimes C_i and C_j based on two important thresholds $\mathcal{S}$ and $\mathcal{P}$, for sufficiently high (strict) coherence; where $\mathcal{S}$ is the interest similarity support measure (significance threshold), and $\mathcal{P}$ is the prevalence support threshold. Therefore, the following definitions follow:

Definition 1. (Instance Feature (IF)) Consider a crime $C_i \in C$, and a feature f. Let $P_f(C_i)$ be the value of the $f^{(th)}$ feature in C_i. For example, if the crime C_2 occurs on a Monday, then $P_{\text{day}}(C_2) =$ Monday.

We define a binary feature similarity function S_f using the Kronecker delta function, where $S_f(c_i, c_j)$ takes on values in $\{0, 1\}$, depending on the outcome of the similarity measure. That is $S_f : C \times C \rightarrow \{0, 1\} \subset \mathbb{N}$ and is based on correlation with other objects or features:

$$S_f(C_i, C_j) = \begin{cases} 1 & \text{if } P_f(C_i) = P_f(C_j) \\ 0 & \text{otherwise} \end{cases}$$

Definition 2. (Coherence) The coherence of a set of crime C_i, C_j is defined as the sum of their pairwise similarities

$$Coherence(C_i, C_j) = \sum_{f=1}^{F} S_f(C_i, C_j). \tag{3.1}$$

Definition 3. (Significance Threshold ($\mathcal{S}$)) The significance threshold $\mathcal{S}$ for a set of crimes, C, in the feature space is defined as the coherence threshold for two crime objects C_i and C_j to be considered similar. That is if the two crimes exhibit sufficient related attributes in common, then we define crime similarity ($\mathfrak{S}$) as follows:

$$\mathfrak{S}(C_i, C_j) = \begin{cases} 1 & \text{if } Coherence(C_i, C_j) \geq \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

The crime similarity for any non-null crime object reference(s) has the following properties:

1. $\mathfrak{S}(C_i, C_i) =$ true (i.e. 1); [reflexive].
2. $\mathfrak{S}(C_i, C_j) =$ true $\iff \mathfrak{S}(C_j, C_i) =$ true ; [symmetry].
3. $\mathfrak{S}(C_i, C_j) \geq 0$; [non-negativity].
4. $\mathfrak{S}(C_i, C_j) = 0, \iff C_i$ and C_j are independent [well-defined].
5. $(\mathfrak{S}(C_i, C_j) = 0)$ || $(\mathfrak{S}(C_i, C_j) = 1)$; [consistency].

Thus, the similarity function implements a "near-equivalence" relation on non-null crime object references $\{C_i, C_j, \ldots\}$. The *consistency* property stipulates that multiple invocations of the similarity function consistently returns true or consistently returns false, provided the similarity information condition used for comparison is not modified. We note that law of transitivity may not necessarily hold in all cases, but may hold in some instances. The transitivity law states that for non-null crime object references, say C_i, C_j and C_k, if $\mathfrak{S}(C_i, C_j)$ returns true and $\mathfrak{S}(C_j, C_k)$ returns true, then $\mathfrak{S}(C_i, C_k)$ should return true. As an example, consider Table 3.4 with three crime objects reference $\{C_i, C_j, C_k\}$, each having eight attributes depicted as a, b, c, d, e, f, g, h. Suppose we set a threshold of *5* for attributes similarity condition, we observe that $\mathfrak{S}(C_i, C_j)$ and $\mathfrak{S}(C_j, C_k)$ return true, since they have at least five attributes $\{a, d, e, f, g, h\}$ and $\{b', c, e, f, g, h\}$ in common respectively. However, $\mathfrak{S}(C_i, C_k)$ would return false because there are fewer than five attributes common to them (note that attribute $a \neq a'$). Thus, the defined similarity relation can be considered as a *tolerance relation*. In mathematics, a "tolerance relation" is defined as a relation that is reflexive and symmetric but need not be transitive. However, an "equivalence relation" is always reflexive, symmetric and transitive.

Table 3.4: A simplified analogy to show that transitivity may not hold.

C_i	a	b	c'	d	e	f	g	h
C_j	a	b'	c	d	e	f	g	h
C_k	a'	b'	c	d'	e	f	g	h

In learning the similarity graph, there is need for a thorough understanding of the domain requirements and task at hand as the resulting similarity function can have a great impact on the resulting clusters. Thus, it is important to first identify or decide what features of the data are most relevant or more promising in realising the target class we want as clusters, so as to optimise the similarity condition around those promising attributes. For example in our specific case of crime series detection, some interesting features to consider are space and time similarity, disguise information, method for victim capture as earlier shown in Table 3.3. This can help to characterise offender's MO.

The significance threshold helps to eliminate the first level of uncertainty between two crime objects, that is

knowing whether the crime objects, say C_i, C_j, are similar enough, to be considered for further analysis. Having decided on the first condition using $\mathcal{S}$, then we move on to the next constraint, which is the prevalence characteristics, threshold $\mathcal{P}$. Having a number of related attributes in common is not sufficient; for two crime objects to be considered highly related for potential series identification, we need additional constraints to ensure that attributes of interest that will assist the analysis and generate meaningful clusters are considered. We thus extend our metric to further estimate the similarity between two crimes as follows:

The prevalence threshold $\mathcal{P}$ for a set of crimes enables us to weight important features highly. The prevalence attribute is examined based on the computed (Euclidean) distance between two crime objects. In computing the distance measure to capture the information for the prevalence threshold, our approach adopts key principles of basic geometry and extends them to the current research in achieving the 2-D components for the day and time attributes (as shown later). The location (loc) attribute typically has the longitude (long) and latitude (lat) as its (2-D) components (X, Y), while that of day and time is computed using the standard geometry concept. It is important to note that for the location attribute we computed the relative distance measure as suburb (quasi-real) data is used. In the case, where an actual crime location (for example, an exact address) and geoid information are known, then an exact measure of the distance between locations can be computed. We note that spatio-temporal attributes have higher level of implication in revealing the characterising feature for a pattern [18, 70, 104] and as such we have given priority to those attributes by weighing them higher than others, using weight (λ_j) as depicted in Equation (3.2). In general, an ideal weight could be learned from the data or suggested based on expert recommendation. The weights considered were based on expert recommendation in this domain.

$$\lambda_j \sqrt{(x_m - x_k)^2 + (y_m - y_k)^2}. \tag{3.2}$$

Where (x_i, y_i) are the 2-D components of the value of $P_f(C_i)$. In learning the prevalence characteristics for the location attribute, the distance between two crime locations considered can easily be determined using the longitude and latitude information as shown in Table 3.5. For instance, assigning the couples $(-33.95003, 18.49586)$ and $(-33.96579, 18.48102)$ respectively to Mowbray and Rondebosch gives the corresponding distance (0.021647) between the locations. These values give an insight into what might be considered as a suitable threshold for

Table 3.5: Computing distance measure for location attribute.

S/N	Locations Apart	Distance Measure
1	Mowbray $\rightarrow$ Rondebosch	0.021647
2	Milnerton $\rightarrow$ Goodwood	0.044721
3	Mitchell's-plain $\rightarrow$ Milnerton	0.196977

checking how far apart two crimes locations are, when all possible distances are computed and the ceil of the quartile information is considered.

The 2-D representation of day is derived by describing a circle with a centre (equator) and radius (r), say $r = 25$. It is important to note that while we have arbitrarily chosen 25 as the radius of the circle, any other positive value could have been used. This is because the assumed radius (value) can be standardised after all. Figure 3.5 can also be visualised as a "unit circle" (with r $= 1$), where the radius $= 1$ has been scaled up. It is important to recognise that this notion of geometric projection has gained prominence in other domains, such as machine vision [27]. Furthermore, there are established proofs of concept for this notion [67], which is beyond the scope of this thesis.

In Figure 3.5, alpha (α) denotes the angle between each pair of days and is determined as shown in Equation (3.3). The circle was partitioned into seven sectors with equal degrees between the sectors. That is by dividing the circle (360 °) by the number of days (7) in a week, this estimate gives an approximate value of 51.42 for α. Note that "Monday" is the reference point in Figure 3.5.

$$\alpha = \left[\frac{360^{\circ}(\text{in a circle})}{7(\text{days in a week})} \right] \approx 51.42^{\circ}. \tag{3.3}$$

Having calculated α, it becomes easy to calculate the 2-D (x and y) co-ordinates for each day. Suppose we want to compute the x and y components for Tuesday (x_t, y_t), we can easily compute θ, which is simply derived by subtracting α from 90° (angle in that quadrant) as shown in Equation (3.4) and Figure 3.5.

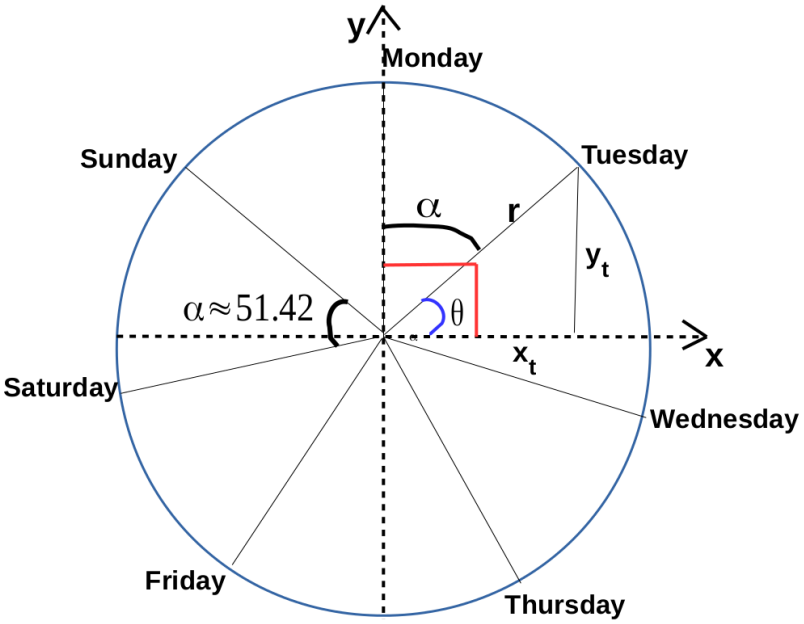


Figure 3.5: Depiction of the 2-D geometric projection for day attributes.

$$\theta = (90 - \alpha)^\circ = (90 - 51.42)^\circ = 38.58^\circ. \tag{3.4}$$

Thus calculating the x, y components for Tuesday, we have the following:

$$\begin{aligned} \sin 38.58 &= \frac{y_t}{25} \\ \implies y_t &= 25 \times \sin 38.58, \\ &= 15.59. \end{aligned} \tag{3.5}$$

Having computed y_t, we compute x_t using Equation (3.5), we have the following:

$$\begin{aligned} \cos 38.58 &= \frac{x_t}{25} \\ \implies x_t &= 25 \times \cos 38.58, \\ &= 19.54. \end{aligned} \tag{3.6}$$

Table 3.6: Angle and corresponding distance (x, y) values for day attribute.

S/N	Day	Angle(θ°)	X(d)	Y(d)
1	Monday	-	0	25
2	Tuesday	38.58	19.54	15.59
3	Wednesday	12.84	24.38	-5.56
4	Thursday	25.74	10.86	-22.52
5	Friday	25.68	-10.84	-22.53
6	Saturday	12.84	-24.38	-5.56
7	Sunday	38.58	-19.54	15.59

The 2-D component is useful because a 1-D component would have assumed that Sunday is far from Monday, which is really not the case. Thus, adoption of the 2-D geometric projection helps to put this into perspective.

Similarly, the same notion of geometric projection is extended to the time attribute as shown in Figure 3.6, in order to obtain a 2-D representation of time information.

Figure 3.6 is simply the graph of the "unit circle" on the $x - y$ coordinate axis, where the unit radius has also been scaled up to 25. It is clear from the figure, that the unit circle is defined as having a radius $r = 1$, instead of 25. Thus, the coordinate points on the axis of the unit circle, counter clock-wisely, will be as follows: $(1, 0), (0, 1), (-1, 0), and (0, -1)$ respectively. These values, when scaled to 25, will correspond to $(25, 0), (0, 25), (-25, 0), (0, -25)$.

Moreover, we use the Z-score scaling (Equation (3.7)) for standardising the data. It is important to note that standardisation is crucial in the analysis, otherwise variables measured at different scales do not contribute equally to the analysis. Thus standardising helps to bring all of the attributes values into proportion with one another.

$$Z = \frac{x - \mu}{\sigma}, \tag{3.7}$$

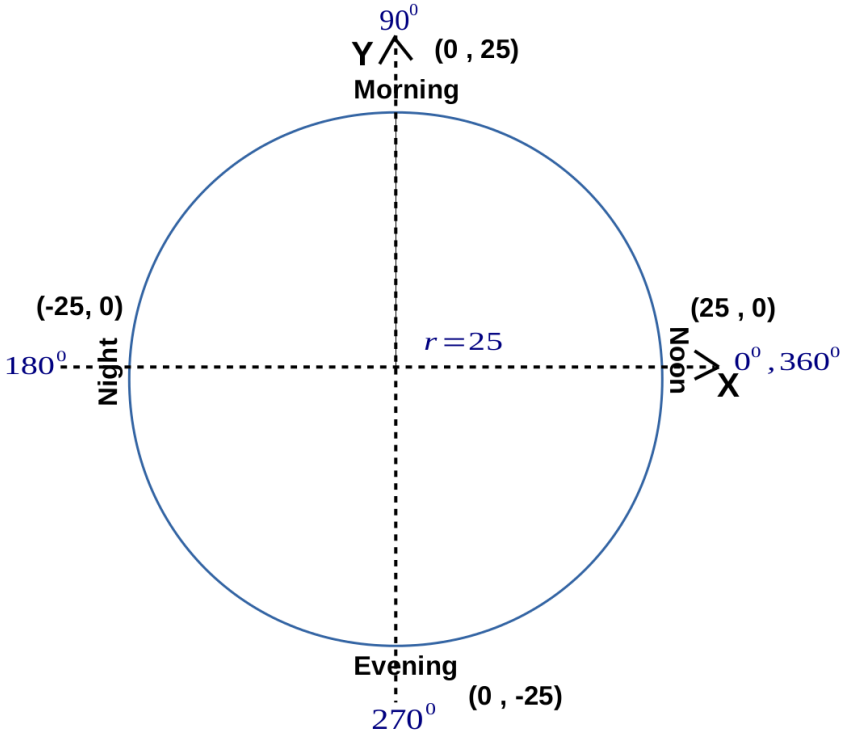


Figure 3.6: Depiction of the 2-D geometric projection for time attribute.

where: μ and σ are the mean and standard deviation of the corresponding attribute values, defined in Equations (3.8) and (3.9) respectively.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3.8}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2}. \tag{3.9}$$

Our threshold is computed based on a sound mathematical principle and crime expert recommendations. The similar objects are then modelled into a graphical structure, to learn a similarity graph that is based on established graph-theoretic model. The significance and prevalence thresholds measure the interest similarity support, and helps to conceptualise the underlying graphical structure, and ensures that a link ensues between two crimes if and only if the support of the similarity attributes is greater than or equal to parameters $\mathcal{S}$ and $\mathcal{P}$. While the parameter $\mathcal{S}$ come from crime intelligence experts as was also done in previous research [61], the coefficient $\mathcal{P}$ is

a parameter we learn from the data. The prevalence threshold considers attributes relating to "day", "time" and "location" information of a crime incident. These features are considered because of their potential characteristics in assisting the analysis; Since a series will happen within a close space-time proximity. In learning a suitable value for parameter $\mathcal{P}$, we consider the data set derived for analysis as shown in Table 3.7 so that distance apart can be derived using the Euclidean formula. For instance, by assigning the couples $(-19.54, 15.59)$ and $(0, 25)$ respectively to Sunday and Monday (from Table 3.6), gives the corresponding distance (21.687778) between the locations as shown in Table 3.8.

Table 3.7: A depiction of the 2-D components for determining prevalence characteristics.

	Geo Loc	Day	Time
C_1	(long, lat)	(x, y)	(x, y)
C_2	(long, lat)	(x, y)	(x, y)
$\vdots$	$\vdots$	$\vdots$	$\vdots$

More formally, $\mathcal{P}$ is set to the 3rd quartile among the set of values computed in the following manner: Consider a crime object $C_i \in C$, we form the $6-$component vector A^i using the 2-D co-ordinates of $P_{day}(C_i), P_{loc}(C_i), P_{time}(C_i)$. Thus,

$$A^i = (P_{day}(C_i), P_{loc}(C_i), P_{time}(C_i)).$$

If $Coherence(C_i, C_j)$ exceeds $\mathcal{S}$ (the significance threshold), we compute the 6D Euclidean distance d_{ij} between A^i and A^j. If the distance is within range, that is not greater than the threshold $\mathcal{P}$, then (C_i, C_j) are connected in the similarity graph. $\mathcal{P}$ is set to the 3rd quartile of the $d_{ij}{}'s$. on the advice of crime experts.

While the work of Porter [85] did not consider temporal relation in computing the Bayes factor in crime linkage analysis, our approach considers this attribute using the prevalence threshold because space-time relation is a key factor in establishing a serial predator or serial crime [20, 28]. We consider the incident day and time-range in our analysis as serial predators tend to establish a time relation and comfort zone in which they perpetrate crime. $\mathcal{S}$ is the interest feature similarity measure, while the prevalence threshold, $\mathcal{P}$, ensures that further attributes of interest such as space (location) and temporal relation are appropriately weighted when establishing that two crimes are

Table 3.8: Sample day attribute characteristics measure.

S/N	Day-Difference	Distance(d)-Apart	Days-Apart
1	Monday-Tuesday	21.687778	1
2	Sunday-Monday	21.687778	1
3	Tuesday-Wednesday	21.69673	1
4	Wednesday-Thursday	21.689444	1
5	Saturday-Sunday	21.69673	1
6	Monday-Wednesday	39.093452	2
7	Monday-Friday	48.750451	4

related. These two thresholds have been invoked based on crime expert recommendations as the location and timing of an ongoing crime series is of vital interest during analysis.

The idea is to try and capture years of human expertise, experience and reasoning into computer models using data mining approaches. It is important to note that we have also worked closely with crime intelligence experts to ensure a valid model and initial concept description, in the context of a developing nation. Thus, the underlying hypotheses or propositions of this research are as follows:

- crime events are seasonal and vary by crime type.
- Most crime patterns exhibit at least a k minimum principal-set that characterise the modus operandi of the offender(s) behaviour.

The database we considered (rape data) has features which comprise location (space), time(t), victim information (i.e. gender, age, etc.), method for victim capture (weapon, kidnapped, lured, etc.), perpetrator information (suspect), and an indicator variable as to whether the rapist was disguised (i.e masked). These features were carefully chosen, with the advice from crime intelligence, due to their capability in characterising the MO of

offenders [28]. Thus, some features in this research are categorical, while some are numeric. It is important to stress that our approach can be applied to other categories of crime (i.e. property related crime, burglary, etc.) and has the capability to extend beyond crime mining domain to other domains that tend to exhibit specific related features that could capture related events. Thus, as nothing specific to rape was used in the model derivation, the model proposed in this research is a general approach.

Thus, we compute a measure that reveals that two crime items or events are similar if they relate to each other in terms of other attributes using a dual threshold scheme (significance ($\mathcal{S}$) and prevalence ($\mathcal{P}$) thresholds). The pattern prevalence information thus induces a similarity graph.

In practice, computing a similarity function or value necessitates finding approximate or exact matches of crime patterns or features in the crime objects being compared. As an example, two crime objects may be considered related if they exhibit a coherent feature values on a number of specified crime features (e.g. happening in the same location, day or time etc.). However, we have used a dual threshold scheme in this research to derive a similarity graph.

Definition 4. (Similarity Graph) A similarity graph is an undirected graph $G = (V, E)$, where V depicts the set of vertices, E depicts the set of edges, $E = \{\{v_i v_j\} : \Lambda(v_i, v_j) \geq \mathcal{S}, (v_i, v_j) \vdash \mathcal{P}, v_i, v_j \in V, v_i \neq v_j\}$.

$$\Lambda(v_i, v_j) = \sum_{f=1}^{F} S_f(v_i, v_j). \tag{3.10}$$

The underlying crime incidents dependency structure can be modelled using a graphical approach as shown in Figure 3.7, based on a metric similarity function, say $\Lambda(v_i, v_j)$ as shown in Equation (3.10). The similarity function, $S_f(v_i, v_j)$, measures the similarity between crimes v_i, v_j in the f^{th} feature.

The representation of the crime similarity data in a similarity graph helps to simplify computation, since we then only need to find sufficiently inter-connected sets of nodes, which is referred to as highly connected sub-graphs (HCS) as shown in Figure 3.8. The discussion and procedure for getting the desired partition (min-cut), which returns the sufficiently connected set of edges, follows shortly.

In CriClust, the graph model adopts HCS, which was originally proposed by Hartuv et al.[37] and has found its

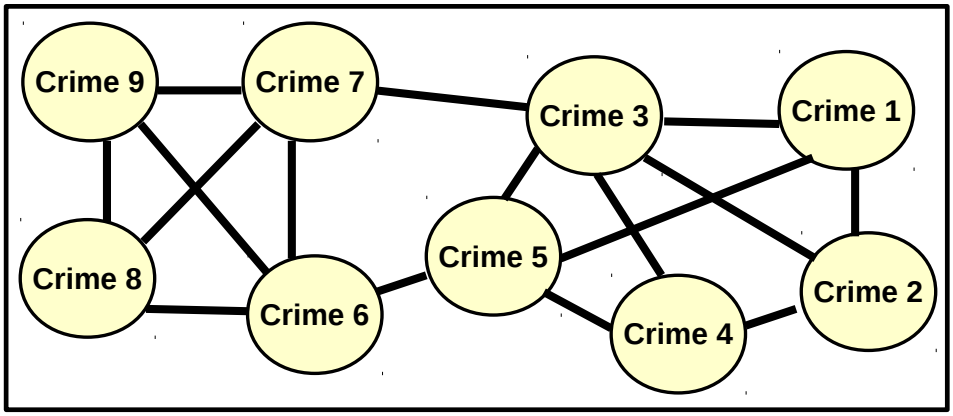


Figure 3.7: Depiction of a crime similarity graph for clustering.

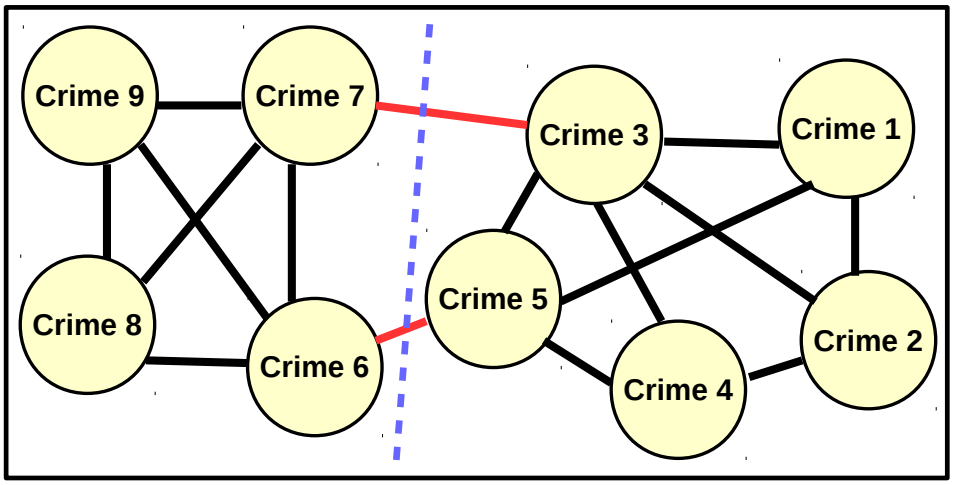


Figure 3.8: Identifying sufficiently connected nodes in a crime similarity graph (red edges are min-cut).

application in many domains that requires identifying such clusters using a graphical approach [40]. Monte Carlo and adaptive graph size heuristics were used to amplify algorithm success. This is to ensure that during series cluster generation, each series' choices expand towards the most promising moves for that particular series. This mirrors the goal of the offenders to maximise their prevalence localities.

For example, the similarity graph in Figure 3.8 (best viewed in colour mode) would result in two HCS when the red links are removed. Thus, each densely connected HCS satisfies the constraints in (3.11) and (3.12). The constraint in (3.11) expresses the fact that in a given HCS, each pair of node satisfies the thresholds $\mathcal{S}, \mathcal{P}$.

Therefore, each cluster inherently has a high intra-class similarity and low inter-class similarity according to these constraints.

$$HCS \vdash \{\mathcal{S}, \mathcal{P}\}. \tag{3.11}$$

$$\frac{|E|}{|V|(|V|-1)} \geq \frac{|V|}{2}. \tag{3.12}$$

Having identified the clusters, each cluster is then considered to have its unique features that will characterise its MO. These features are referred to as the peculiar features for that particular series and may vary for each series.

Definition 5. (Peculiar Feature (PF)) Consider a cluster $\mathcal{C}$ and a feature f. Let $P_f(C_i)$, $C_i \in \mathcal{C}$, be the value of C_i in the $f^{(th)}$ feature. Equation (3.13) returns the values (possibly replicated) of the features f in the cluster $\mathcal{C}$.

$$\lambda_f(\mathcal{C}) := [P_f(C_i), C_i \in \mathcal{C}], f = 1, \cdots, F. \tag{3.13}$$

For a given feature f, the most frequent values of $\lambda_f(\mathcal{C})$ characterise the MO of a particular series. For example, several rape incidents happening within a close space-time proximity and with the same method for victim capture and trailing features would suggest that the same culprit is involved in the crime. In this particular instance, the features "location", "time/day apart between crimes," and "method of victim capture" characterise the MO for the series. Thus, each unique feature values in this instance will reflect the prevalence information for that series. Figure 3.9 depicts the major phases involved in the process of series identification.

3.2.2 Cluster Identification Using Highly Connected Sub-graphs

A graph partitioning problem entails "cutting" a (similarity) graph into two or more "good" pieces (sub-graphs), such that each partition (cluster) is connected and relatively well-defined [37]. In CriClust, clusters correspond to highly connected sub-graphs. Typically, what defines a cluster would depend on the goal of the analysis and the focus of the analyst. Thus, constraints and parameter specifications may differ in various clustering tasks.

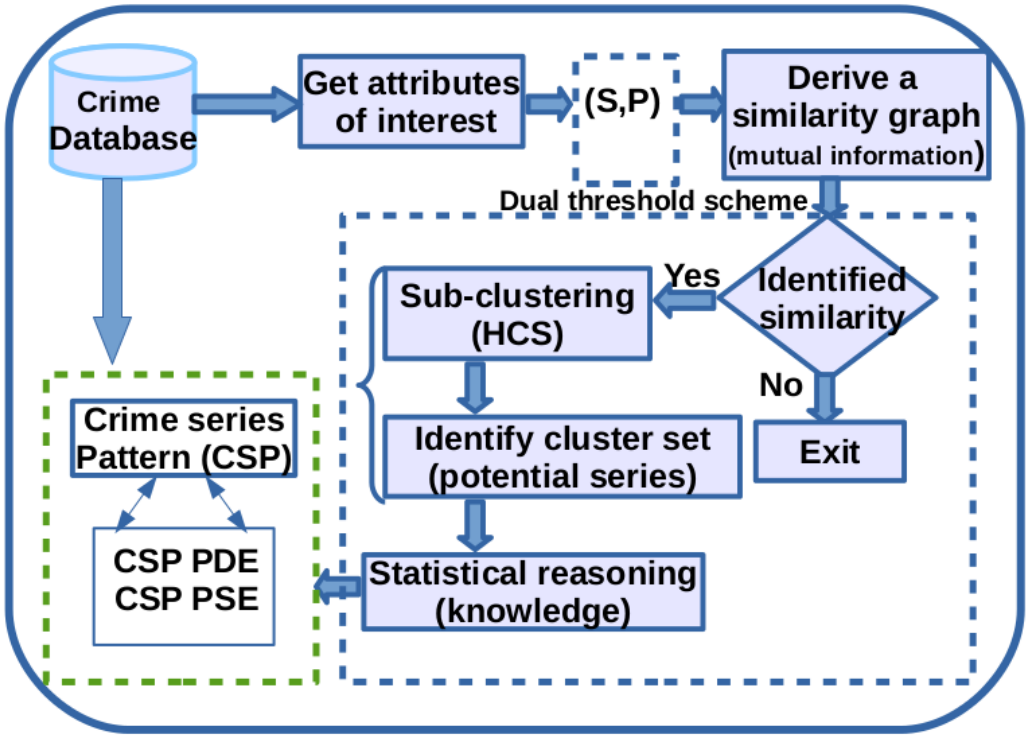


Figure 3.9: Process of crime series detection with CriClust model.

In order to pose the graph partitioning problem more formally, otherwise known as clustering problem [99], the following definitions follow:

Definition 6. (Connected Graph)

1. A graph is a representation of a set of objects where some pairs of objects are connected by edges. The interconnected objects are called nodes (also known as vertices), and the links that connect some pairs of nodes are called edges (also called arcs). More formally, a graph G can be represented as $G = (V, E)$, where V is the set of vertices and E the set of edges that connects the vertices.

2. A graph is connected if there exists a path between every pair of vertices. That is in a connected graph, every vertex is reachable from any other vertex. A disconnected graph is one that is not connected. By convention, a graph with a single vertex is considered connected. An edgeless graph with two or more vertices is obviously disconnected.

Definition 7. (Graph Cut) Given $G = (V, E)$, a graph cut is a partition of V into two disjoint non-empty sets,

$T, \bar{T} = V \setminus T$, where the edges of the cut $(T, V \setminus T) = \{(u, v) : (u, v) \in E, u \in T, v \in V \setminus T\}$. The number of edges in the $(T, V \setminus T)$ cut is called the *size* of the cut.

We are interested in the task of computing a cut in the graph with the minimum size, commonly referred to in the literature as the *minimum cut (min-cut) problem* [12, 37]. That is a cut in the graph whose number (cardinality) of edges is minimum. This helps us to optimise the connected crime objects which form the clusters (C^i).

The minimum cut problem has found application in many research domains, such as in networking (e.g. exploring the theory of network flow problems [81]). It is important to recognise that different algorithms have been devised to solve the min-cut problem. Generally, the minimum cut problem is defined as follows: find the cut of "smallest" edge weight in G, where G could be weighted or unweighted, and directed or undirected.

In general, the min-cut problem is described as the max-flow problem, solved with a deterministic estimate with running time $O(n^3)$ [13]. However, one of the famous approaches to identifying the min-cut is that of Karger [49]. The Karger approach uses a random sampling strategy to identify a potential min-cut. The approach randomly selects an edge from the set of possible edges in a graph G, and uses the contraction process described in Algorithm 1 to reduce G down to "two" vertices with a single cut left between the two vertices. The contraction process simply repeatedly collapses an edge, which means combining two vertices to produce a single vertex.

3.2.3 CriClust Problem Specification

- Specifying the Similarity Graph

  For a crime object C, let $L_d(C)$ denote the 2-D (x, y) component of day attribute, $L_t(C)$ - the 2-D (x, y) component of time attribute, and $L_l(C)$ - the 2-D (longitude, latitude) component of location attribute. Moreover, let $L_i(C)$ be the i^{th} attribute of the crime object C. The specification for CriClust model is as follows:

  $m : \mathbb{N}$ $\rightsquigarrow$ number of attributes ($\geq$ 9)

  $\mathcal{S} : \mathbb{N}$ $\rightsquigarrow$ significance threshold ($\geq$ 7)

  $\mathcal{P} : \mathbb{N}$ $\rightsquigarrow$ prevalence threshold ($\leq$11)

  $N : \mathbb{N}$ $\rightsquigarrow$ number of crime objects

$C : \mathbb{R}^m$ $\rightsquigarrow$ crime objects with m features each

G : array $[N, N)$ of $\mathbb{R}^m \times \mathbb{R}^m$ $\rightsquigarrow$ similarity graph (originally empty)

For $C^q, C^r \in C$

if $|\{i : L_i(C^r) = L_i(C^q)\}| \geq \mathcal{S}$ $\rightsquigarrow$ sufficiently similar crime objects

$A^q := (L_d(C^q), L_t(C^q), L_l(C^q))$

$A^r := (L_d(C^r), L_t(C^r), L_l(C^r))$

if Euclid-dist$(A^q, A^r) \leq \mathcal{P}$

$G+ = (C^q, C^r)$ }

Return G $\rightsquigarrow$ Similarity graph

- Specifying the Minimum Cut (min-cut) using Karger's algorithm

minCut$(G = (V, E))$

$G_0 \leftarrow G$

j = 0

While G_j has $|V| > 2$ do

pick an edge e_j from G_j at random

$G_{j+1} \leftarrow G_j \setminus e_j$

$j \leftarrow j + 1$

$\{T, V \setminus T\}$ is the cut in the original graph that corresponds to the min-cut of G_j

- Specifying the Highly Connected Sub-graph (HCS)

For a graph G, let $E(G)$ be the edge set, $V(G)$ be the vertex set and C be a minimum cut [4]

$HCS(G(V, E))$

begin

$\{H, \bar{H}, C\} \longleftarrow$ minCut(G)

if G is highly connected

then return (G)

else

[4] A minimum cut, C, is the minimum number of edges which separates G into sub-graphs H and $\bar{H}$.

$HCS(H)$

$HCS(\bar{H})$

end if

end

The correctness and completeness of the approach considered in this research can be seen as a direct consequence of the correctness and completeness of the methods and algorithms utilised for achieving the target task [50].

3.2.4 Adaptive Graph-Size-Based Contraction Operations in Crime Series Detection

There are basically two known challenges with Karger's approach for finding min-cut:

1. Since it randomly selects an edge for the contraction process, it can end up with a cut that is not a min-cut and hence needs to be repeated to improve success rate.
2. It may also return a singleton partition, which is not desirable in this research as we are mostly concerned with series-like pattern.

Thus, Karger-Stein (KS) algorithm [50], which is a refinement of the Karger algorithm, was introduced to tackle some of the problems in the Karger method. KS uses a branching amplification technique as an advancement on Karger, which clones the graph and contracts random edges in both copies once fewer than $\frac{N}{\sqrt{2}}$ vertices are remaining in the graph [13]. This helps to reduce the probability of selecting a wrong min-cut .

In order to address the two earlier mentioned problems, we adopt an adaptive graph size (AGS) approach that helps to ensure that an appropriate edge candidate is selected in the contraction process. Combining Monte Carlo heuristics help to promote the selection of an ideal edge candidate from the possible edge set. Adaptivity in this context entails assisting the process adjust its parameters appropriately during the contraction process. Thus, in order to bound the error probability appropriately, the next edge for contraction is selected randomly, a concurrent contraction operation is performed and then the final result is selected based as the minimum of the resulting cuts. The typical contraction process, which basically transforms the graph into two connected components, is

described as shown in Algorithm 1. The remaining edges at the end of the process are those connecting the two components, which partitions the graph.

Algorithm 1 Contraction Process ($G = (V, E)$)

1: repeat until $|V| = 2$

2: select an edge $e \in E$ (uniformly at random)

3: Let $G \leftarrow \mathrm{G}/e$ (collapse-edge)

4: **return** G

In assisting the process with identifying the minimum cut, we employ an adaptive graph-size-based contraction operation that borrows from the Karger-Stein (KS) method [50]. The KS method on the other hand uses a recursive contraction approach to contract random edges until the vertices reduces from n to $\frac{n}{\sqrt{2}}$, where n is the number of nodes. The method recursively calls itself twice on the residual graph and returns the minimum of the resulting cuts (min-cut). The KS method improve the probability of success from $O\left(\frac{1}{n^2}\right)$ to $O\left(\frac{1}{\log n}\right)$, and runs in $O(n^2 \log n)$ [13].

The proof of correctness that the KS method produces a valid cut with high probability has been established [50]. If we are certain that a candidate edge of a minimum cut is not contracted then the resulting output is 100 % correct [13]. In order to mitigate the risk of selecting a wrong candidate as the minimum cut two edge selections are made at random and the minimum cut amongst the two is considered. This is done recursively during the process. We augment this method using Monte Carlo heuristics to amplify the algorithm success. While the heuristic may result in a time-cost trade-off, that is increase in algorithm running time, the chances of deriving the right partition for cluster identification is increased. It is important to note that we are more concerned about getting the crime patterns correctly. This heuristic (minDiff) minimises the node size discrepancy after the merge once the number of nodes is below a threshold. Thus, the desired cluster is derived accordingly as summarised in Algorithms 2 and 3. This heuristic helps to increase the probability of success.

A simple illustration is depicted in Figure 3.10, which shows two supernodes where 9 and 25 nodes have previously been merged, and a third singleton node. The graph on the left has the capability of returning three different possibilities ((i), (ii), (iii)) as highlighted in Figure 3.10, if a random edge is contracted (note that self loops are ignored during the contraction operation). The heuristic selects a combination that promotes our objective, for

Algorithm 2 AGS-mincut(G,N)

1: $G = (V, E)$

2: **if** $|G| \leq 2$ **then**

3: return G

4: K $= \dfrac{|G|}{\sqrt{2}}$

5: Do while $|G| > K$

6: select $e \in E$ (uniformly at random)

7: **if** $|G| < N$ **then**

8: $minDiff(G, e)$

9: **end if**

10: $G \leftarrow \mathrm{G}/e$

11: end while

12: $G \leftarrow \mathrm{A}GS - mincut(G, N)$

13: $G' \leftarrow \mathrm{A}GS - mincut(clone(G), N)$

14: **end if**

15: **return** min $\{G, G'\}$

example in this case, the first possibility (labelled (i)). This is because the merge operation in the first case would return the minimum discrepancy (i.e 25 - 10 = 15 nodes) among the set of possibilities $\{15, 17, 33\}$ respectively. It is important to note that this will not affect resulting clusters since required constraints are checked before a partition (min-cut) is considered at the end of the process.

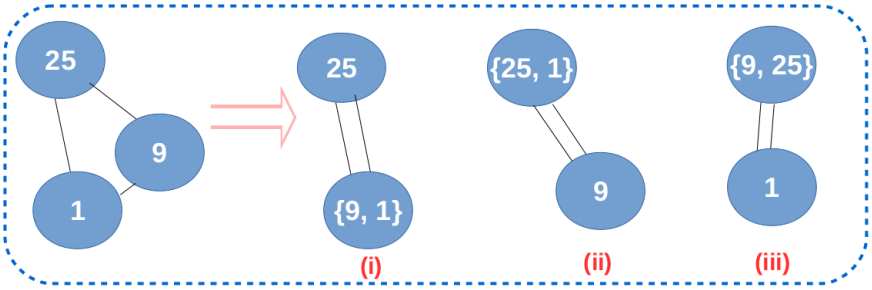


Figure 3.10: A depiction of an adaptive process in promoting highly connected sub-graphs.

Algorithm 3 newMincut (G, T)

1: $minSize = \infty$

2: $N = |\frac{G}{2}|$

3: repeat T times

4: $size =$ AGS-mincut(G, N)

5: **if** $size < minSize$ **then**

6: $minSize := size$

7: **end if**

8: end repeat

9: **return** $minSize$

3.3 Reasoning: Statistical Interest Measures on Identified Potential Series

This section describes the two measures considered useful for actionable solutions in public safety improvement. They are: (i) Proportion Difference Evaluation (PDE), which reveals the propagation effect of a series; and (ii)

Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series.

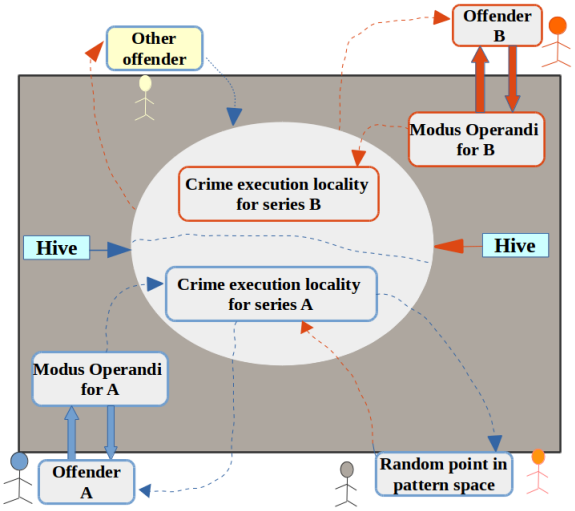


Figure 3.11: A depiction of serial predator colony.

Crime predators (agents) have behaviours (MO) that could be characterised by simple rules, and interactions with other entities, which consequently influence their behaviours (see Figure 3.11, which reveals how crime serial predators interact with their (conducive) localities). Offenders A and B each have a unique way of operating (committing crime) and a "suitable" location in space. We are interested in identifying how these agents (crime predators) repeatedly execute their interactions and behaviours (Modus Operandi). The recurrent nature of a serial crime reveals the propagation effect, (P_ϵ), of the emerging series at the corresponding location as shown in Figure 3.12. The crime predator or serial offender (agent) could take two different forms [20]:

- Single offender, who repeatedly perpetrates a crime.
- Multiple co-offenders, who occasionally act together and sometimes independently.

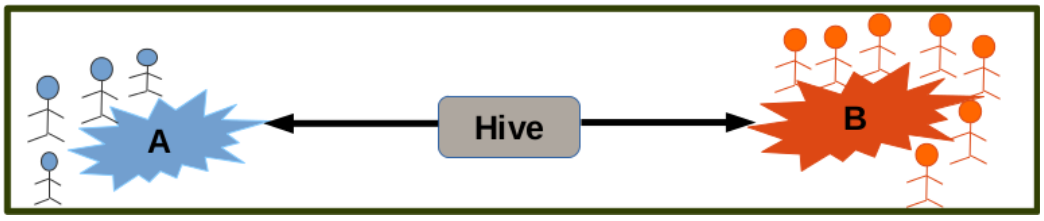


Figure 3.12: A depiction of the propagation effect of two crime series.

3.3.1 CSP Proportion Difference Evaluation

We explore the proportion difference between series at the same location, using relative measures rather than absolute measures. That is the within cluster and between cluster proportion difference. This provides a better insight into the "peculiar attributes" for each location and causative factors for perpetuating the crime. Hence promoting suspect targeting and prioritisation; that serves as a guide for police as to which perpetrators to hunt down first.

In situations where more than one crime series (CS) pattern emerges at a location, the propagation effect (P_ϵ) helps to identify the dominant series (CS_d), through evaluating the PDE of those series. Identifying a dominant series is very useful for investigative and suspect prioritisation. Consider $\mathcal{C}^1, \mathcal{C}^2, \ldots \mathcal{C}^i$ to be crime series clusters, and $|\mathcal{C}^i|$ the number of crime objects of the i^ith cluster (series), the following definitions follow:

Definition 8. (Propagation Effect (P_ϵ)) The propagation effect of a series is a measure of the dominating power or collective strength of the series as a result of the compulsive recurrent nature of the crime committed. P_ϵ is more formally described in Equation (3.14).

$$P_\epsilon(\mathcal{C}^i) = \frac{|\mathcal{C}^i|}{\sum_j |\mathcal{C}_j|}. \tag{3.14}$$

Definition 9. (Dominant Series (CS_d)) A dominant series, in the set of clusters (CSP) identified at a location l, CSP_l, is the series with the highest propagation effect, that is highest proportion of instances identified.

More formally, CS_d at a location (l) is defined in Equation (3.15):

$$P_\epsilon(CS_d) := \max_i P_\epsilon(\mathcal{C}^i). \tag{3.15}$$

In a situation where two series have the same value, then $P_\epsilon(CS_d)$ assumes the value of ∞, until a series emerges as dominant.

Therefore, the PDE entails identifying the propagation effect of a particular series as well as the dominating power of that series at a particular locality.

The proportion difference satisfies the axioms of a probability measure, this means that it is always positive and cannot be greater than one. Where there are only two series identified at a location, the proportion difference for the other series would be defined as given in Equation (3.16).

$$CSP^{(2)} = 1 - \left(\frac{|CSP^{(1)}|}{|CSP^{(1)}| + |CSP^{(2)}|} \right). \tag{3.16}$$

It is instructive to examine how the derivation fares for the case of a k-series pattern, where $k \geq 2$. For example, consider a 2-series pattern, say $CSP_l = \{CSP^{(1)}, CSP^{(2)}\}$, identified at a location (l). If the two series, CSP_1, CSP_2, are perfectly positively correlated in terms of P_ϵ (i.e. have the same propagation effect), then the dominant series for the two corresponding series is set to ∞, which implies that we do not currently know which of the series is dominant. This is due to the fact that there is a tie between the probability of occurrence of both series at the location. So, any of the series could be tackled first by public safety agencies. However, in such instance one might then want to consider the peculiar features in order to use the strongly indicative item-sets (attribute values) of the series as a yardstick for crime control prioritisation. The pattern space enumeration (PSE) reveals such strongly indicative features or attribute values.

3.3.2 CS Pattern Space Enumeration

In the crime series detection problem, our investigation is driven by two main concepts for series identification, which are: (i) seasonality of crime event, and (ii) defining principal feature set. That is most crime patterns exhibit at least a k minimum principal-set that characterise the modus operandi of the offender(s) behaviour. Moreover, in addition to the PDE of a series, it is interesting to know actual features that characterise the MO of a series, which relates to the PSE. Therefore, understanding for example that the probability that series-A is often instigated by gender status could be a (hot prey) pointer to the fact that such gender category needs to be watchful in the areas where series A is dominant or prevalent. Moreover, recognising or identifying other intrinsic properties could lead to an investigation of regional differences between crime scenes (i.e. particulars of the crime location). The MO of a series may be perceived as a direct consequence of the following two factors considered from Figures 3.11 and 3.12 :

- Site productivity (i.e. crime attractors).
- Behavioural themes and organisational qualities of offenders.

That is, a typical serial offender would have an organised way of perpetuating his crime, which is further characterised or determined by the crime attractors at that location. Crime attractors refer to sites, properties, objects, and locations which offenders are very familiar with and offer several criminal opportunities. We are interested in identifying the collective strength of attributes in a series and perhaps what actually uniquely characterise its MO (peculiar features).

Definition 10. (Strongly Indicative Item-set ($\mathcal{SI}$)) An item-set I of attribute values is said to be strongly indicative in a cluster C^i if at least half the cluster members possess all attribute values in I. Thus the following conditions hold:

1. The indicative strength or measure $\mathcal{SI}(I)$ of the item-set is at least J. That is the support count, say K, is $\geq J+1$, where $J = \frac{|C^i|}{2}$
2. "Hereditary" property holds. suppose an item-set (night, masked, black-male) is frequent in a particular series, then every subset of the item-set is also frequent. Conversely, if an itemset is infrequent, then all its supersets should obviously be infrequent.

3.4 Evaluation Metrics

In order to quantify the quality of series patterns generated and reliability of the model, the evaluation metrics considered for the experimental results are in two folds:

1. Statistical accuracy metrics (SAM): mean absolute error (MAE) is an established SAM. It measures the object-level performance of the model by comparing the observed cluster scores to the actual expected scores. That is the proportion of expected clusters correctly detected by CriClust, as shown in Equation (3.17), where $m_i - n_i$ measures the difference between expected and observed cluster scores. A low MAE value implies a high accuracy and vice-versa. We examined this across the 40 locations considered and further

employ n-fold cross-validation which derive a more accurate estimate of model prediction performance by combining (averages) measures of the model.

$$\text{MAE} = \frac{\sum_{i=1}^{N} |m_i - n_i|}{N} \tag{3.17}$$

2. Precision and recall measure: while the PDE (that is % difference) is useful for providing an object-level information of the series cluster, we recognise that it is not sufficient to take this as a final consideration. What is important is to examine the clusters whether they define grouping of the similar crimes; that is have high intra class similarity and low interclass similarity according to some similarity condition. Thus, at the pattern level we evaluate the precision (Prec) and recall (Rec) performance of the model satisfying the strongly indicative itemset $\mathcal{SI}$ as shown in Equations (3.18) and (3.19). Where $|C^i|$ refers to cardinality of crimes in a pattern and c_i is an instance in the pattern identified by CriClust. P_f is the peculiar features for pattern i and |PSE| is the entire series pattern derived at a location.

$$\text{Prec} = \frac{1}{|C^i|} \sum_{i=1}^{|C^i|} (c_i \in \text{PSE}^i \vdash (\mathcal{SI}_{(P_f)})). \tag{3.18}$$

$$\text{Rec} = \frac{1}{|\text{PSE}|} \sum_{i=1}^{|C^i|} (c_i \in \text{PSE}^i \vdash (\mathcal{SI}_{(P_f)})). \tag{3.19}$$

These metrics are established statistical measures in information retrieval.

3.5 CriClust Algorithm

Algorithm 4 describes the major phases in the criclust model.

Algorithm 4 Crime Series Detection (CriClust) Algorithm

Require:

- Spatial framework with features and instances embedded within the framework
- A similarity relation, S_f
- Some prevalence threshold, $(\mathcal{S}, \mathcal{P})$

Ensure: Generate all CSP with interest measure $\vdash (\mathcal{S}, \mathcal{P})$

1: Initialise threshold parameter, $(\mathcal{S}, \mathcal{P})$

2: Learn a similarity graph based on a join under similarity relation S_f on the spatial dataset

3: Do While (stopping criteria not met)

// Clustering cycle loop

4: Derive the HCS

// get HCS based on AGS-mincut

5: Do until (each $\lambda(C^i)$ is examined)

// M.O investigation

6: Series cluster (CSP) update

//Analyse CSP, location (l), estimate PDE

7: C^i= sort by series type and prevalence locality (CSP) //to enhance visualisation

8: **for all** $C^i \in CSP_l \geq 2$ **do**

9: Estimate P_ϵ

//dominating power: # instance in a series

10: D = $\max_i P_\epsilon(\mathcal{C}^i)$ //unique max value for dominant series

11: **if** count$(D) = 1$ **then**

12: $|\mathcal{CS}_d| = D$

13: Else $|\mathcal{CS}_d| = \infty$ //possibility of a tie between PDE of series

14: **end if**

15: **end for**

16: $get\ PSE$ (pattern level information)

17: Determine peculiar features on $\lambda(C^i)$ // defining features

18: **return** CS_d, $C^i_{(l)}$; // potential series across locations.

3.6 Chapter Summary

In this chapter, the model formulation and properties were presented. This research focuses on identifying crime series, which is considered to be a smart way of achieving crime deterrence, particularly in resource constrained settings as there is evidence that most crimes happen as a result of repeat offenders. Let C be a set of crime items or objects, where each crime object, say $C_i \in C$, is defined by a set of attributes A. Our interest lies in crime objects that exhibit a coherent pattern on a subset of attributes of A. The model specification uses a dual threshold scheme namely: significance (S) and prevalence (P) thresholds, to guide the symmetric similarity relation, which is used to derive or learn a similarity graph. Clusters are then identified using the similarity graph by adopting an adaptive graph contraction method that borrows from the established Karger-Stein method of identifying a good partition (min-cut) in a graph. Two interest similarity measures are then used to further explore the derived clusters and present actionable knowledge for assisting public safety agencies in crime deterrence. The interest similarity measures are: (i) Proportion Difference Evaluation (PDE), which reports the propagation effect of a series; and (ii) Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series, in order characterise modus operandi. Lastly, we present the evaluation metrics used in this research. The results of the CriClust experiment are documented in the subsequent chapter.

Chapter 4

RESULTS, DISCUSSIONS AND EXPERIMENTAL EVALUATION

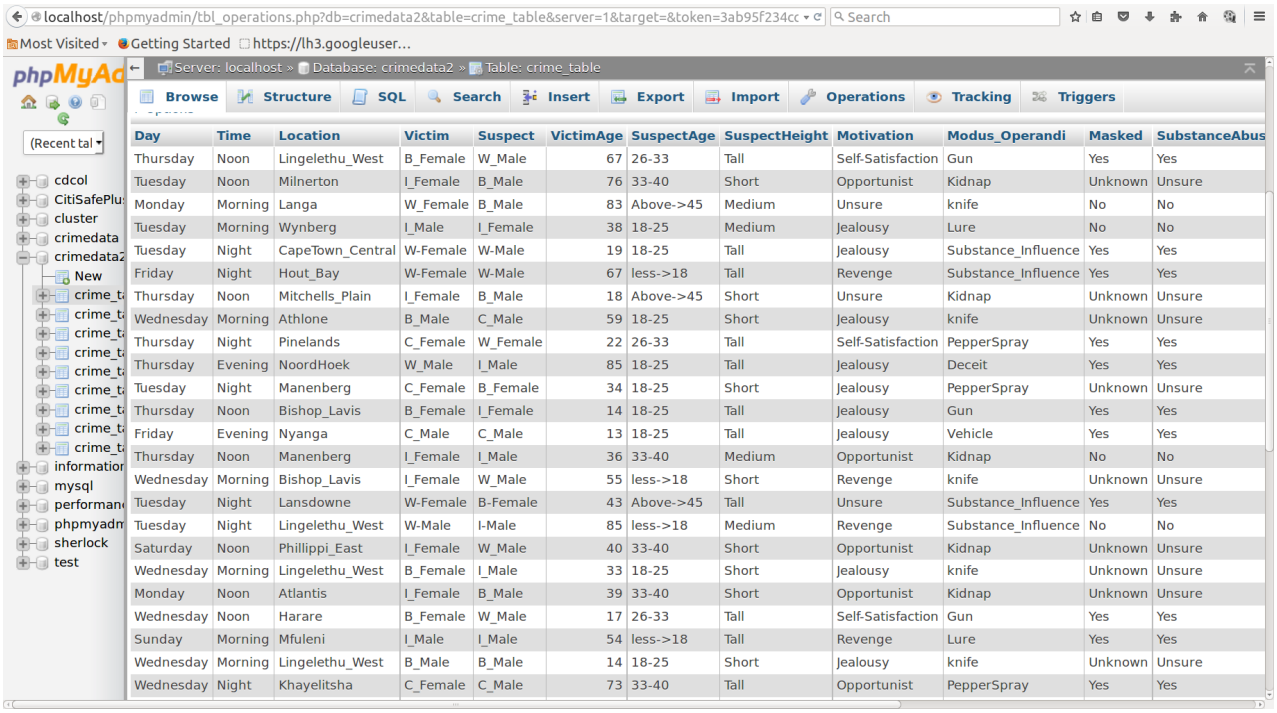
CriClust is designed to identify patterns of crime thought to have been committed by repeat offenders. To establish the reliability and effectiveness of CriClust in promoting safe cities and assisting public safety agencies in knowledge support for crime control, experimental results are presented in this chapter. A summary of the nature of data considered as well as a "controlled" experiment performed with CriClust is documented herein. Furthermore, the comparison of CriClust with competing baselines and common clustering techniques is also presented.

4.1 Sample Rape Data For Experiment

In the past, the only gender that came to mind as a potential victim whenever a rape incident occured was a female. However, in 2007 the previous definition of rape was revised by the Sexual Offences Amendment Act of South Africa [93] to include *"all forms of sexual penetration without consent irrespective of gender"*. This means that a "male" rape is possible. Hence, the rape data for experiment has the male gender attribute as part of potential suspect as well as victim. The attributes include day, time and location information, the capture

method involved in the offending process, whether the suspect was disguised (for example, masked), traces of substance-abuse and victim information.

Figure 4.1 depicts the sample rape dataset and attributes considered for experiment. It is important to note that while this research explores a rape database for identifying crime series, the idea considered in this study can be extended to other forms of crime. The motivation for considering a rape database is the fact that despite the incessant efforts in tackling sexual assault and violence, South African communities happen to be a place where the rape of people (and particularly women and children) is common [77]. Further information about the attributes and the database is described and documented in section 3.1.1. We carefully and deliberately try to focus our analysis on relevant attributes that will assist the analysis according to crime expert advice.



Day	Time	Location	Victim	Suspect	VictimAge	SuspectAge	SuspectHeight	Motivation	Modus_Operandi	Masked	SubstanceAbus
Thursday	Noon	Lingelethu_West	B_Female	W_Male	67	26-33	Tall	Self-Satisfaction	Gun	Yes	Yes
Tuesday	Noon	Milnerton	I_Female	B_Male	76	33-40	Short	Opportunist	Kidnap	Unknown	Unsure
Monday	Morning	Langa	W_Female	B_Male	83	Above->45	Medium	Unsure	knife	No	No
Tuesday	Morning	Wynberg	I_Male	I_Female	38	18-25	Medium	Jealousy	Lure	No	No
Tuesday	Night	CapeTown_Central	W-Female	W-Male	19	18-25	Tall	Jealousy	Substance_Influence	Yes	Yes
Friday	Night	Hout_Bay	W-Female	W-Male	67	less->18	Tall	Revenge	Substance_Influence	Yes	Yes
Thursday	Noon	Mitchells_Plain	I_Female	B_Male	18	Above->45	Short	Unsure	Kidnap	Unknown	Unsure
Wednesday	Morning	Athlone	B_Male	C_Male	59	18-25	Short	Jealousy	knife	Unknown	Unsure
Thursday	Night	Pinelands	C_Female	W_Female	22	26-33	Tall	Self-Satisfaction	PepperSpray	Yes	Yes
Thursday	Evening	NoordHoek	W_Male	I_Male	85	18-25	Tall	Jealousy	Deceit	Yes	Yes
Tuesday	Night	Manenberg	C_Female	B_Female	34	18-25	Short	Jealousy	PepperSpray	Unknown	Unsure
Thursday	Noon	Bishop_Lavis	B_Female	I_Female	14	18-25	Tall	Jealousy	Gun	Yes	Yes
Friday	Evening	Nyanga	C_Male	C_Male	13	18-25	Tall	Jealousy	Vehicle	Yes	Yes
Thursday	Noon	Manenberg	I_Female	I_Male	36	33-40	Medium	Opportunist	Kidnap	No	No
Wednesday	Morning	Bishop_Lavis	I_Female	W_Male	55	less->18	Short	Revenge	knife	Unknown	Unsure
Tuesday	Night	Lansdowne	W-Female	B-Female	43	Above->45	Tall	Unsure	Substance_Influence	Yes	Yes
Tuesday	Night	Lingelethu_West	W-Male	I-Male	85	less->18	Medium	Revenge	Substance_Influence	No	No
Saturday	Noon	Phillippi_East	I_Female	W_Male	40	33-40	Short	Opportunist	Kidnap	Unknown	Unsure
Wednesday	Morning	Lingelethu_West	B_Female	I_Male	33	18-25	Short	Jealousy	knife	Unknown	Unsure
Monday	Noon	Atlantis	I_Female	B_Male	39	33-40	Short	Opportunist	Kidnap	Unknown	Unsure
Wednesday	Noon	Harare	B_Female	W_Male	17	26-33	Tall	Self-Satisfaction	Gun	Yes	Yes
Sunday	Morning	Mfuleni	I_Male	I_Male	54	less->18	Tall	Revenge	Lure	Yes	Yes
Wednesday	Morning	Lingelethu_West	B_Male	B_Male	14	18-25	Short	Jealousy	knife	Unknown	Unsure
Wednesday	Night	Khayelitsha	C_Female	C_Male	73	33-40	Tall	Opportunist	PepperSpray	Yes	Yes

Figure 4.1: Sample rape database for experiment

4.2 Empirical Observations on Experimental Data

We provide considerable empirical evidence to support our arguments and model derivations, and present statistically interpretable crime series patterns for actionable solution in a qualitative and quantitative manner. This is fundamentally important as patterns in the clusters need to exhibit some level of coherence based on a defined similarity measure; requiring high intra-cluster similarity and low inter-cluster similarity according to defined similarity condition. Furthermore, clusters that cannot be interpreted will not assist public safety agencies in effective decision making and crime deterrence targets. First, we ensure that crimes with similar attributes are well identified and connected according to our experts' knowledge and through the use of a similarity function and dual threshold mechanism, which results in a similarity graph. Second, we ensure that the selection of a potential partition that ensures the recovery of sufficiently connected edges in the crime similarity graph is uniformly chosen from the available set of possible choices, with minimal constraints. The process is further reinforced using the Monte-Carlo approach and heuristics to minimise singleton clusters where possible. This is achieved using our adaptive graph size mechanism. Our confidence with the generated patterns lies in the promising empirical observations, and was substantiated with the optimistic reaction and input we received from experts in this domain.

The experimental set-up was implemented using Java NetBeans platform with multi-threading, Apache Derby Network Server 10.10.2.0., with security manager installed using the basic server security policy. The experiment was conducted on an Inspiron-7347 DELL machine, Intel(R) Core(TM) i5-4210U CPU @ 1.70GH. The data considered for the experiment consist of 5500 rape crime records across 40 locations (suburbs) in Western Cape, South Africa, comprising of 13 attributes of relevant features as prescribed by the crime intelligence unit. In what follows, we present some results to show the potential usefulness and reliability of the CriClust model.

4.2.1 Background Level Cluster Information: Graduated Colour Map of Crime Information

Among other functionalities and solution, the system offers a high level view of information concerning the level of crime committed at a particular location, adapted from [63]. For example, Figure 4.2 presents a graduated

colour map of crime locations (synonymous to hotspots; to examine geographic areas in relation to crime) in the Western Cape, South Africa. The crime density levels are depicted using five different categories and the crime density information was derived by apportioning corresponding percentage evaluation of contributing spatial crime information per location.

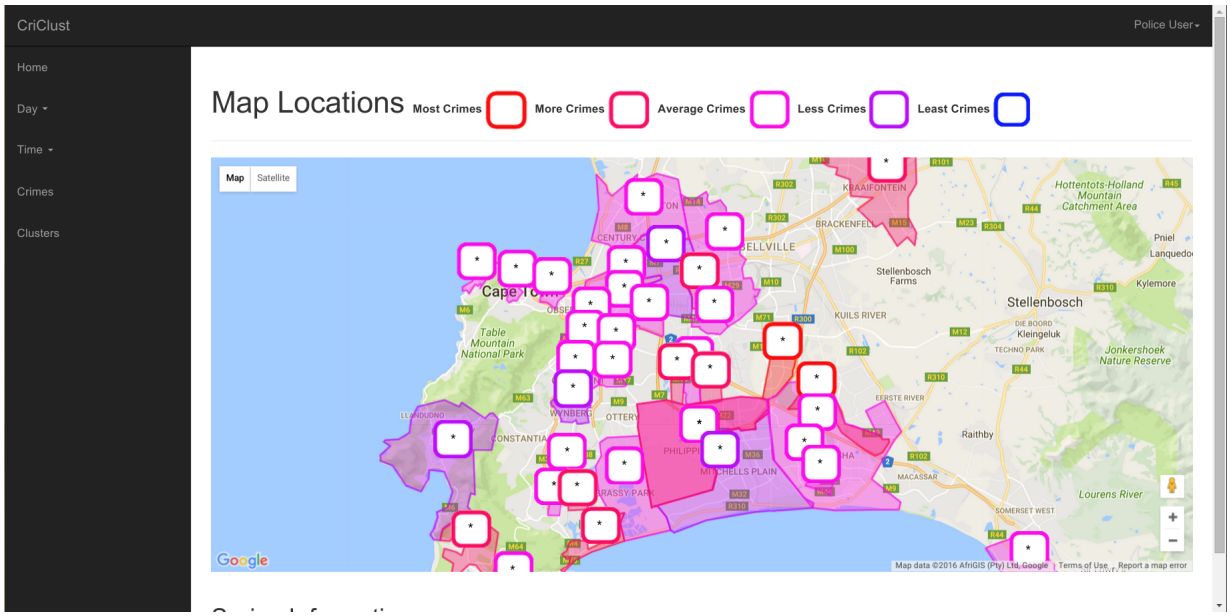


Figure 4.2: graduated colour map of identified locations of crime

Clicking each node in the map (Figure 4.2) further reveals the associated spatial crime information (statistic) for any suburb in Western Cape as shown in Figures 4.3, 4.4 and 4.5. This type of information corresponds to part of what the police consider standard product analysis, where they try to derive basic cluster information on different categories of crime committed. This standard product analysis further involves identifying problematic times of day, days of week and locations.

For example, Figures 4.3, 4.4 and 4.5 presents the general background level information which corresponds to varying crime density levels at three and six different locations in Western Cape, South Africa, respectively. Moreover, the three locations (Nyanga, HoutBay, Strand) in Figure 4.3 have different crime density levels of 31, 18 and 25 respectively, on the 1000 records of crime examined. Similarly on 1500 crime records examined, another set of three locations (Milnerton, Dieprivier, Mfuleni) highlighted in Figure 4.4 presents the corresponding statistics.

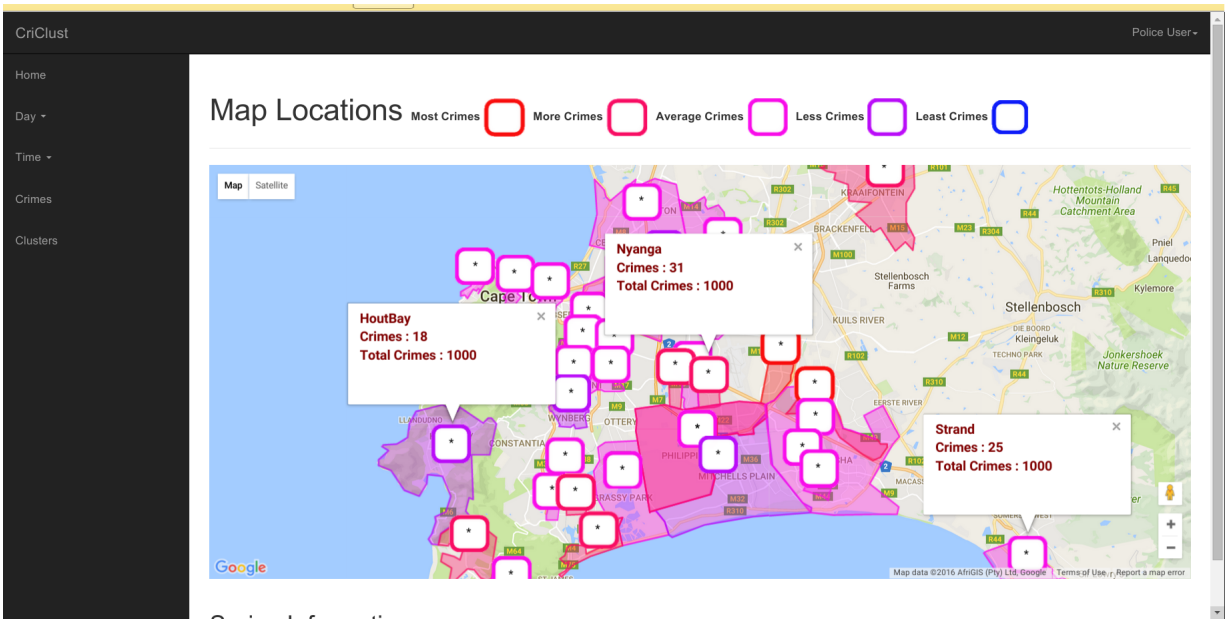


Figure 4.3: The locations of crimes with corresponding crime densities: graduated colour map of cluster information on 1000 records

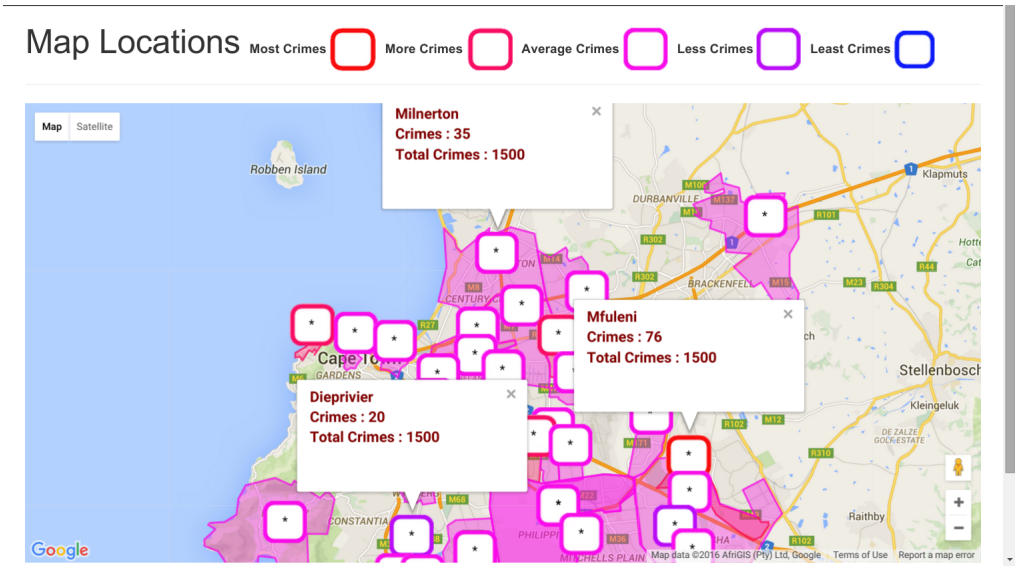


Figure 4.4: The locations of crimes with corresponding crime densities: graduated colour map of cluster information on 1500 records

In general, we note that as the number of records increased, the crime levels also rapidly increased across locations (see Figure 4.6), which is not an unusual trend. Furthermore, the spatial information in Figures 4.3, 4.4 and 4.5 provides a quick high level information about important areas requiring immediate intervention from security

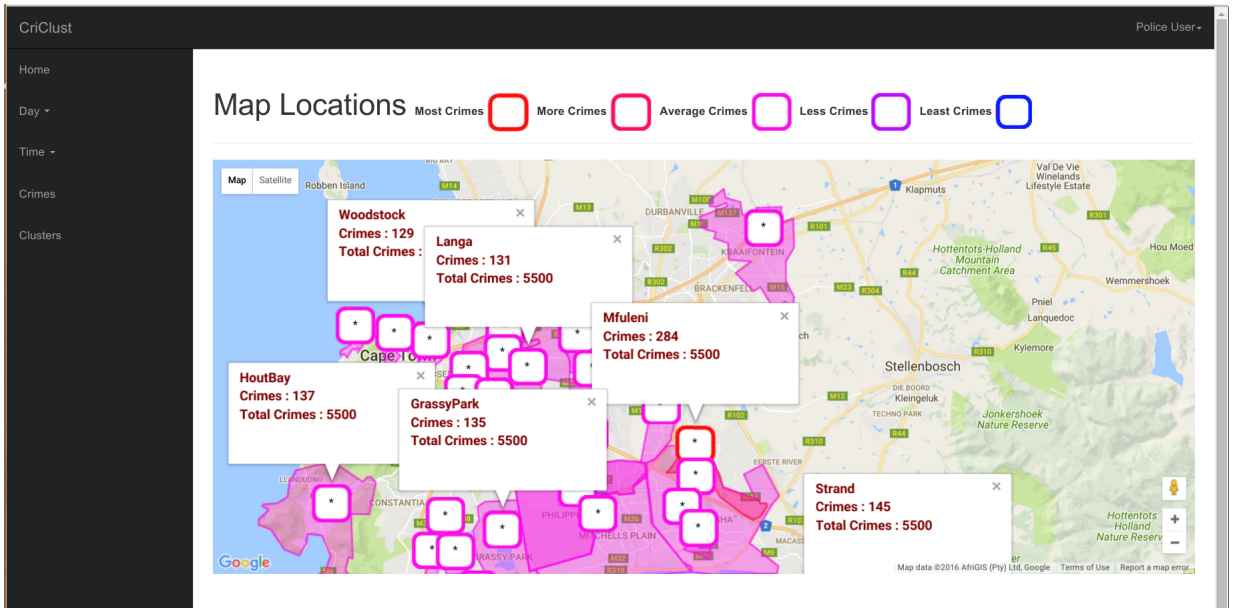


Figure 4.5: The locations of crimes: graduated colour map of cluster information on 5500 records

agencies.

From an operational perspective, these varying density levels provide quick background-level insightful information that enables public safety agencies to promptly discover critical areas demanding attention. However, it is worth noting that such general background-level information may not be directly actionable for crime deterrence, while identifying the exact MO of offenders is directly actionable [9].

It is thus worth mentioning that the background level information provided is only an additional feature of the CriClust system since the major interest and focus is on series cluster identification. Series identification entails isolating individual crimes, by revealing the exact correlation between features or attributes involved in such cluster. Moreover, such correlation helps to clearly identify the MO of the perpetrators. Hence, identifying exact features in patterns (that is raw data rows in crime series) is of practical importance for actionable information in promoting citizen-centred safety. In what follows we present empirical results on series identification with CriClust.

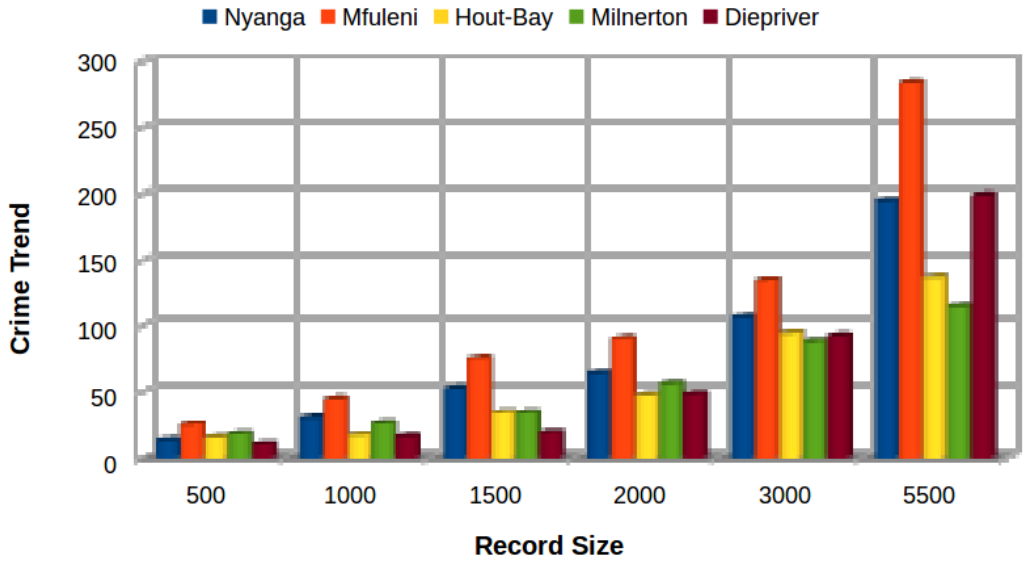


Figure 4.6: Crime trend across locations with varying record size

4.3 Empirical Analysis of Series Information Across Locations

Crime series refers to crimes thought to have been perpetuated by the same offender [78], that is repeat offender(s). Furthermore, research has shown that many crimes that happen regularly across different locations (perhaps on the order of 60 %) could have been committed by a serial criminal [30, 57]. This makes the identification of potential crime series of practical relevance and critical importance for public safety. In what follows, we present results on how CriClust attempts to promote proactive policing to achieve crime control targets in a smarter way.

4.3.1 Identified Series Information Across Locations

CriClust uniquely presents cluster information in a manner that can be understood by a novice public safety personnel, with no expert domain knowledge. Figure 4.7 presents the identified locations with at least one series. This gives a quick high level insightful information on areas with repeat offenders. However, it is worth mentioning that the map is able to reveal more information about the series clusters. This graduated colour map can show the following for any suburb:

- the number of series at a particular location (as seen in Figure 4.7).

- proportion difference evaluation (PDE) across series identified at a specific location (a pie chart with % per series), that is the propagation effect of each series.
- pattern space enumeration (PSE), revealing attribute information and the peculiar features that characterise a particular series.

These details are further explored in the succeeding sections to express their potential and usefulness in actionable knowledge support.

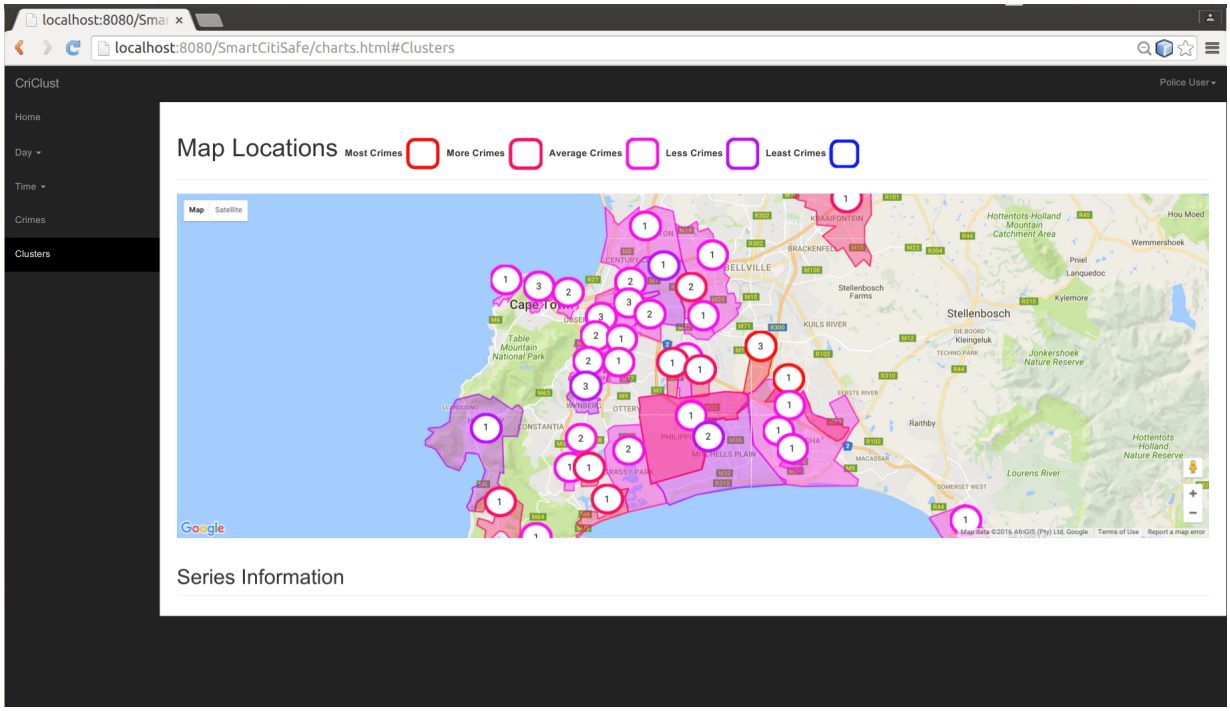


Figure 4.7: The locations of crimes series.

4.3.2 PDEs of Identified Series Information Across Locations

Figure 4.7 presents the number of series identified at locations. At the locations with more than one series it is interesting to know the dominant series amongst them as well as factors that contribute to the spatial differences (i.e. MO) at those locations. Thus, the PDE helps to differentiate or classify the series at that location by revealing the proportion difference (%) of the series based on the propagation effect of each series as presented in Figure 4.8. The propagation effect tells us which of the series has a high dominating power (dominant series) at

a particular locality. This measure can help to guide decisions as to which series to track down first. While Figure 4.7 presents locations of crimes with at least one ongoing series, wherein node values indicate the number of identified series at the corresponding location, Figure 4.8 reveals an instance of PDEs associated with two series at a particular location. Thus, clicking each node in the map (Figure 4.7) reveals the corresponding PDEs for the location highlighted as shown in Figure 4.8, which reveals that there are two ongoing series at the highlighted location (Mowbray) with PDEs 35% (series 4) and 65% (series 3) respectively. In this instance, it is clear that "series 3" is the dominating (dominant) series, which should ideally be tackled first.

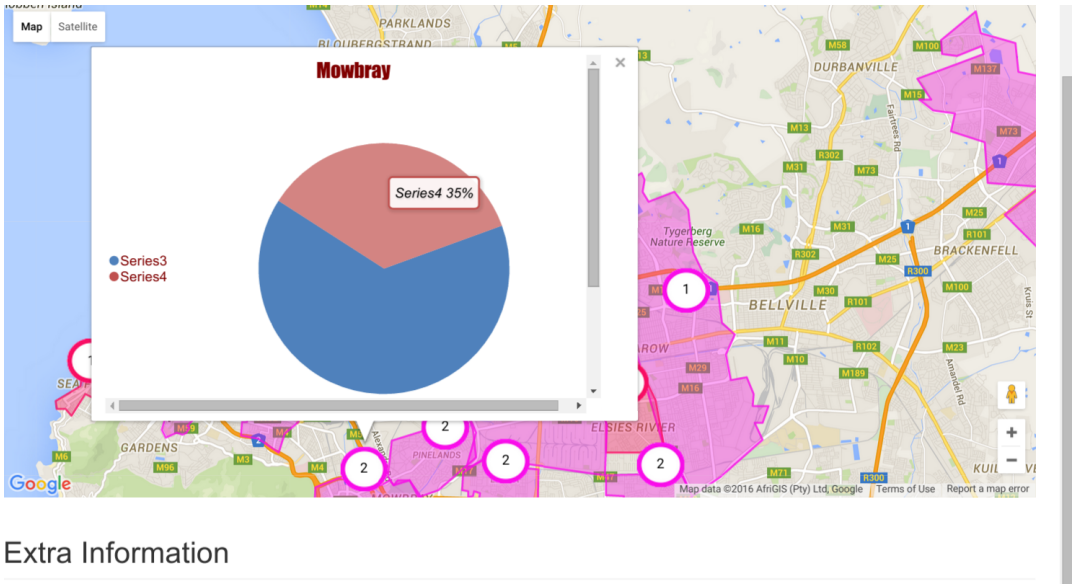
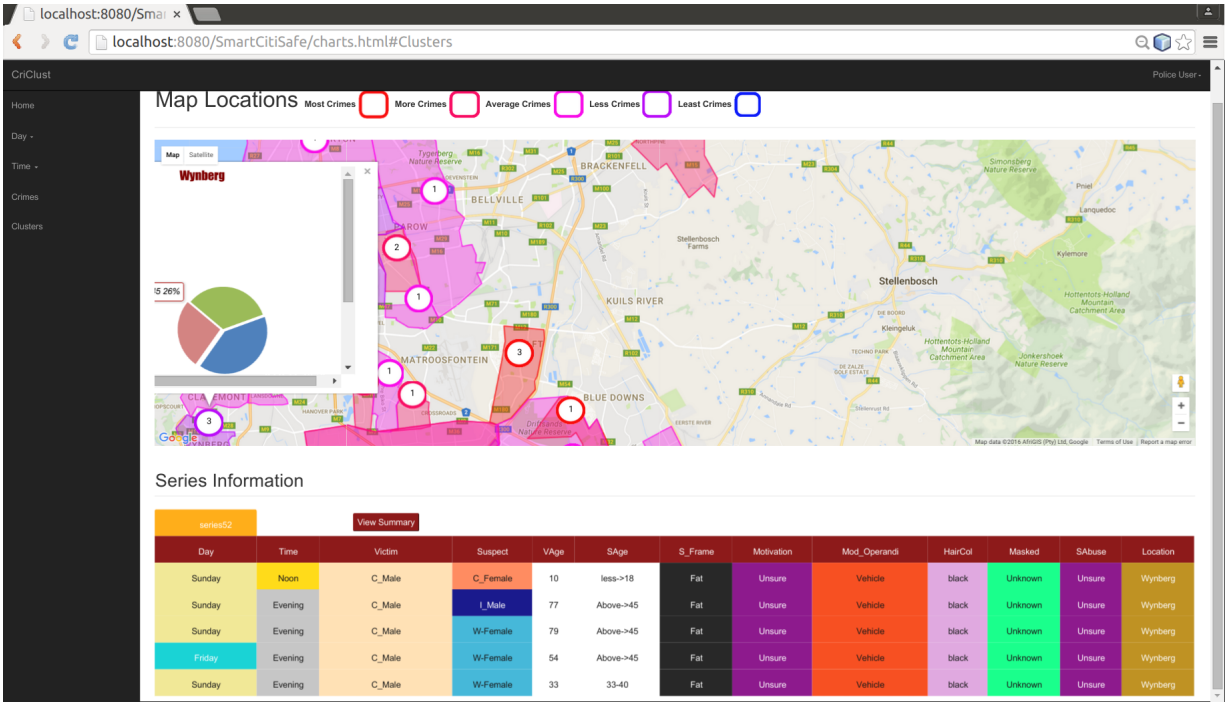


Figure 4.8: Visualisation of the propagation effect and proportion difference of corresponding series at Mowbray.

Furthermore, Figure 4.9 shows another instance of three series identified at Wynberg location, with PDEs 26 %, 34 % and 40 % respectively. Note that Figure 4.9 reveals a section of raw-data information (PSE) for an instance of the highlighted series. Generally, clicking the series PDE information (that is the pie chart) reveals the pattern space information of the corresponding (PDE) series at that location. The pattern space enumeration (PSE) gives much more in-depth information about attribute values characterising a series

It is more useful and proactive for public safety and security agencies to: (i) not just identify specific locations with crime series (as seen in Figure 4.7), but to also (ii) get the corresponding PDEs; and (iii) understand the specific MO involved in each of the series, which PSE reveals.



Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Sunday	Noon	C_Male	C_Female	10	less->18	Fat	Unsure	Vehicle	black	Unknown	Unsure	Wynberg
Sunday	Evening	C_Male	I_Male	77	Above->45	Fat	Unsure	Vehicle	black	Unknown	Unsure	Wynberg
Sunday	Evening	C_Male	W-Female	79	Above->45	Fat	Unsure	Vehicle	black	Unknown	Unsure	Wynberg
Friday	Evening	C_Male	W-Female	54	Above->45	Fat	Unsure	Vehicle	black	Unknown	Unsure	Wynberg
Sunday	Evening	C_Male	W-Female	33	33-40	Fat	Unsure	Vehicle	black	Unknown	Unsure	Wynberg

Figure 4.9: Visualisation of the propagation effect and proportion difference of corresponding series at Wynberg.

A notable characteristic and observation emanating from these results and worth re-iterating is the fact that despite the high crime density levels identified at some locations that are originally categorised (colour-coded) as "most" crimes (see Figures 4.3, 4.4 and 4.5), the number of series identified at some of those locations is relatively lower or smaller than those found in some other locations initially classified as areas with "less" or "least" crimes. For example, observe in Figure 4.7 that some areas colour-coded as "most" crimes have only 1 or 2 ongoing series identified there, while some other ones with "less" crimes have up to 3 ongoing series (see Table 4.1). This indicates that a serial crime may not necessarily happen at hotspot locations, as rightly established in the introductory section of this thesis, which makes series identification a unique problem. Moreover, there is also a possibility that some of the locations categorised as hotspots potentially have many opportunist offenders perpetuating crimes there, as explained by the law of criminology.

Furthermore, we note that when the crime records increases the number of series identified across most of the locations remains as it was (2 or 3 series). This means that increase in crime record does not necessarily always imply increase in the number of identified series at the locations or emergence of a new series, as depicted in Figure 4.10 and Table 4.2. Observe the overlap in Figure 4.10 as the identified number of series is consistently

Table 4.1: Locations revealing that series may not necessarily happen at hotspots ("most crimes") locations.

Locations	Categorisation	No of Series Identified
Wynberg	Less crimes	3
Observatory	Most crimes	3
Parow	Most crimes	2
Mowbray	Average crimes	2
Blues-Downs	Most crimes	1

maintained across some on the locations with varying record set. For example, Wynberg maintained two series between 2500 and 3500 records (overlapping with GrassyPark and CapeTown-central locations) before a third series (intersecting with Observatory) emerged at 5500 record. However, it was observed that the intensity level of some of the series (PDEs) increased and varies as crime record increases (Table 4.2). This suggests that a series that emerges with a particular pattern almost always retains or sticks to the pattern, but could now have a higher level of instances in its pattern, which points to increase in the propagation effect of the series at the corresponding location. Thus, knowing the exact characterising features ($PSEandP_f$) for a series is useful since the pattern is what is basically propagated as crime increases, hence intervention strategies can be planned around these features.

Table 4.2: Trend of PDEs across locations with increasing crime record.

S/N	Location	RecordSize	Number of Series	PDEs (%)		
				S1	S2	S3
1	Mowbray	1500	2	85	15	-
		2500	2	67	33	-
		3500	2	53	47	-
		5500	2	35	65	-
2	Wynberg	1500	1	100	-	-
		2500	2	40	60	-
		3500	2	45	55	-
		5500	3	40	34	26
3	CapeTown Central	1500	2	35	65	-
		2500	2	41	59	-
		3500	2	47	53	-
		5500	2	50	50	-
4	Grassy Park	1500	2	50	50	-
		2500	2	64	36	-
		3500	2	69	31	-
		5500	2	79	21	-
5	Observatory	1500	2	46	54	-
		2500	2	40	60	-
		3500	3	40	35	25
		5500	3	32	38	30

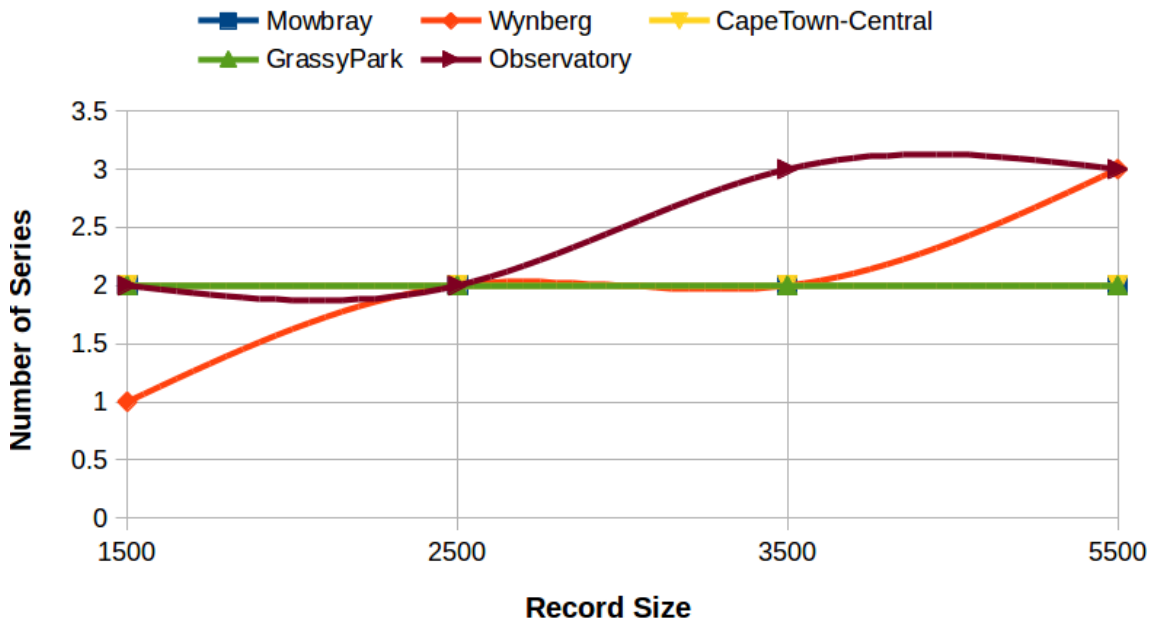


Figure 4.10: Trend of series observed across locations with varying data size

4.3.3 CriClust PSE Identified Across Locations

The pattern space enumeration (PSE) is useful for identifying the nature of the contributing factors to crimes committed at a location, as well as understanding the specific MO the perpetrator is using. This is expedient because beyond the high level numerical information, it is possible to visualise and explore exact attribute values in a particular series cluster. Figures 4.13 and 4.14 present two series identified for Cape Town central. Each series has a PDE of 50 %. This is interesting because both series are "equally likely" series and thus there is no significant dominating series at that location. Thus the dominant series takes on the value ∞. In this instance we do not currently know which of the series is dominant since there is a tie between them. However, it is possible that as crime level further increases in that area one of the series could emerge as the dominant series. This is a typical scenario in which PSE shines as it reveals more information about the nature and sensitivity of attributes involved in such series. By sensitivity of attributes we mean the level of implications and negative impact of the peculiar features emerging in the series. The peculiar features are the strongly indicative item-set (attributes) in a series. For example, a series that has most of its victims as females that are within the teenage age-category could suggest that a serial rapist has a strong desire or passion for abusing young females. Therefore such series may need to be tackled first before another with females or males victims in the adult age-category. Furthermore,

a cluster instance with a significant level of substance abuse might also require urgent attention as people under substance influence (drugs) are highly susceptible to acting more irrationally, thereby committing more crimes with high negative impact. Moreover, research has shown that there is a complex relationship between drugs and crime [1] and that criminality and substance abuse are highly linked.

Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Friday	Night	W-Female	B_Female	28	Above->45	Moderate	Unsure	Substance_Influence	brown	No	No	Mowbray
Sunday	Noon	W-Female	B_Male	5	Above->45	Moderate	Self-Satisfaction	Substance_Influence	brown	No	No	Mowbray
Friday	Night	W-Female	B_Female	3	Above->45	Moderate	Unsure	Substance_Influence	brown	No	No	Mowbray
Friday	Night	W-Female	B_Female	22	Above->45	Moderate	Self-Satisfaction	Substance_Influence	brown	No	No	Mowbray
Friday	Night	W-Female	B_Female	10	Above->45	Moderate	Unsure	PepperSpray	brown	No	No	Mowbray
Friday	Night	W-Female	B_Female	28	less->18	Moderate	Unsure	Substance_Influence	unknown	No	No	Mowbray
Friday	Night	W-Female	B_Female	63	Above->45	Moderate	Self-Satisfaction	Vehicle	brown	No	No	Mowbray

Figure 4.11: Series cluster-1 identified for Mowbray location.

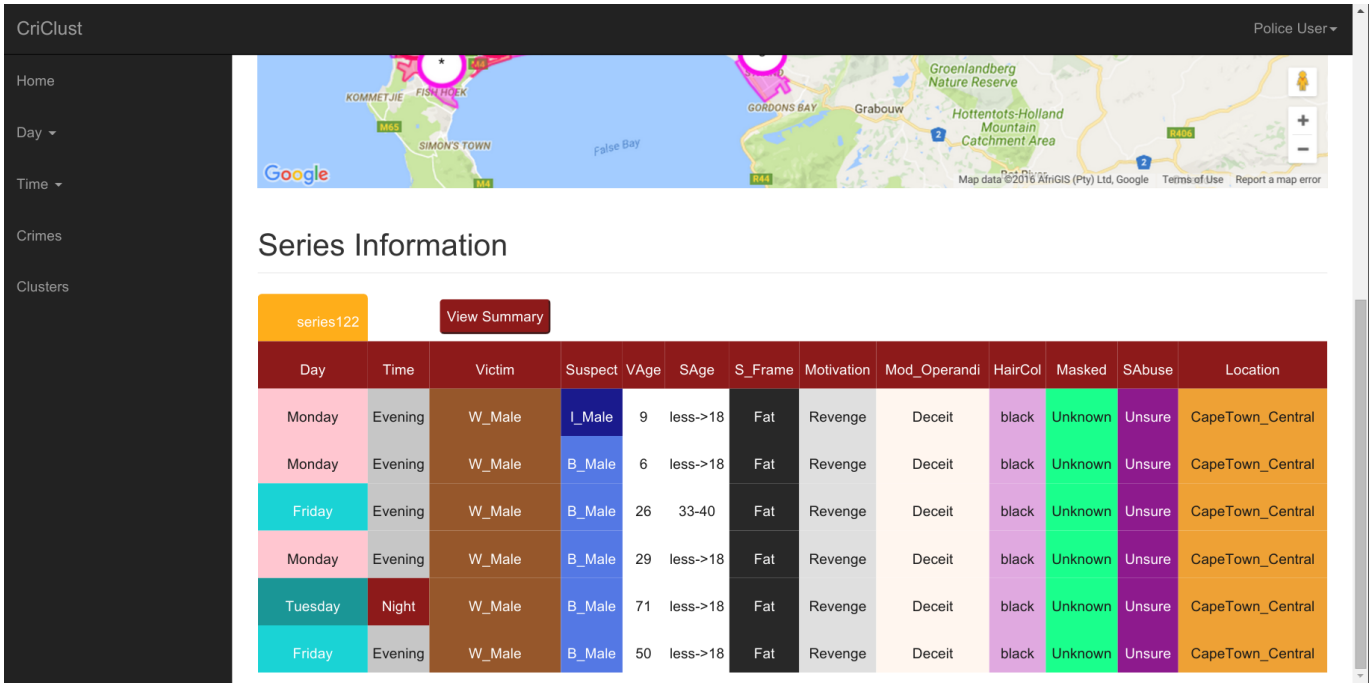
Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Thursday	Evening	I_Female	B_Male	33	33-40	Slender	Opportunist	Deceit	unknown	Yes	Yes	Mowbray
Thursday	Evening	W_Male	B_Male	14	33-40	Slender	Opportunist	Deceit	unknown	Yes	Yes	Mowbray
Thursday	Evening	W_Male	B_Male	0	33-40	Slender	Opportunist	Deceit	unknown	Yes	Yes	Mowbray
Thursday	Evening	W_Male	B_Male	54	33-40	Slender	Opportunist	Deceit	red	Unknown	Yes	Mowbray
Thursday	Evening	W_Male	B_Male	41	33-40	Slender	Opportunist	Deceit	unknown	Yes	Yes	Mowbray
Thursday	Evening	W_Male	B_Male	49	33-40	Slender	Opportunist	Deceit	unknown	Yes	Yes	Mowbray

Figure 4.12: Series cluster-2 identified for Mowbray location.

Figures 4.11 and 4.12 reveal the patterns of two different series identified at Mowbray, where one occurs mostly at night time with the suspect unmasked, the other occurs mostly at evening time with suspect masked. In the first series, however, there is effectively no trace of substance abuse by the perpetrator while a significant level of

[1] http://www.bjs.gov/content/pub/pdf/DRRC.PDF

substance abuse is perceived in the second series. These variations further confirms and points to the dynamics of offenders' behaviour and subsequently give an insight into ways to strategise in tackling specific attributes emerging in each series.

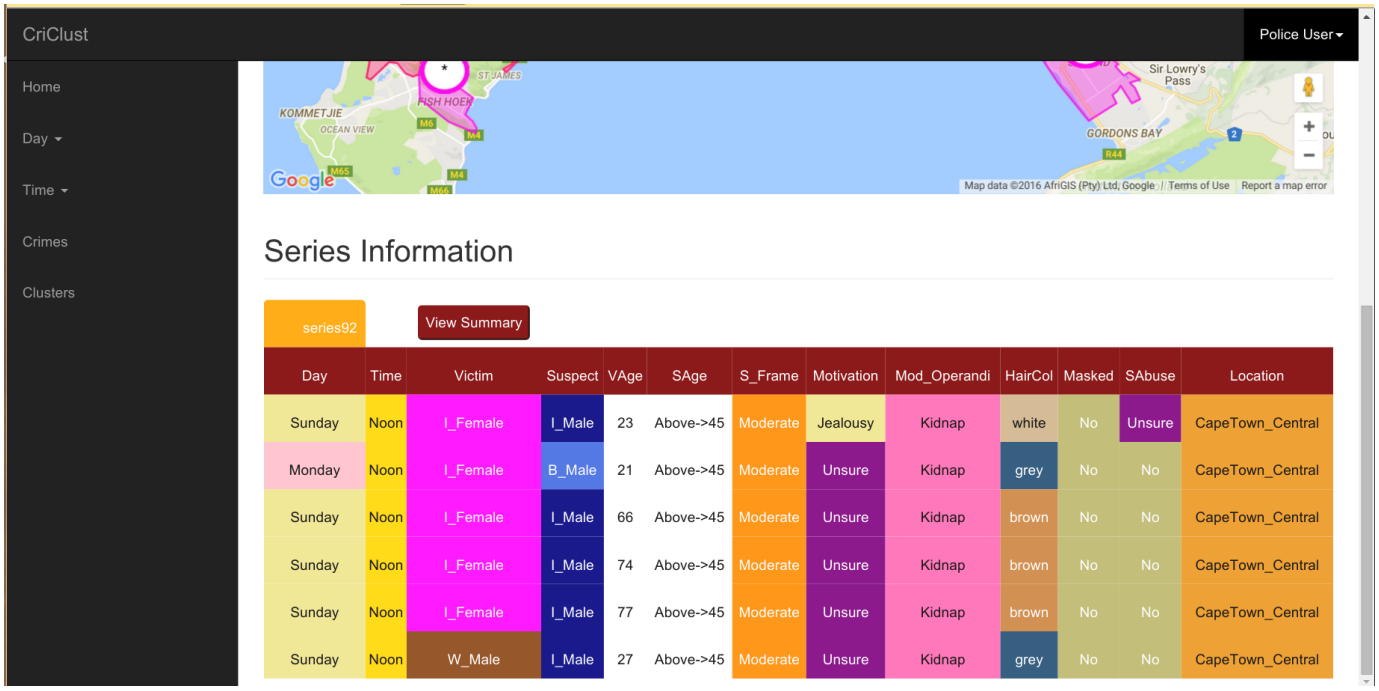


Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Monday	Evening	W_Male	I_Male	9	less->18	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central
Monday	Evening	W_Male	B_Male	6	less->18	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central
Friday	Evening	W_Male	B_Male	26	33-40	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central
Monday	Evening	W_Male	B_Male	29	less->18	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central
Tuesday	Night	W_Male	B_Male	71	less->18	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central
Friday	Evening	W_Male	B_Male	50	less->18	Fat	Revenge	Deceit	black	Unknown	Unsure	CapeTown_Central

Figure 4.13: Series cluster-1 identified for Cape-Town-Central location.

In Figures 4.13, the series is noted to be perpetrated mostly on Monday/Friday evenings by a fat black male (B-Male) targeting white males (W_Males) as victims, while the other series in Figure 4.14 is mostly perpetuated at noon time and the suspect is an Indian male (I_Male) with a moderate frame who uses kidnap method for victim capture. Furthermore, the victim in the second series is mostly Indian-females (I-female). Figures 4.15 and 4.16 reveal the peculiar (characterising) features for the two series; peculiar features for a series can be visualised by clicking "View Summary" on the CriClust system.

While these two series are prevalent in the same locality, it is important to recognise that there are slight variations in their characterising (peculiar) features. These variations further confirm the earlier proposition in this research, which anticipated that there might be slight changes in an offenders' behaviour because past learning, current targets, situational and geographical attributes typically influence each crime outcome. Thus, seeing different



Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Sunday	Noon	I_Female	I_Male	23	Above->45	Moderate	Jealousy	Kidnap	white	No	Unsure	CapeTown_Central
Monday	Noon	I_Female	B_Male	21	Above->45	Moderate	Unsure	Kidnap	grey	No	No	CapeTown_Central
Sunday	Noon	I_Female	I_Male	66	Above->45	Moderate	Unsure	Kidnap	brown	No	No	CapeTown_Central
Sunday	Noon	I_Female	I_Male	74	Above->45	Moderate	Unsure	Kidnap	brown	No	No	CapeTown_Central
Sunday	Noon	I_Female	I_Male	77	Above->45	Moderate	Unsure	Kidnap	brown	No	No	CapeTown_Central
Sunday	Noon	W_Male	I_Male	27	Above->45	Moderate	Unsure	Kidnap	grey	No	No	CapeTown_Central

Figure 4.14: Series cluster-2 identified for Cape-Town-Central location.

characteristics emerge in the MO of the series is not too surprising. However, knowing the detail information about each series is not trivial and also much more insightful for knowledge support than typical high level background information.

Furthermore, Figure 4.17 shows the series information identified at Grassy-park location. In this instance, the perpetrator is a coloured-male (C-male) and an opportunist who usually mask when committing rape crime, mostly on Tuesday mornings. The target victim here is mostly Indian-male and at one instance it was a white-male. This series could suggest, for example, a gay serial rapist (hot prey, see Table 2.1). As clearly seen in the cluster examples presented, the results also agree with the significance threshold constraint that similar crimes must have more than 6 attribute values in common, while the prevalence (space-time) condition is also enforced. Furthermore, we based these attributes on expert recommendation.

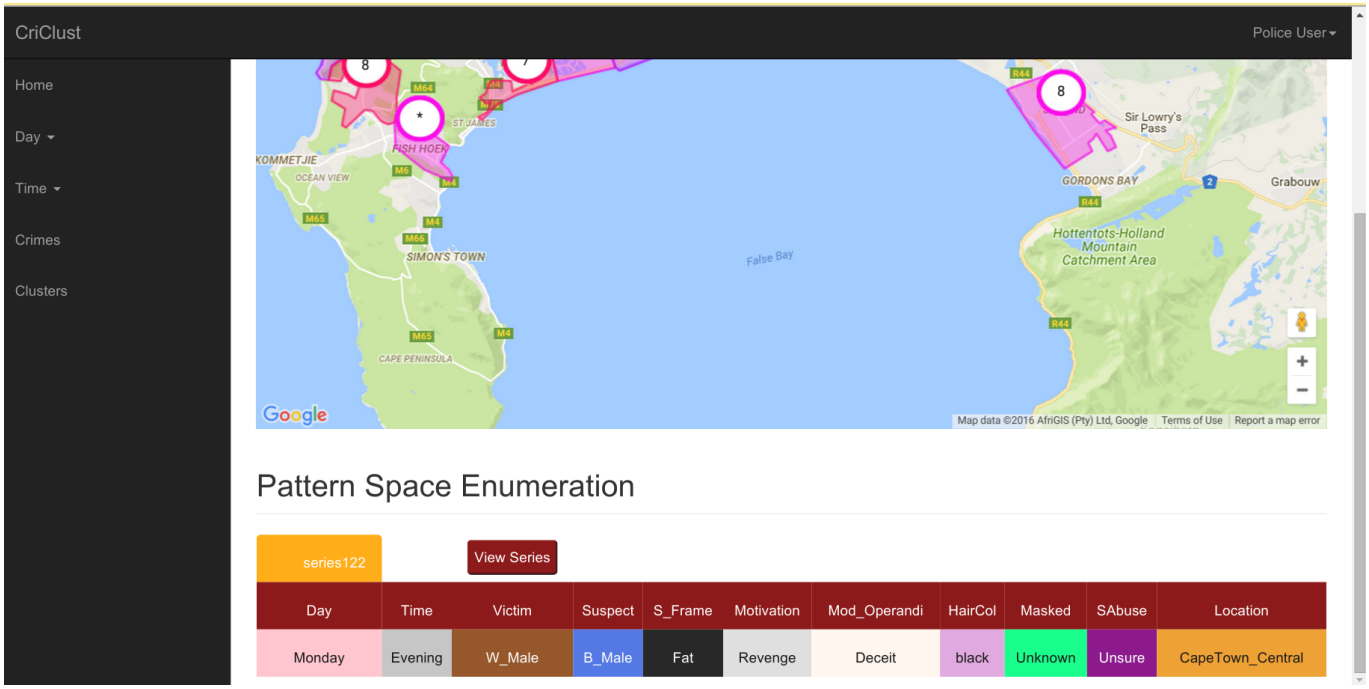


Figure 4.15: Corresponding PSE for cluster-1 identified for Cape-Town-Central location

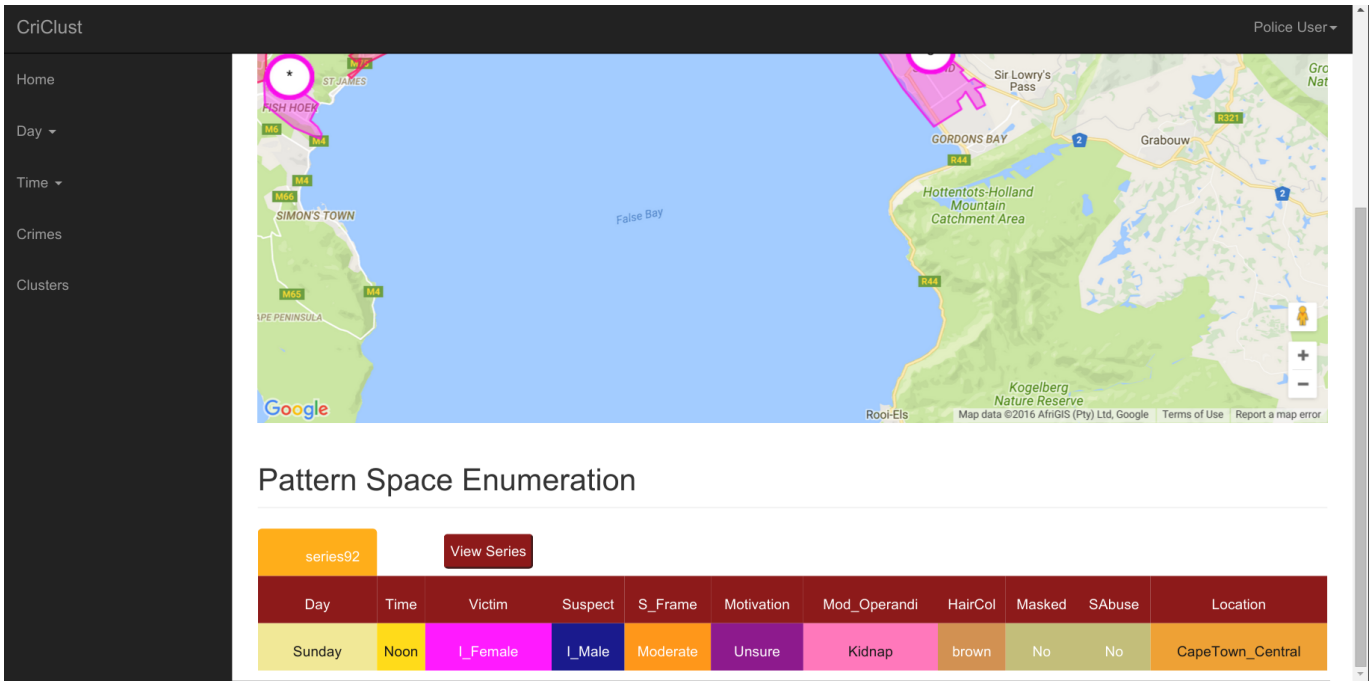
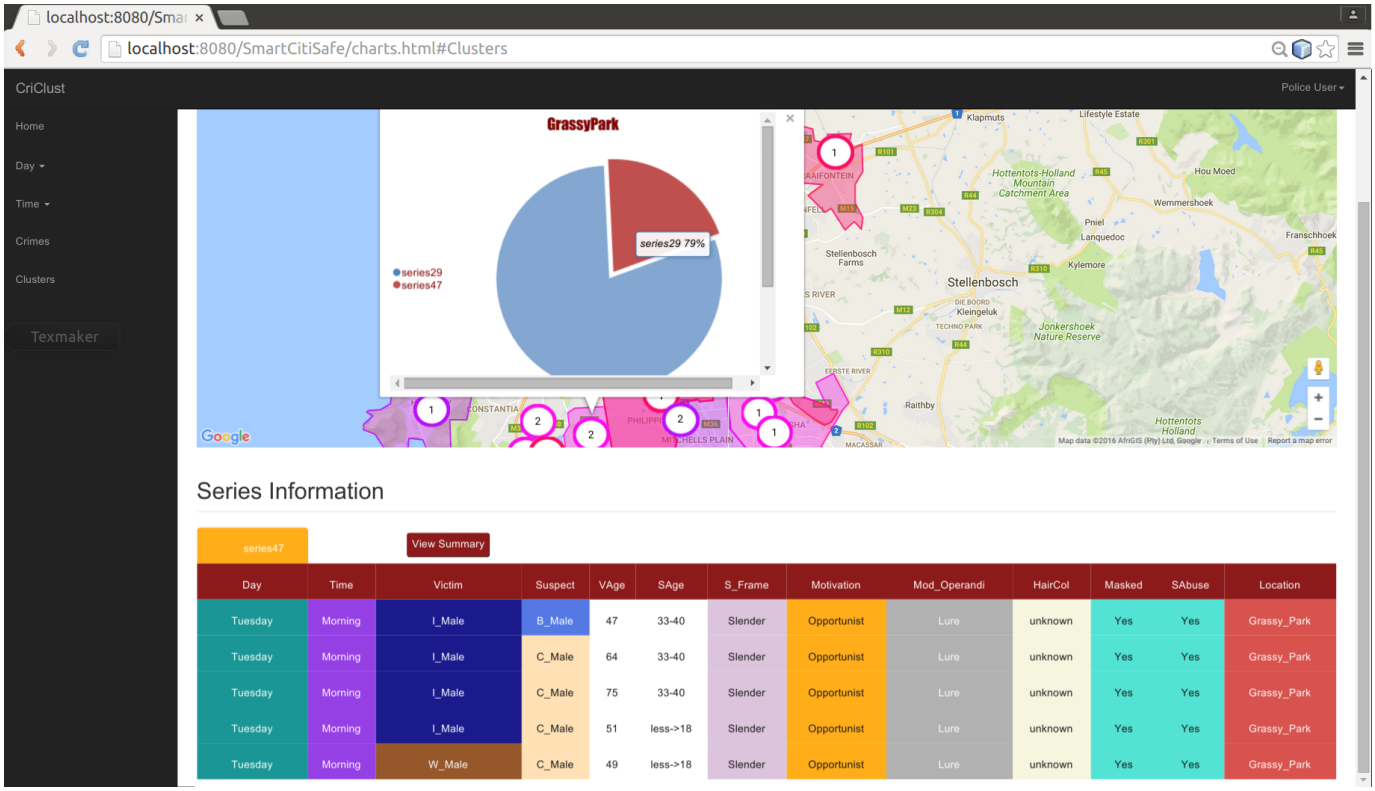


Figure 4.16: Corresponding PSE for Cluster-2 identified for Cape-Town-Central location



Day	Time	Victim	Suspect	VAge	SAge	S_Frame	Motivation	Mod_Operandi	HairCol	Masked	SAbuse	Location
Tuesday	Morning	I_Male	B_Male	47	33-40	Slender	Opportunist	Lure	unknown	Yes	Yes	Grassy_Park
Tuesday	Morning	I_Male	C_Male	64	33-40	Slender	Opportunist	Lure	unknown	Yes	Yes	Grassy_Park
Tuesday	Morning	I_Male	C_Male	75	33-40	Slender	Opportunist	Lure	unknown	Yes	Yes	Grassy_Park
Tuesday	Morning	I_Male	C_Male	51	less->18	Slender	Opportunist	Lure	unknown	Yes	Yes	Grassy_Park
Tuesday	Morning	W_Male	C_Male	49	less->18	Slender	Opportunist	Lure	unknown	Yes	Yes	Grassy_Park

Figure 4.17: A depiction of the PDE and PSE of series at Grassy-Park location

4.4 Quantitative Evaluation of CriClust Model

Table 4.3 describes the peculiar features that characterise each series, denoted $(S1, S2, S3)$. The markers "1" (presence) and "0" (absence) respectively denote emergence or disappearance of a corresponding feature. "Disappearance" in this sense means a scenario where the value of the feature is relatively "undefined" or not consistent enough to be considered as a characterising feature for the series. The emergence of a feature does not necessarily mean that the feature has the same "value" across all the series highlighted in Table 4.3 as the opportunities available to potential offenders vary across different spatial space. Thus having the indicator "1" for lines (S/N) 1 and 2 for the "Day" attribute does not mean S1 and S2 at Mowbray always happen on the same day as they are two different series, but emerging at the same locality. Furthermore, the suspect frame (SFr) attribute emerges for both S1 and S2, but with unique values "moderate" and "slender" respectively as seen in Figures 4.11 and

Table 4.3: A depiction of the characterising (peculiar) features emerging for each series (S_i).

S/N	Location	PDE(%)	Day	Time	Vic	Sus	VAge	SAge	SFr	Mot	MO	HCol	Mask	Sub-Ab
1	Mowbray	35 (S1)	1	1	1	1	0	0	1	0	1	1	0	0
2		65 (S2)	1	1	1	1	0	1	1	1	1	1	1	1
3	CapeTown	50 (S1)	0	1	1	1	0	0	1	1	1	1	0	0
4	Central	50 (S2)	1	1	1	1	0	1	1	0	1	0	0	0
5	Wynberg	40(S1)	1	1	1	1	0	1	1	0	1	1	0	0
6		34(S2)	1	1	1	0	1	0	0	1	1	0	1	1
7		26(S3)	0	1	0	1	0	1	1	0	1	1	0	0
8	Grassy-	21(S1)	1	1	1	0	0	0	0	1	1	1	0	1
9	Park	79(S2)	1	1	1	1	0	1	1	1	1	0	1	0

4.12. Also note that the "motivation" feature emerges for $S2$ but did not emerge for $S1$. Hence, this feature has the indicators "0" and "1" for $S1$ and $S2$ respectively. Each of the series has at least six features characterising the offender's MO, which aligns with the initial proposition and similarity (threshold) condition for this research.

It is clear from Table 4.3 that the operating times for the series and the capture method (called MO in the data as this is the term police use for it) are features that are highly consistent throughout the identified series, which is as anticipated, while some other features such as motivation (Mot) and Victim age (VAge) are not very consistent across series clusters as observed in Figures 4.11 and 4.12. These varying observations agree with the fact that each series is likely different and has its unique MO, since the opportunities available to potential offenders vary across different spatial space due to differences in spatial factors. While the capture method emerges across all the series, it is also interesting to note that the corresponding feature values vary for different series. For example, the method for victim capture is through "kidnapping' in one instance and "substance-influence" in some other series. These varying features confirm the standard theories of environmental criminology, such as crime pattern theory and routine activity theory [18] (see Chapter 2). In summary, the essence of knowing and understanding these features and varying characteristics in the analysis is to achieve the target end, which is actionable knowledge

(e.g. suspect prioritisation) for crime deterrence in resource constrained settings.

In order to amplify the success rate (SR) of the algorithmic process and ensure connected series are identified with a high probability, we employ an adaptive graph-size (AGS) contraction operation with a Monte Carlo heuristic in the process of crime series detection. We validated the approach and heuristics by examining the output with a training sample dataset of 80 % and a test set of 20 %. Further validation was conducted to check the accuracy of the algorithm wherewith the appropriate partitions are known before hand, using a controlled dataset as discussed in Section 4.4.2. Figure 4.18 shows the outcome of the experiment.

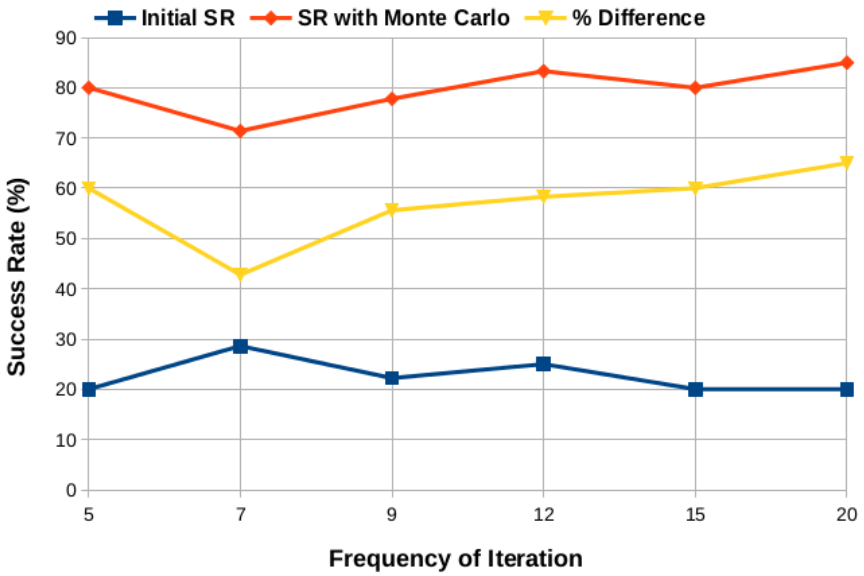


Figure 4.18: An assessment of the amplified success rate (SR) with Monte Carlo and adaptive graph size approaches

The experiment shows that the adaptive graph size approach together with Monte Carlo heuristics yields a much higher success rate (SR) for identifying clusters. Figure 4.18 presents the success rate (SR) of the initial Karger method for deriving the appropriate partition for cluster identification, and that of the improved approach (adaptive graph size) that borrows from the Karger Stein method using Monte Carlo heuristics, which is adopted in this research. The figure reveals the initial SR of the algorithm in identifying an appropriate partition, as well as percentage difference after SR became amplified.

Figure 4.19 further presents the reciprocal rank measures of the cluster identification process, given by the average

reciprocal rank (ARR) in Equation (4.1) where $rank_i$ is the position of the first relevant cluster partition identified in each iteration W. In practice, an ARR of 1 is perfect, while an ARR close to 0 is very bad. It is clear from Figure 4.19 that our ARR values are between 0.6 and 0.8, and thus have a reliable success rate.

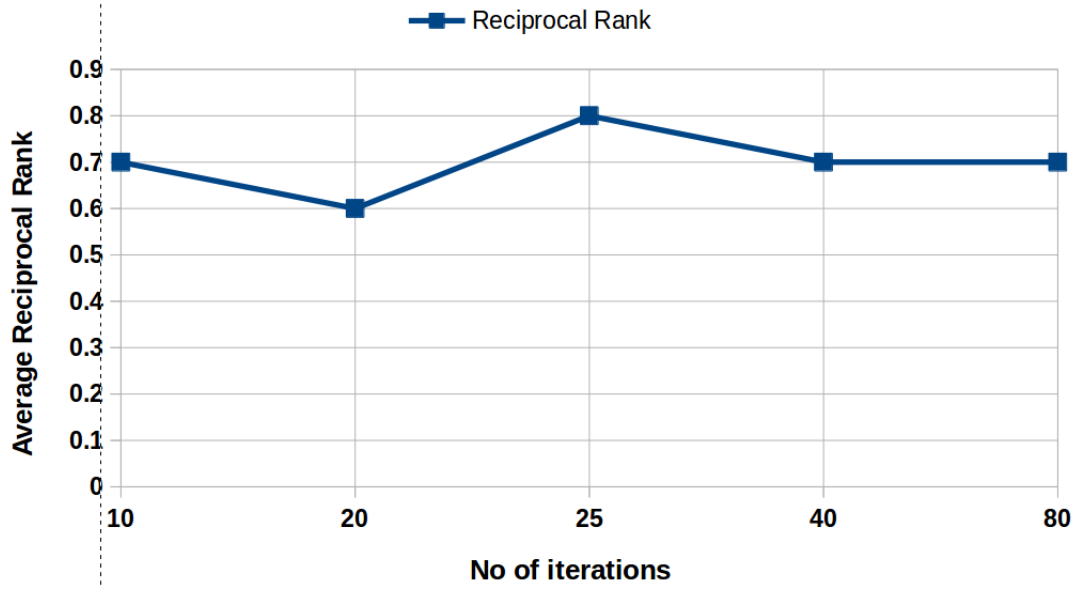


Figure 4.19: Average Reciprocal Rank on CriClust Model

$$ARR = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{rank_i} \tag{4.1}$$

While we also observe that the amplified success rate comes with some trade-off in terms of speed (time), we note that this is still within a reasonable limit especially because we prioritise identifying correct clusters with very high intra-class similarity and low inter-class similarity. Figure 4.20 shows the time cost trade-off of both the original karger-stein and improved algorithm employing the adaptive graph size (AGS) heuristics. The multi-threading mechanism adopted in the algorithmic process also assisted the reasonable time limit. As a proof of concept, one can infer from Figure 4.20 that the AGS approach has a reasonable timing, while serving the dual purpose of promoting algorithmic success as presented in Figure 4.18.

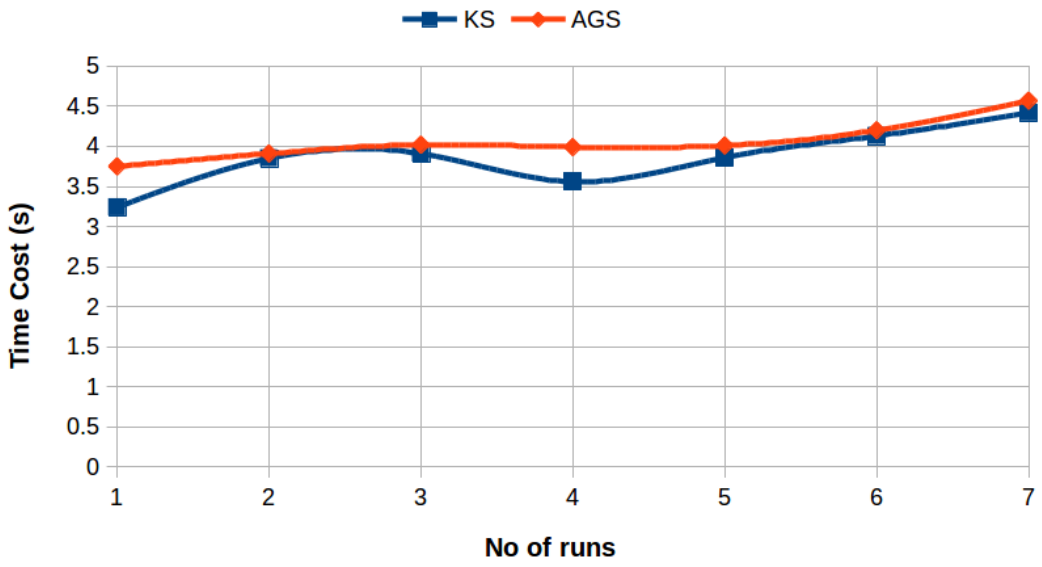


Figure 4.20: Time cost comparison of KS and AGS on training data

This research further evaluates the reliability of the model, in terms of the ground truth identification of clusters whether it satisfies the earlier stated hypothesis. That is whether crime series identified in a cluster have high intra-class similarity. To do this the precision of the algorithm in terms of level of coherence was checked on identified clusters. The common measures presented in this work assume a ground truth notion of relevancy, whereby a cluster is expected to meet the earlier established propositions in the research.

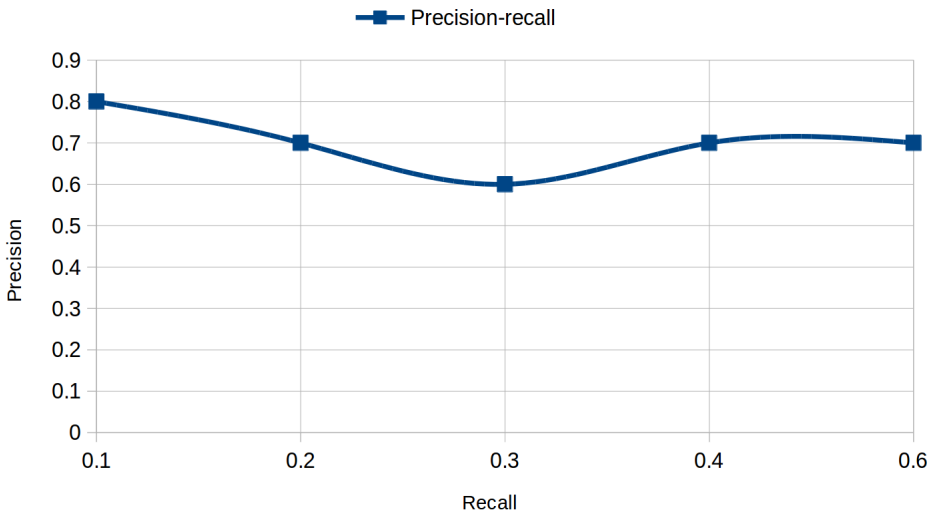


Figure 4.21: Precision and recall measure on CriClust model

4.4.1 Assessment of Peculiar Features in CriClust Patterns

We anticipate that the identified peculiar features for a series will increase as the the strongly indicative feature sets increases, which is dependent on the significance threshold because that forms the core basis of the similarity condition. Figure 4.22 shows how the pattern level precision of the series grows with varying $\mathcal{S}$ conditions. The number of discovered series for $\mathcal{S} = 6$ is 118, $\mathcal{S} = 7$ is 91, and for $\mathcal{S} = 8$ is 43. There were too few patterns with a higher value of $\mathcal{S}$, which could mean that certain series would be omitted at a very high threshold level. Moreover, with such high value there is no indication of any significant precision gain as seen in Figure 4.22. Hence a reasonable threshold that will assist the analysis is crucial. The threshold considered in this research is however recommended by crime experts.

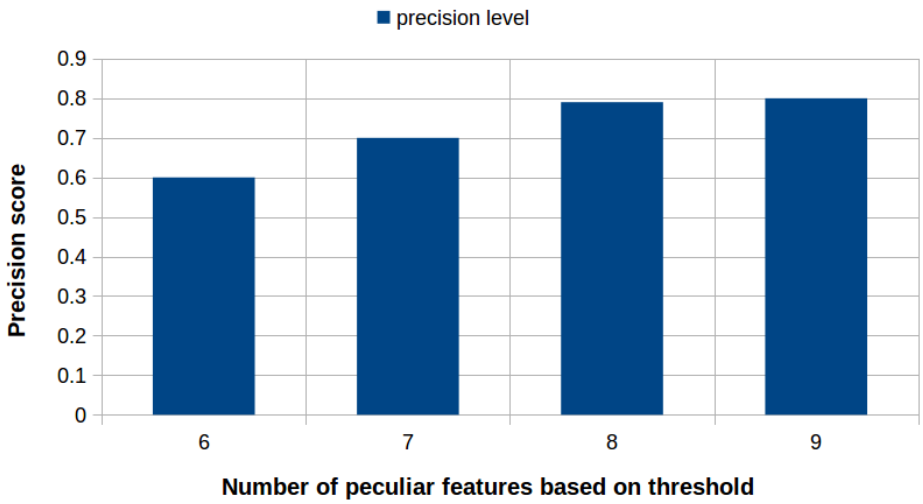


Figure 4.22: Precision score for peculiar features based on significance threshold

4.4.2 Performance of CriClust on Controlled Experimental Data

The reliability of the CriClust model is based on its ability to determine or identify all inherent crime series information. Thus as a further examination, we investigated the performance of CriClust on some controlled experimental data, in order to assess the reliability of the model in generating all inherent crime series as well as highly correlated clusters.

For the control experiment, some pre-defined hand-crafted (auto-generated) crime series were inserted into the

data and output was checked to see if they were identified as clusters. This is achieved by simulating a control set of data, whereby a number of potential series had been pre-defined in the data. In this case we generated data based on some a priori knowledge, being aware that a potential series will mostly occur in very close proximity (e.g. location and time), and with similar MO. We then derived the control experimental data by generating a pool of (unique) random clusters in which known offender characteristics are put into consideration. We assign a unique code, say k, (in form of a numeric value) to every location (l). Then we pre-define the average number of clusters (n) to be generated for each location. The set of possible values for other attributes, which is stored in an array, is then referenced based on this unique code, using the modulo operator on the unique code. Thus we use the formula $k_l \ \% \ n$ to reference each attribute index value while populating the control data set. We performed experiment by setting the value of n to 2, 3, and 4. We performed experiment by setting n to 2 at the first instance as shown in Figures 4.23 and 4.24.

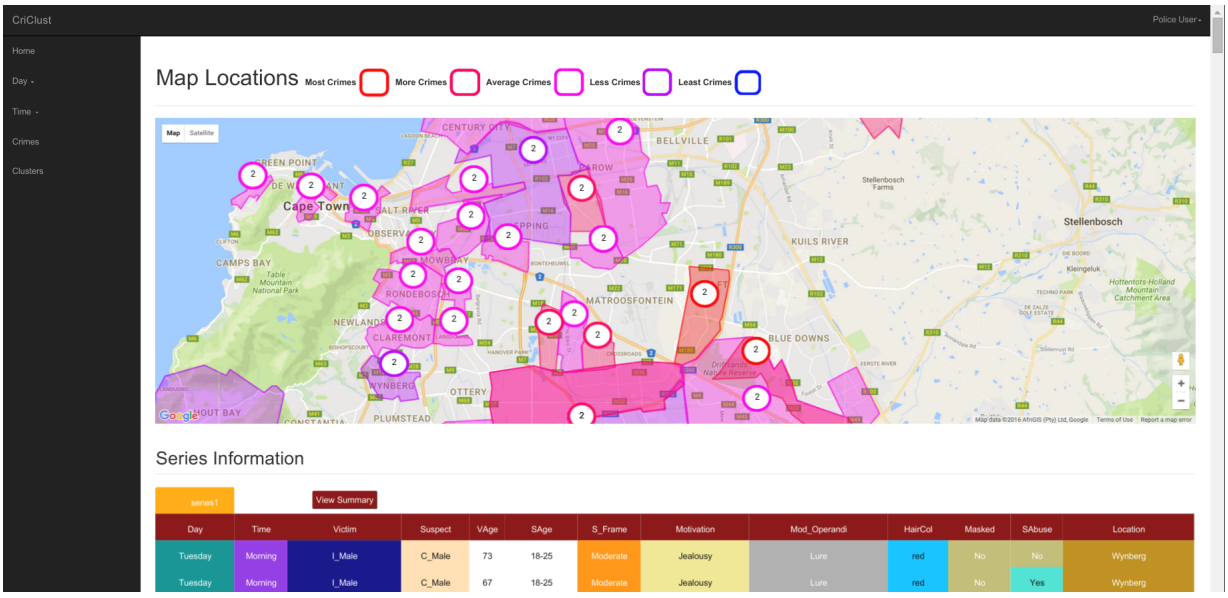


Figure 4.23: Node-value indicator for number of series identified per location

Furthermore, we repeated the same procedure for $n = 3$ and $n = 4$. This means that 3 and 4 clusters are expected per location respectively. Figure 4.25 gives the corresponding output for when $n = 3$. From qualitative point of view, it is clear from Figures 4.24 and 4.25 that the inherent series as well as the expected number of clusters per location were adequately identified in most of the locations. The generated clusters also show that CriClust is reliable in identifying crimes with similar attribute characteristics. It is important to note that the corresponding

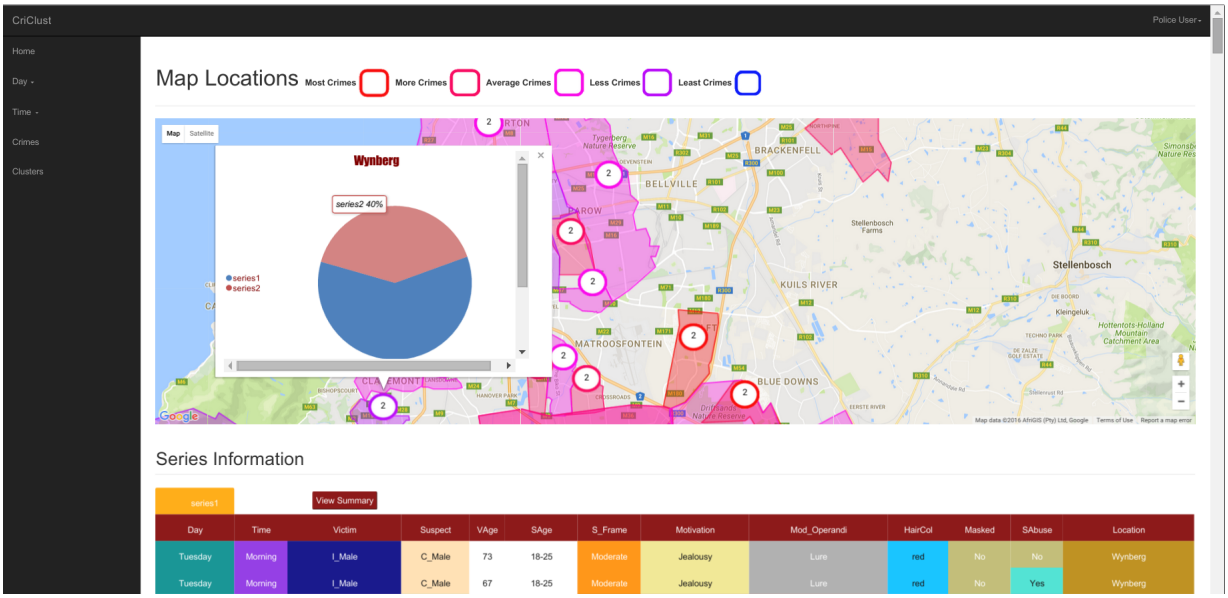


Figure 4.24: CriClust performance on controlled experimental crime dataset: The case of 2-series clusters (showing PDE) identified per location

PDE outputs were not known before-hand based on the fact that the crime data was completely auto-generated using the defined number of corresponding series expected per Location (with the modulo operator) as a yardstick.

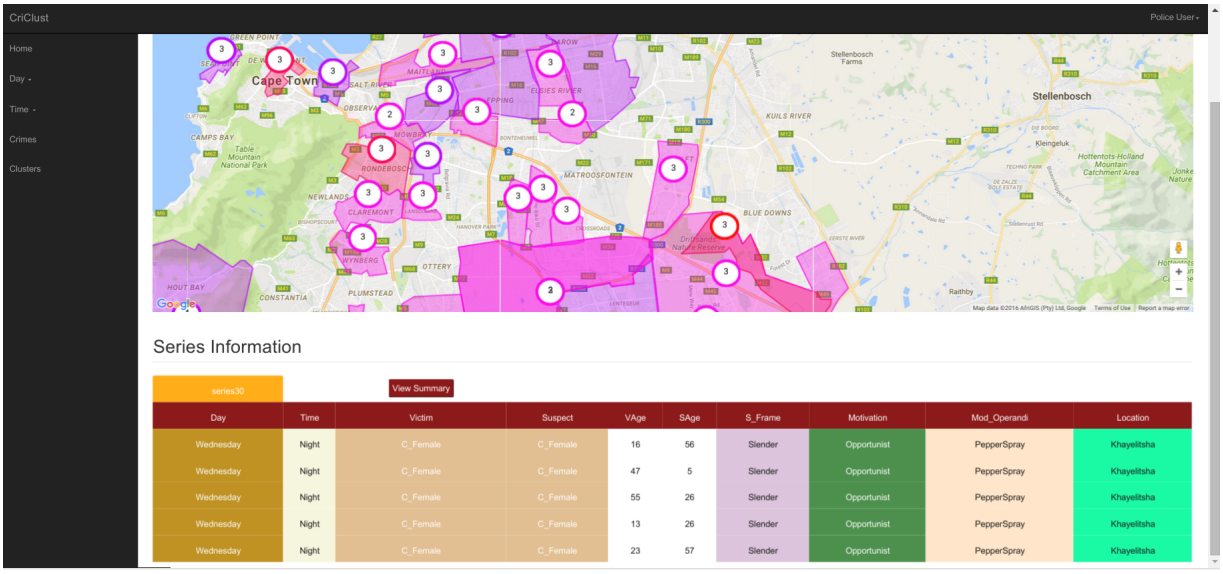


Figure 4.25: CriClust performance on controlled experimental crime dataset: The case of 3-series clusters identified per location

Table 4.4: Performance of CriClust on controlled experimental data.

S/N	No of Loc	Record Size	n-values	Expected No of Series	Observed Series	% Error	% Success
1	40	1000	2	2×40	2×39	2.5	97.5
2	40	1500	2	2×40	2×40	0.0	100
3	40	2000	2	2×40	2×37	7.5	92.5
4	40	1000	3	3×40	3×40	0.0	100
5	40	1500	3	3×40	3×40	0.0	100
6	40	2000	3	3×40	3×38	5.0	95
7	40	1000	4	4×40	4×40	0.0	100
8	40	1500	4	4×40	4×39	2.5	97.5
9	40	2000	4	4×40	4×40	0.0	100

To further quantify the performance of the model, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. We then compute the error percentage for the 40 locations examined as shown in Equation (4.2). Table 4.4 shows the performance of CriClust in generating all inherent series clusters on controlled experimental data. The resulting information on the experiment when the number of expected clusters n is 2, 3 and 4 is presented in Table 4.4. The number of "observed series" across locations is expected to correspond to the "expected series" at the location. Note that "loc" means location in Table 4.4 as there are 40 locations. According to the estimates in Table 4.4, the averages of % error ($\frac{17.5}{9}$) and % success ($\frac{812.5}{9}$) are 1.94 % and 98.06 % respectively. Table 4.4 suggest that CriClust is promising in identifying all inherent series in a dataset with minimal error.

$$\%\text{Error} = \frac{|\text{Expected-Value} - \text{Observed-Value}|}{\text{Expected-Value}} \tag{4.2}$$

4.5 CriClust Scalability Trend on Cluster Identification

Furthermore, we conducted experiments to analyse the scalability of the cluster identification process. Scalability can be measured from two different perspectives: (i) data size; and (ii) hardware capacity with respect to number of processors. However, this research only considered scalability from the data size perspective. Figure 4.26 reveal the runtime performance of CriClust as dataset size increases. These times are averages over multiple runs against each dataset. We note that run-time increases approximately linearly as the data size increases, which is typical of most data dependent applications. The time to deploy indicates the time it takes for the application to establish connection with the database and to be ready for cluster processing, while the runtime indicates the actual time it takes to process the clusters. It is important to note that applications such as CriClust could achieve even better runtime performance when deployed on a very powerful (high-capacity) machine, which is a typical consideration in parallel processing mechanism and several big data applications. This means that the performance gain in terms of runtime is relatively dependent on the capacity of the machine on which it is deployed, among others.

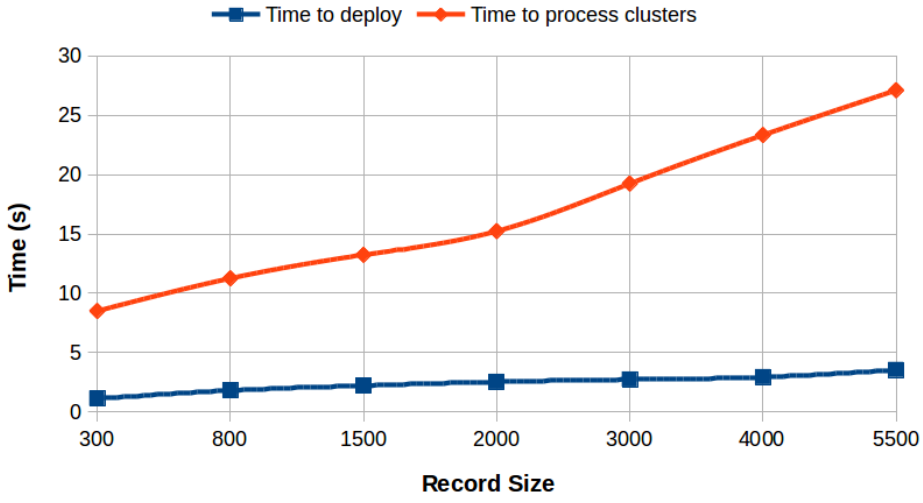


Figure 4.26: Scalability assessment of CriClust with increasing data size

4.5.1 CriClust Analysis Class Model

The class diagram showing the interaction of objects in the CriClust model is presented in Figure 4.27. A multi-threading mechanism that implements the ForkJoinPool framework was used to achieve maximum efficiency,

where the adaptive graph size method and Monte-Carlo heuristics were adopted to amplify the success rate of the algorithm in determining sufficiently connected edges and the right partitions. The proposed CriClust model for series identification has seven integrated classes, which include CrimeCollection, CrimeNode, Graph, CutGraph, DistanceMatrix, HCSGenerator and KargerStein.

The CrimeCollection class contains the methods for establishing conections with the database and retrieving data from the database for storage and cluster processing. The DistanceMatrix is responsible for creating the 2d ArrayList distances for the crime nodes. The ArrayList stores the edge connections between the crime nodes. The values in the ArrayList is one of the following: $\{-1, 0, 1\}$. The value -1 means there is no edge between two nodes, value 0 always holds for the same node, and value 1 means there is an edge between the nodes. The decision for linking edges in the graph is based on the similarity conditions established. The distance data structure is achieved using the significance and prevalence information. Each node in the graph is an instance of the CrimeNode class, which contains parameters that give information about the crime. There is also a MasterNode in the form of an arrayList, which contains an instance of the CrimeNode itself and all other nodes merges into it. The Graph class consist of instance of CrimeNode in an ArrayList, while the CutGraph consists of CrimeNodes instances that have been partitioned (cut) from the main graph. This class also contains a mincut variable that indicates the number of edges that have been detached to obtain the subgraph, since the idea is to identify the minimum amongst these numbers. The highly connected subgraph generator (HCSGenerator) is the main class that implements the clustering algorithm, which takes the graph constructed from the DistanceMatrix class. The KargerStein class which implements a recursive task is called by the HCSGenerator. The recursive task is useful for identifying the smallest minimum cut (ideal partition) when the Monte Carlo approach is invoked during the adaptive graph size process.

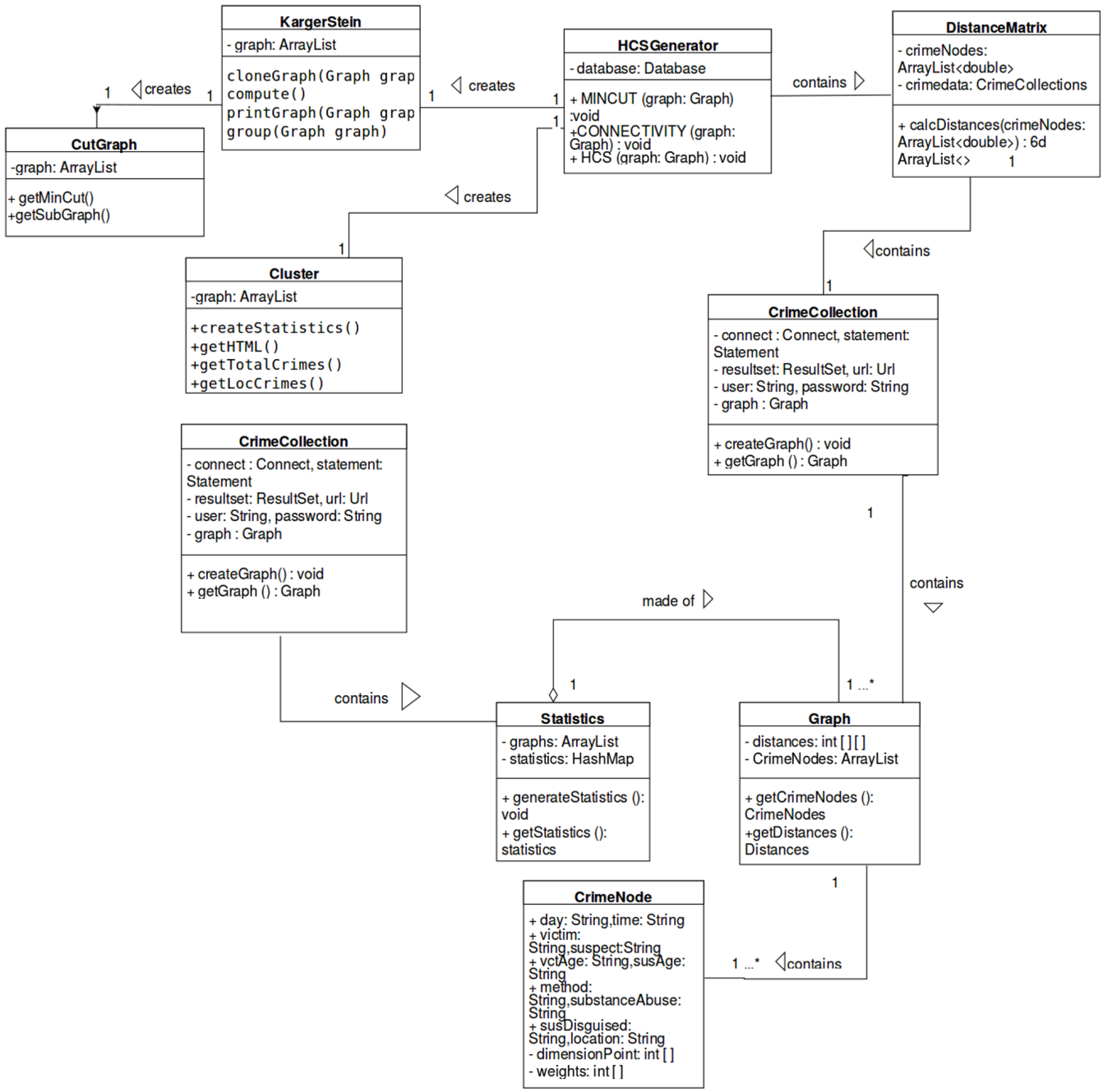


Figure 4.27: Analysis class model showing interaction between objects in CriClust

4.6 Systematic Comparison of other Series Detection Models with Our Proposed Approach

Considering a further investigation on similarities and differences between our proposed CriClust model with other existing models for series (or series-related) identification, Table 4.5 presents a number of criteria that were explored. This comparison also shows how our proposed approach contributes to the body of knowledge on crime series identification for improved public safety outcomes.

We compare existing related series detection approaches with our proposed method of crime series detection using five main features, which include: exploratory basis, domain (nature or category of crime analysed), modelling approach, techniques used, and finally empirical observations and results presented for knowledge support. While most of the efforts in Table 4.5 are research in series detection, it is important to note that not all actually present discovered patterns in a manner that will assist public safety, particular a person with no domain knowledge. CriClust uniquely present its results to sensitise security agencies and guide on what series cluster to examine and track down first, by using the PDE and PSE concept. Furthermore, to the best of the researcher's knowledge the dual threshold mechanism and adaptive graph size with Monte Carlo heuristics considered in the CriClust model have never been considered for series identification.

4.7 Benchmarking: Baseline Comparison with Common Clustering Techniques

As a baseline for performance and relevance examination, we compare our proposed algorithm with some baseline models and algorithms. The problem addressed in this research is essentially a clustering problem, although the similarity between clusters is somewhat supervised based on some a priori knowledge on what constitutes a potential crime series. Three different clustering techniques were selected to investigate and study for comparison, these are: (i) K-means; (ii) hierarchical clustering (HAC); and (iii) Nearest neighbour. These techniques were selected based on the following major criteria, among others:

Table 4.5: Comparative evaluation of existing crime series detection systems with CriClust System

S/N	Features	Crime Linkage [85]	Mining Rotten Core [107]	Serial Crime Pattern [20]	Crime Linkage[119]	CriClust Model (our work)
1	Exploratory Basis	Crime linkage	Crime series detection	Serial criminal patterns detection	Crime linkage	Crime series detection
2	Nature of Crime Explored	Breaking & Entering crimes	Burglary (housebreaking)	Armed Robberies	Burglary crimes	Sexual crime (rape)
3	Modelling Approach	Statistical approach	Conventional optimisation	Neural Network (NN)	Bayesian Network	Dual threshold scheme & graphical model
4	Techniques Used	Bayes factor, Hierarchical clustering	Integer linear programming, clustering, BFS	Cascaded network of Kohohen NN	Bayes Network (BN)	Geometric-projection, HCS clustering & Monte-Carlo
5	Emperical Observation	Posterior odds, Bayes factor & number of clusters	Map locations of series, pattern space, precision & recall	Percentage of predicted & actual patterns	Poterior probabilities & BN	Map (PDE, PSE) of potential series, scalability, precision & recall

- applicability
- popularity
- flexibility
- relevance to current study
- potential for scalability

The HAC seeks to build a hierarchy of clusters by starting with each crime as a singleton cluster. At each iterative step, the most similar clusters (based on the defined similarity condition) are then greedily merged to form a new cluster. This reduces the number of singleton clusters at each iterative level and imposes a hierarchical structure on the dataset. The metric typically used in HAC is the distance between pairs of observations (crimes) and a linkage condition that defines the dissimilarity of observations as a function of the pairwise distances of crimes in the database. K-means on the other hand simply partitions a dataset into k clusters in which each instance of the dataset belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The iterative nearest neighbour (NN) uses a nearst neighbour metric for selecting and grouping objects. The Waikato Environment for Knowledge Analysis (WEKA [2]) workbench for machine learning was used to compare these algorithms, using a supervised attribute instance. We note that most existing clustering techniques could be considered, however, there is need for relevant concept description in order to achieve meaningful results as regards the target task in this research. It is clear that clusters identified would not be a good representation of the expected partition as most of the algorithms have varying means of processing clusters. Crime series identification is a unique problem that is different from a universal clustering problem, which most of the compared algorithms are known for.

while K-means, NN and CriClust are able to cope with datasets consisting of at least 2000 records, Hierarchical clustering method became intractable at this level. We also note that existing clustering techniques naturally fail to identify clusters of interest (series cluster) in the dataset, but perhaps could perform better if parameters are appropriately supervised in the algorithm. Hence the importance of supervising the similarity conditions to meet the target goal. Tables 4.6 and 4.7 show sample clusters generated by both K-Means and Hierarchical clustering respectively. It is clear from the resulting clusters in these tables that it is hard to highlight correlated clusters that satisfy the objective of this research (potential series identification). An important observation emanating from Tables 4.6 and 4.7 is that the generated clusters fail to identify potential series and did not take the space-time attribute into consideration in its analysis. On the other hand, CriClust generated clusters with an important consideration for the space-time information as earlier explained in Section 3. We however recognise that these baselines have the potential to perform better if their parameters are appropriately tuned and channelled towards the goal of the analysis. While K-means is a fast algorithm, the major limitation in the K-means algorithm for

[2]http://www.cs.waikato.ac.nz/ml/weka/

this kind of task is the fact that it is an unsupervised technique and the number of clusters needs to be specified at the onset of the analysis. Figure 4.28 shows the general scalability of the models compared in terms of record size and processing time. We note that K-means is generally linear as opposed to Hierarchical with much higher times. CriClust performed better for crime series detection than these methods.

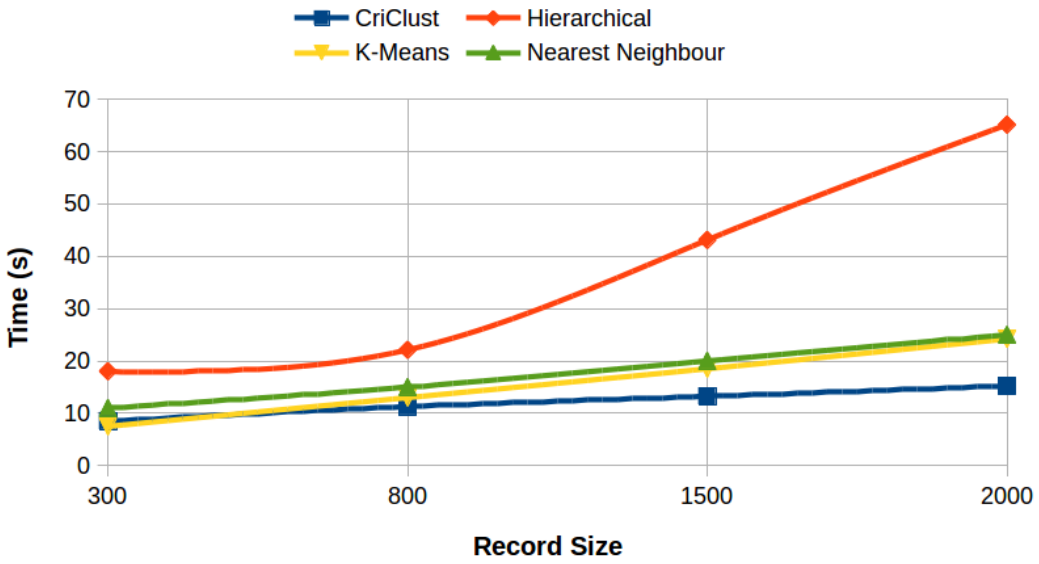


Figure 4.28: Time cost comparison for common techniques compared

Day	Time	Location	Victim	Suspect	V-Age	S-Age	S-Frame	S-Height	Motivation	MO	Mask	S-Abuse
Fri	Evening	Strand	I-male	I-male	82	33-40	Moderate	Medium	Jealousy	Kidnapped	Unknown	Yes
Wed	Night	Atlantis	I-male	I-Female	59	Above-45	Slender	Short	Self-Satisfaction	PepperSpray	Unknown	No
Thu	Evening	Nyanga	I-Female	C-Female	21	Above-45	Slender	Medium	Jealousy	Gun	Yes	No
Wed	Morning	Nyanga	W-Female	C-Male	63	18-25	Slender	Medium	Revenge	Vehicle	Unknown	No
Sat	Evening	Mowbray	B-Female	W-Male	65	33-40	Moderate	Short	Revenge	Kidnapped	Unknown	No
Thu	Evening	Elsies_River	B-Female	C-Female	85	less-18	Slender	Short	Self-Satisfaction	PepperSpray	Unknown	Yes
Fri	Night	Khayelitsha	W-Male	B-Female	81	less-18	Slender	Medium	Jealousy	Deceit	Unknown	Yes
Sun	Noon	Mfuleni	I-male	I-Female	20	33-40	Slender	Short	Jealousy	Lure	Unknown	No
Thu	Evening	Atlantis	W-Male	W-Male	29	Above-45	Slender	Short	Revenge	Kidnapped	Unknown	No
Fri	Noon	Goodwood	I-male	C-Female	32	18-25	Moderate	Medium	Self-Satisfaction	SubstanceAbuse	Unknown	No
Wed	Noon	Rondebosch	I-Female	C-Female	61	26-33	Slender	Short	Jealousy	Lure	Unknown	No

Table 4.6: Sample Cluster: K-Means

Day	Time	Location	Victim	Suspect	V-Age	S-Age	S-Frame	S-Height	Motivation	MO	Mask	S-Abuse
Thursday	Night	Langa	C_Female	C_Male	29	less-18	Slender	Medium	Opportunist	knife	Unknown	Yes
Tuesday	Morning	CapeTown_Central	I_Female	B_Male	55	33-40	Moderate	Short	Self-Satisfaction	Deceit	No	Unsure
Friday	Night	Dieprivier	W-Female	B_Female	10	18-25	Fat	Tall	Revenge	Deceit	Unknown	No
Wednesday	Noon	Khayelitsha	B_Male	I_Male	12	33-40	Fat	Tall	Jealousy	Lure	Yes	No
Friday	Noon	Phillippi_East	C_Male	W-Female	81	33-40	Slender	Medium	Jealousy	Substance	No	Unsure
Tuesday	Noon	Lingelethu_West	B_Male	B_Female	76	33-40	Fat	Short	Jealousy	Substance	Yes	Unsure
Tuesday	Morning	CapeTown_Central	W-Female	B_Male	22	Above-45	Fat	Medium	Jealousy	Kidnap	Yes	Yes
Sunday	Noon	Langa	C_Male	B_Female	36	Above-45	Fat	Tall	Revenge	Gun	Yes	Yes
Thursday	Morning	Kraaifontein	I_Female	C_Male	21	18-25	Slender	Medium	Revenge	Lure	No	Unsure

Table 4.7: Sample cluster: Hierarchical clustering approach.

4.8 Chapter Summary

This chapter presented quantitative and qualitative experimental results, both with quasi-real and "controlled" experimental data. The results revealed the potential usefulness of CriClust as a means to complement the desired safety goals and objectives in developing nations aimed at promoting smart city development. Map locations of potential series identified together with the proportion difference (PDE) derived for those series, as well as pattern space information (PSE) for peculiar features identification was presented. The peculiar features are important for public safety to understand and focus on driving factors for specific series at a particular location, while proportion difference identifies the series with a high dominating power to track down first. As we have shown through series cluster examples, CriClust provides richer insights into specific attributes and MO of each series than existing crime systems with only background level information. Furthermore, a high level feature comparison with other research on series detection was presented, followed by comparison with common (baseline) clustering techniques: KMeans, NN and Hierarchical clustering (HAC). Experimental results reveal that these techniques are generally not able to identify (potential) series as CriClust can. These results demonstrate the reliability of CriClust in assisting public safety agencies to achieve crime deterrence targets with a specific focus on potential series identification.

It is worth noting that existing crime data available in open repositories such as UCL repository [3] was not considered in this research due to the fact that the data lacks detail attribute information about victims and suspects that is considered in this research. Also, other repositories such as Crime Hub [4], which is managed by the Institute for Security Studies, only provide general background information that will not help our analysis. While we made many repeated efforts to collect real crime data from the police, this has not been successful despite the promises we received. The bureaucracy involved in the data collection process, as well as continued delay in releasing crime data was responsible for this. However, it is important to recognise that part of the originality of our ideas includes relevant concept description, illustrations and the semantics one can associate with them. Moreover, the quasi real data were generated based on crime expert recommendations. The results of the empirical observations show that CriClust is promising in promoting smart public safety. The next chapter presents a reflection on how

[3] https://archive.ics.uci.edu/ml/datasets.html

[4] https://issafrica.org/crimehub/stats/

the empirical findings in this research addressed the research questions and then concludes the thesis.

Chapter 5

CONCLUSION, RECOMMENDATION AND FUTURE WORK

The motivation for this research is the incessant challenge to tackle crime faced by public safety agencies, particularly in resource constrained settings such as in developing nations, which is an impediment to realising smart city development targets. The proposition of this research was that identifying crime series patterns (as defined by the international association of crime analysts (IACA) [78]) offers a smart way to tackle crime. This is because many crimes are committed by repeat offenders and most crime patterns exhibit at least a k minimum principal set that characterise the modus operandi (MO) of the offenders' behaviour, which can assist in the identification of crime series patterns. Furthermore, past literature suggests that there is a paucity of research work on the area of series identification from developing countries. To address this proposition, two research questions were considered:

- *Firstly, how can crime mining and machine learning support and promote the identification of crime series patterns (CSP) to provide actionable insight from crime data, and what heuristics can be used to augment such analysis?*
- *Secondly, what level of confidence can be expressed in such a model and how does the technique scale-up*

with increasing numbers of crime records?

This chapter summarises and concludes this research. The discussion in this chapter reports how the empirical analysis and findings addressed the research questions. Contribution of the research to the notion of smart city development in developing nations follows. Thereafter, suggested recommendation and opportunities for future research is documented.

5.1 Research Summary and Conclusion

Data mining is the process of uncovering useful, understandable patterns in data using machine learning, statistics, artificial intelligence and database systems. Crime data mining reveals previously unknown useful and understandable information about crime events. Thus, it helps to channel public safety resources more effectively (optimally) across geographical locations. The importance of crime mining in achieving sustainable crime reduction and control in a smart city development cannot be overemphasised. This is particularly crucial in resource-constrained settings, such as in developing nations, where police are short-staffed and the available resources have proven to be insufficient for deterring crime. This further necessitates the need for effective intervention strategy to promote optimal allocation of public safety resources in such settings. Identifying serial (repeat) offenders, referred to as crime series, is considered one of the smart ways of achieving crime control targets in developing nations. While the identification of repeat offenders patterns is crucial for promoting actionable solution in resource constrained settings, our findings reveal that there is currently no specific tool used by the police to derive such patterns. Moreover, the effort in such area has largely remained a manual process, which is tedious, resource-consuming and error-prone.

While this research considered a rape database as an application scenario for identifying crime series, the idea considered in this research can be extended to other forms of crime such as fraud detection and burglary, to mention a few. Furthermore, the model has the capability to extend beyond crime mining to other domains that tend to exhibit specific related features that could capture related events. The motivation for considering a rape

database is the fact that despite the increased sensitivity and understanding about sexual assault and violence [1], South Africa as a developing nation happens to be a place where sexual offence is common [2] [3].

5.2 Synthesis of Empirical Findings

5.2.1 How can crime mining and machine learning support and promote the identification of crime series patterns (CSP) to provide actionable insight from crime data, and what heuristics can be used to augment such analysis?

Crime series identification is investigated as there is evidence (from repeat victimisation analysis) that a high percentage of crimes happening at specified locations are committed by repeat offenders [57, 30]. However, current literature suggests that despite the critical importance of crime series identification, there is relatively little research exploration on the topic, particularly in developing nations, when compared to other spatio-temporal analysis, such as hotspots [107].

In this work, we propose a promising approach that effectively captures inherent crime series patterns, while particularly maintaining desirable statistical properties. To identify crime series in a (rape) crime database, a hybrid model called CriClust, which combines similarity concepts and graph connectivity (highly connected subgraphs) was adopted. CriClust is augmented with a dual threshold scheme (significance and prevalence thresholds), and considers established theoretical concepts in order to conceptualise the research framework. A dual threshold scheme is used to identify mutual information between similar crime objects, as well as preserve the pattern space localities. The similarity data is derived based on some similarity condition. The representation of the crime similarity data in a similarity graph helps to simplify computation, since we then only need to consider sufficiently connected sets of nodes, which is referred to as highly connected sub-graphs (HCS). In this research, spatial and temporal data is represented so as to facilitate difference computations by using geometric projection. The similar objects are then modelled into a graph structure, which is then partitioned into highly connected sub-graphs of

[1] http://rapecrisis.org.za/

[2] http://www.iol.co.za/news/crime-courts/raped-teen-left-for-dead-under-rocks-1884760

[3] http://www.news24.com/SouthAfrica/News/Girls-abducted-from-home-raped-20150410

related crime. The graph model adopts HCS, which was originally proposed by Hartuv et al.[37] and has found its application in many domains that require identifying such similar clusters using a graphical approach. Monte Carlo approach and adaptive graph-size contraction heuristics are employed to amplify the algorithm success. The coherence and high intra-class similarity in the clusters obtained illustrate the suitability of the CriClust model in identifying crime series and thus promoting pro-active crime control strategies. Hence we conclude that clustering of crime-similarity graphs using the HCS algorithm can support crime series detection if CriClust's similarity metric and adaptive-graph size heuristics are used.

5.2.2 What level of confidence can be expressed in such a model and how does the technique scale-up with increasing numbers of crime records?

Our approach uses a crime similarity function to connect crime instances that share related attribute information, based on the dual threshold scheme. The similar objects are then modelled into a graph structure which is then partitioned into HCS of related crime. In addition, two new interest measures were considered to further augment the analysis: (i) Proportion Difference Evaluation (PDE), which measures the propagation effect of a series; and (ii) Pattern Space Enumeration (PSE), which reveals underlying strong correlations and defining features for a series, in order to quantify the strength of the prevalence localities. Heat maps are then used to enhance the visualisation of pattern specific locations where respective series activities are prevalent as shown in Figures 4.7, 4.8 and 4.9. The PDE helps to identify the dominant series at locations where there is more than one series. Furthermore, the PSE assists in identifying the MO of perpetrators, by displaying the raw data rows of any selected crime series, as presented in Figures 4.11, 4.12, 4.13, 4.16, and the characterising (peculiar) feature emerging for different series identified. The PSE is critical for actionable information in promoting citizen-centred safety.

The research produce much confidence with the generated patterns (in both quantitative and qualitative manner), which was substantiated with the optimistic reaction and input we received from experts in this domain. Diagrammatic representation of concepts and ideas are used to explain how pieces of crime information connect or link together in a descriptive manner. Furthermore, the generated patterns agrees with the initial concept description (hypothesis) and threshold constraints earlier established in this research, wherein at least up to 6

defining attributes must be present in the identified series. Furthermore, the research use a training sample,for which the correct cluster partition had been pre-determined, to evaluate the success rate of the adaptive graph size (AGS) approach considered. Figures 4.18 and 4.20 shows that the adopted AGS is promising in generating the correct clusters with a reasonable time cost trade-off. Moreover, the level of coherence in the generated cluster based on the qualitative visualisation presented (for example, see Figure 4.17), as well as the validation of the derived series patterns across locations as presented in Figure4.21 reveals that the required cluster partition is returned with a confidence level of 95%. These findings reveal that the proposed CriClust model is promising. Furthermore, the scalability assessment in terms of increasing volume of crime as shown in Figure 4.26 reveals that CriClust performance in realistic high volume of crime is promising.

The usefulness of this research in terms of clusters generated and relative scalability can only be appreciated if one considers the effort crime analysts usually have to go through if they were to identify crime series clusters in even hundreds of record using Excel, a common tool currently in most SA police stations. This would otherwise be tedious, error-prone and time-consuming if not assisted with effective models such as CriClust. This research has demonstrated the usefulness of CriClust and inspired confidence on the generated clusters as one could see that generated clusters are highly correlated and agree with the initial threshold conditions set out as constraints in the study. While we note that common clustering algorithms are useful, they have failed to capture crime series patterns and the prevalence information that is useful for knowledge support in the specific scenario of crime series detection. This is achieved in the CriClust model by using the dual threshold scheme.

It is recognised that crimes recorded based on simple location proximity and counting frequency of similar crimes in hotspots will not suffice in assisting crime intelligence to deter crime [16, 41]. In summary the following are the key benefits of the CriClust system:

- Timely series pattern discovery: security agencies can stop a crime if they timeously identify the pattern of such crime, leveraging these to inform and influence actionable safety goals or targets.
- Quality of series generated: the relevance of the generated series to user's (public safety) information need is clearly seen. The coherence in the patterns generated and correlation with initial threshold conditions set out for the analysis are evident. This also proves the correctness of our approach.

- Statistically interpretable patterns and visualisation: CriClust pays special attention to systematically presenting series information such that a novice (public safety personnel) in the crime mining field can easily understand what the trend is saying. This is achieved using the Google map application programming interface (GMAPI), which helps to enhance visualisation of locations where series activities are prevalent. Furthermore, the notion of the PDE and PSE information which reveals the propagation effect (dominance) of a series and characterising feature for a series, aid actionable knowledge support. Part of the novelty of this research includes the various description, formalism, illustrations and the semantics one can associate with them.

To the best of the researcher's knowledge, besides the fact that there is paucity of research in crime series identification in developing nations, there is also no significant research effort in presenting series information in terms of the proportion difference (PDE) and pattern space enumeration (PSE) as presented in this research, with visualisation enhancement. Furthermore, the heuristics (dual threshold mechanism, geometric projection and AGS) and use of HCS are unique to CriClust.

5.2.3 Contribution to Smart City Development in Developing Nations

Smart city development is an emerging phenomenon that is driving much information and communication technology (ICT) research in recent times. This phenomenon is also currently a major focus in most developing nations of the world, and has varying interpretations by different researchers [33, 35]. While smart city generally focuses on transforming existing cities into better and more intelligent ones, its development is specifically concerned with two major objectives, which are: (i) increase or promote the quality of life of people; and (ii) improve the quality and efficiency of the services rendered by government entities and decision makers.

Security and effective policing has been noted as one of the best means, amongst others, in which data analytics could have a great effect on "smart cities" [29]. In developing nations where there is limited available resources to security agencies, coupled with little or no capital outlay to acquire armed weapons and related materials aggravates the challenge of crime. Moreover, existing software (e.g. Analyst's Notebook) that could reveal pattern in crime data are very expensive to purchase and requires critical training or a domain expert, which

poses serious constraints on developing nations. Our findings indicate that such tool is only available at the headquarters/provincial level in SA. Thus, local stations only make use of basic Excel software for filtering data and identifying patterns, which is cumbersome and time consuming. In some situations, local stations transfer crime data accumulated over a period of time to the provincial authority for analysis, since there are few experts and that is where the more advanced tool is available. This is a great limitation to effective policing because if at local levels, police are able to derive patterns in a timeous manner, then they can act to stop such patterns. However, in situations where they will have to wait a couple of days or weeks to get the analysed pattern from provincial level, crime could have worsened during the waiting period. Thus, in identifying crime series CriClust is a promising model for revealing pattern information in crime data which, unlike the high level summary statistics provided by official reports, is useful for public safety strategic and tactical planning.

Lastly, we note that patrols randomly carried out by the police are less effective [4], if not well articulated. Identifying series pattern is of great importance since repeat victimisation study has proven that many crimes that happen are perpetuated by the same offender (called serial offender)[57, 30]. CriClust is presented as a means to complement safety goals in a smart city development and achieve crime deterrence targets through place-oriented preventive patrol, among others. This is one of the smart ways to deal with crime and achieve crime deterrence targets in a resource constrained setting such as in developing nations.

5.3 Recommendation

Crime is currently a global problem that needs to be curbed. Crime experts need to work closely with crime analysts in order to find an effective solution to this global problem [68]. The need to adopt effective crime mining solution as a useful way of achieving sustainable crime deterrence in developing nations cannot be overestimated. While we have not presented smart data analysis or the CriClust system as a panacea, the solution presented in this research is more than a case study and is applicable to other crime domains. This can help in pro-actively improving public safety, particularly in resource-constrained settings such as in developing nations.

Crime information archived by public safety domains can be effectively explored, using data mining approaches, to

[4] http://renegadenoble.com/weblog/evaluating-the-effectiveness-of-random-preventive-patrol/

gain insight into crime committing trends or patterns [78]. Our research seeks to unlock transformational public safety value within complex crime data environment. From an operational and tactical response perspective, the solution presented in this research promise to be the beginning of a smarter approach to achieving crime reduction and deterrence targets in resource constraint environments. CriClust can assist analysts in suspect prioritisation, predicting and responding to patterns that anticipate crime before it happens. This will consequently help to tackle under-performance in certain core responsibilities of the police and help to develop evidence-based policies. In conclusion, crime data mining shows promise in helping to fight crime, however the availability and accessibility of accurate and up-to-date crime information is important if we are to address the crime problem in developing nations more effectively.

5.4 Limitations of Research

In this study, the focus was on identifying smarter ways of dealing with crime particularly in developing nations. Based on the recommendation of the international association of crime analysts [78], crime series pattern identification was investigated as a smart means of assisting crime intelligence and public safety agencies in resource constrained environments. By resource constrained environment, we mean situations in developing nations, Africa, where police are short-staffed and have limited resources in fighting crime. Therefore, the context, concept description and attribute selection were based on recommendations from experts in a developing nation, South Africa. However there is a great possibility to extend or generalise this notion to suit any (developing) nation or emerging scenario.

While it is a known fact that the availability and accessibility of accurate and up-to-date crime information is important if we are to address the crime problem (in developing nations) more effectively, lack of data access still remain a universal challenge. There exist many repositories of data such as the crimeHub [5], however the level of information provided on this platform cannot help the kind of analysis focused in this research. Thus, this universal problem of data accessibility did not leave this research as an exception. This means that this study could not successfully access and use real (rape) crime data for analysis, however, the quasi-real data used

[5] https://www.issafrica.org/crimehub/

for the experiment were: (i) guided by inputs and recommendations from crime experts; and (ii) confirmed as representative of what the police keep in realistic scenarios of crime. This is because the efforts made in accessing real data could not be finalised in good time and is still not finalised as at yet, to suit the timeline of this research. Furthermore, existing repositories (such as crimeHub [6]) were not very helpful since they only provide high level information of crime statistics as opposed to the raw information required for the analysis considered in this research. It is important to recognise that beyond the high level (map) information, and statistics provided by most crime information repositories, which may not be directly actionable, there is a need to identify raw patterns of crime and specific attributes thereof in order to achieve crime reduction. However, it is imperative to note that the context, initial concept description and attributes selection in this research were advised by crime experts, and as such the quasi-real data considered in this research is representative and sufficiently close to what the police archive as real data.

Finally, the research could not evaluate the system on a longitudinal basis. This means that the research did not provide a long-term impact of the use of the system on the eventual statistics or emerging trend of crime information in South Africa as a developing nation. This was not evaluated because crime intelligence currently have a means of analysing crime information and it would have been difficult to determine whether the use of CriClust system is what directly influenced current or emerging crime trends or reports. Furthermore, deploying such a system would involve quite a time-dependent process, due to bureaucratic constraints. Nonetheless, considering more time and resources, such a long-term research-impact measure is feasible and has been considered as part of future research.

5.5 Opportunities for Future Research

5.5.1 Possible Expansion of the CriClust System

The system developed in this research is a proof-of-concept solution that focus on the identification of crime series pattern (in a rape crime data set) for knowledge support. There is great potential in extending the current research

[6] https://www.issafrica.org/crimehub/

for more general pattern derivation and future adoption. Future research could consider combining mining of text and visual information, following a more extensive consideration for promoting effective investigative solution. For instance, attributes relating to suspect information (e.g tattoo, masked) could perhaps be translated into visual information (identikits) to mine suspect information and gain better insight into the crime data. Furthermore, other tactical means of integrating crime data sources for tactical analysis (such as hot-product, hot-place, to mention a few, as suggested by IACA) could be considered as an extension in the CriClust system.

5.5.2 Consideration for Streaming Data

Considering today's information age, mining of crime data streams -such as footage from video cameras in public places, text messages from victims, etc- needs some considerable attention. This could be supported in the form of cascaded mechanism for knowledge discovery and decision support. The major challenges include limited resources of memory and time, data complexity and arrival rate. Most data mining operations have conventionally been carried out over static datasets, where data is assumed to be resident in memory and mining algorithms can afford to access the input data several times. Streaming data requires the ability to make sure that data is continuously processed in real time without any visible time lag or information loss.

While no technique is universally applicable to a crime clustering or data mining solution, current trends indicate that enormous challenges remain open and need to be addressed in this domain of interest. These include effectively managing complex data formats (in its variety) during analysis, and developing scalable programming models for strategic or tactical crime analysis. Addressing these challenges necessitates a paradigm shift in the mind of crime analysts and data mining researchers. Hence, research should continue in this domain of interest.

5.5.3 Consideration for Real Crime Data

This research was only able to use a quasi-real dataset that was generated based on expert (police) recommendation to fit with police descriptions and data occurrences which they archive. Knowing what the official sources report or document is imperative to assisting in tackling the challenge of crime and coming up with viable initiatives. This essentially means that actual crime reports archived by the police should be accessible by crime analysts and

data mining researchers for effective knowledge support. Thus future research could explore deploying CriClust using actual crime records.

5.5.4 Long Term Qualitative Consideration and Evaluation

A long term qualitative evaluation of CriClust system with crime intelligence expert is worthwhile. This could be achieved by actually deploying the system with crime intelligence units in developing nations and evaluating its usefulness over a period of time, in terms of crime control and deterrence targets achieved in such a nation.

References

[1] Statistics South Africa. Victim of crime survey (VOCS), 2014.

[2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the SIGMOD International Conference on Management of Data*, pages 207–216. ACM, 1993.

[3] M. Ankerst, M. Breunig, H. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the International conference on Management of Data*, pages 49–60. ACM, 1999.

[4] M. Ashby and J. Bowers. A comparison of methods for temporal analysis of aoristic crime. *Springer Open Journal of Crime Science*, 2(1):1–16, 2013.

[5] K. C. Baumgartner, S. Ferrari, and C. G. Salfati. Bayesian network modeling of offender behavior for criminal profiling. In *Proceedings of the IEEE Conference on Decision and Control and the European Control Conference*, pages 2702–2709. IEEE, December 2005.

[6] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2008*, pages 243–254. Atlanta, Georgia, USA, April 2008.

[7] N. Bouhana, S. D. Johnson, and M.D. Porter. Consistency and specificity in burglars who commit prolific residential burglary: Testing the core assumptions underpinning behavioural crime linkage. *Legal and Criminological Psychology*, 2014.

[8] G. Breetzke. Geographical information systems (GIS) and policing in South Africa: a review. *Policing: An International Journal of Police Strategies and Management*, pages 723–740, 2006.

[9] N. Brodie. Guide: Understanding crime statistics in South Africa - what you need to know, 2015. Retrieved from: http://www.unfpa.org/swop (accessed on 14 October 2015).

[10] A. Buczak and C. Gifford. Fuzzy association rule mining for community crime pattern discovery. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, pages 1–10. ACM Washington D.C. USA, 2010.

[11] S. Chainey and J. Ratcliffe. *GIS and Crime Mapping*. Wiley Online Libraries, West Sussex, England, 2005.

[12] G. Chartrand. A graph theoretic approach to a communication problem. *In Journal of Applied Mathematics*, 14(4):778–781, 1966.

[13] C. S. Chekuri, A. V. Goldberg, D. R. Karger, M. S. Levine, and C. Stein. Experimental study of minimum cut algorithms. In *Proceedings of the Eight Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'97)*, pages 324–333. ACM, New York 1997.

[14] H. Chen, W. Chung, Y. Qin, M. Chau, J. Xu, G. Wang, R. Zheng, and A. Homa. Crime data mining: An overview and case studies. In *Proceedings of the Annual National Conference on Digital Government Research*, pages 1–5. Boston, 2003.

[15] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau. Crime data mining: A general framework and some examples. *In Journal of IEEE Computer*, 37(4):50–56, April 2004.

[16] H. Cheng, X. Yan, and J. Han. IncSpan: Incremental mining of sequential patterns in large database. In *Proceedings of the IEEE International Conference on Knowledge Discovery and Data Mining*, pages 527–532. Seattle WA, August 2004.

[17] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–268. ACM New York (NY) USA, 2001.

[18] V. R. Clarke and M. M Felson. Introduction: Criminology, routine activity, and rational choice. *International Journal of Advances in Criminological Theory: Routine Activity and Rational Choice*, 5:1–14, 1993.

[19] T. Cocx, W. Kosters, and J. Laros. An early warning system for the prediction of criminal careers. *In Advances in Artificial Intelligence: Lecture Notes in Computer Science*, 5317:77–89, 2008.

[20] K. Dahbur and T. Muscarello. Classification system for serial criminal pattern. *Artificial Intelligence and Law*, 11(4):251–269, 2003.

[21] S. Deshpande and V. Thakare. Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS)*, 1(1):32–44, 2010.

[22] B. Devesh. Emerging trends in utilisation of data mining in criminal investigation: An overview. *International Journal of Environmental Science, Computer Science and Engineering and Technology (JECET)*, 1(2):124–131, 2012.

[23] H. Deylami and Y. Singh. Adaboost and SVM based cybercrime detection and prevention model. *International Journal of Artificial Intelligence Research*, 1(2):117–130, 2012.

[24] S. Dirks, C. Gurdgiev, and M. Keeling. Smarter cities for smarter growth: How cities can optimize their systems for the talent-based economy, 2010. Retrieved from:ftp://public.dhe.ibm.com/common/ssi/ecm/en/gbe03348usen/GBE03348USEN.PDF. (accessed on 25 April 2013).

[25] M. R. D'Orsogna and M. Perc. Statistical physics of crime: A review. *Phys Life Rev: Elsevier*, pages 1–21, 2014.

[26] Y. Doytsher, P. Kelly, R. khouri, R. McLaren, H. Mueller, and C.A. Potsiou. Rapid urbanization and mega cities: The need for spatial information management. *FIG Commission*, 3(48):1–91, 2010.

[27] T. Dray and C. A. Manogue. The geometry of the dot and cross products. *Journal of Online Mathematics and Its Applications*, 1156(6):1–13, June 2006.

[28] H. Ellingwood, R. Mugford, C. Bennell, T. Melnyk, and K. Fritzon. Examining the role of similarity coefficients and the value of behavioural themes in attempts to link serial arson offences. *Journal of Investigative Psychology and Offender Profiling*, 10(1):1–27, January 2013.

[29] Innovation Enterprise. 7 uses for analytics in smart cities, 2015. Retrieved from: https://channels.theinnovationenterprise.com/articles/158-7-uses-for-analytics-in-smart-cities (accessed on 5 November, 2015).

[30] L. W. Evett, G. Jackson, D. V. Lindley, and D. Meuwly. Logical evaluation of evidence when a person is suspected of committing two separate offences. *Journal of Science and Justice*, 46(1):25–31, Elsevier 2006.

[31] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence.*, 17(3):37–54, 1996.

[32] L. Lewis G. Jakobson, J. Buford. Situation management: Basic concepts and approaches, information fusion and geographic information systems. In *Lecture Notes in Geo-information and Cartography*, pages 18–33, 2007.

[33] Seoul Metropolitan Government. *Smart Cities-Seoul: A Case Study*. ITU-T Technology Watch Report, 2013.

[34] E. R. Groff and N. G. La Vigne. Forecasting the future of predictive crime mapping. *Crime Prevention Studies*, 13:29–57, 2002.

[35] C. Hafedh, G. Ramon, A. Theresa, N. Taewoo, M. Sehl, J. Hans, W. Shawn, and N. Karine. Understanding smart cities: An integrative framework. In *Proceedings of the Hawaii International conference on System Sciences (HICSS)*, pages 2289–2297, 2012.

[36] J. Han, J Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation. *ACM International Journal of Data Mining and Knowledge Discovery*, 8(1):53–87, Kluwer Academic Publishers Netherlands 2004.

[37] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *In Journal of Information Processing Letters*, 76(4-6):175–181, 2000.

[38] Craig Haskins. Crime in Cape Town: Some strategic considerations with respect to the use of information, 2007.

[39] M. Helbich, J. Hagenauer, M. Leitner, and R. Edwards. Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. *International Journal of Cartography and Geographic Information Science*, 40(4):1–11, Taylor and Francis group 2013.

[40] F. Húffner, C. Komusiewicz, and M. Sorge. Finding highly connected subgraphs. In *Proceedings of the 41st International Conference on Current Trends in Theory and Practice of Computer Science, Pec pod Sněžkou, Czech republic*, pages 1–20, January 2015.

[41] O. Isafiade and A. Bagula. Citisafe: Adaptive spatial pattern knowledge using Fp-growth algorithm for crime situation recognition. In *Proceedings of the IEEE International Conference on Ubiquitous Intelligence and Computing*, pages 551–556. IEEE, December 2013.

[42] O. Isafiade, A. Bagula, and S. Berman. A revised frequent pattern model for crime situation recognition based on floor-ceil quartile function. In *Proceedings of the 3rd International Conference on Information Technology and Quantitative Management (ITQM)*, pages 251–260. Elsevier, August 2015.

[43] O. E. Isafiade, A. B. Bagula, and S. Berman. *On the use of Bayesian Network in Crime Suspect Modelling and Legal Decision Support*. IGI-Global USA, 2016.

[44] O. E. Isafiade and A.B. Bagula. *Data Mining Trends and Applications in Criminal Science and Investigations*. Advances in Data Mining and Database Management. pp. 1-386, IGI Global, 2016.

[45] J. Ferreira J, P. Joáo, and J. Martins. GIS for crime analysis: Geography for predictive models. *In Electronic Journal of Information Systems Evaluation*, 15(1):36–49, 2012.

[46] G. Jiji and S. Anantharadha. Automatic tracking of criminals using data mining techniques. *Journal of the Institution of Engineers (India)*, 93(4):217–221, Springer-Verlag 2013.

[47] B. Kadhim. A proposed framework for analyzing crime data set using decision tree and simple k-means clustering algorithms. *Journal of Kufa for Mathematics and Computer*, 1(3):8–24, 2011.

[48] L. Kangas. Artificial neural network system for classification of offenders in murder and rape cases. *National Criminal Justice Reference Service (NCJRS)*, 20849-6000:1–38, 2001.

[49] D. R. Karger. Global min-cuts in RNC, and other ramifications of a simple min-cut algorithm. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 21–30. ACM, New York 1993.

[50] D. R. Karger and C. Stein. A new approach to the minimum cut problem. *Journal of ACM*, 43(4):601–640, 1996.

[51] M. Kaur, S. Vashisht, and K. Saurabh. Adaptive algorithm for cyber crime detection. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 3(3):4381–4384, 2012.

[52] I. Kenji, Y. Ryuichi, and K. Kaoru. Current status of implementation for smart and resilient community in Japan. In *Proceedings of the IEEE International Conference on Smart Grid Engineering (SGE'12) UOIT, Oshawa, ON*, pages 1–8, August 2012.

[53] M. Keyvanpour, M. Javideh, and M. Ebrahimi. Detecting and investigating crime by means of data mining: a general crime matching framework. In *Proceedings of the World Conference on Information Technology*, pages 872–880. Elsevier, ScienceDirect 2010.

[54] N. Khan and V. Bhagat. Effective data mining approach for crime-terrorpattern detection using clustering algorithm technique. *International Journal of Engineering Research and Technology (IJERT)*, 2(4):2043–2048, April 2013.

[55] R. Krishnamurthy and J. Kumar. Survey of data mining techniques on crime data analysis. *International Journal of Data Mining Techniques and Applications*, 1(2):117–120, 2012.

[56] C. Ku, A. Iriberri, and G. Leroy. Crime information extraction from police and witness narrative reports. In *Proceedings of the IEEE International Conference on Technologies for Homeland Security*, pages 193–198. Boston MA, 2008.

[57] P. A. Langan and D. J. Levin. Recidivism of prisoners released in 1994, 2002. Number 193427 in BJS Special Reports, U.S. Department of Justice: Retrieved from: http://www.bjs.gov/content/pub/pdf/rpr94.pdf (accessed on 22 January 2016).

[58] J. L. LeBeau and M. Leitner. Introduction: Progress in research on the geography of crime. *The Professional Geographer*, 63(2):161–173, 2011.

[59] H. Li, L. Xue, Y. Zhu, and C. Yang. The application and implementation research of smart city in china. In *Proceedings of the IEEE International Conference on System and Science and Engineering*, pages 288–292. Dalian Liaoning, June 2012.

[60] S. Li, S. Kuo, and F. Tsai. An intelligent decision-support model using FSOM and rule extraction for crime prevention. *International Journal of Expert Systems with Applications*, 37(10):7108–119, 2010.

[61] S. Lin and D. E. Brown. An outlier based data association method for linking criminal incidents. *Decision Support Systems*, 41(3):604–615, March 2006.

[62] F. Lourenço, V. Lobo, and F. Bação. Binary-based similarity measures for categorical data and their application in self-organizing maps. In *Proceedings of the XI Jornadas de Classificao e Anlise de Dados Lisboa*, pages 1–18. JOCLAD, April 2004.

[63] T. Madzingaidzo, T. Chirema, and T. Mokwena. Sherlock: Crime data mining. *UCT Computer Science Capstone Mini-Project*, pages 1–20, 2015.

[64] A. Malathi and S. Baboo. Enhanced algorithm to identify change in crime patterns. *International Journal of Combinatorial Optimization Problems and Informatics (IJCOPI)*, 2(3):32–38, 2011.

[65] A. Malathi and S. Baboo. An enhanced algorithm to predict future crime using data mining. *International Journal of Computer Applications*, 21(1):1–6, 2011.

[66] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons-New York, 1996.

[67] T. Van Mele, L. Lachauer, M. Rippmann, and P. Block. Geometry-based understanding of structures. *Journal of the International Association for Shell and Spatial Structures*, 53(4):285–295, October 2012.

[68] A. Milgram. Why smart statistics are the key to fighting crime, 2013. TED Talk, Retrieved from: https://www.ted.com/talks/anne-milgram-why-smart-statistics-are-the-key-to-fighting-crime (accessed on 5 November, 2015).

[69] A. Mohammad, J. Mohsen, E. Martin, G. Uwe, and F. Richard. Crimewalker: A recommendation model for suspect investigation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 1–8. ACM, October 2011.

[70] P. Mohan, S. Shekhar, J. A. Shine, J. P. Rogers, Z. Jiang, and N. Wayant. A neighborhood graph based approach to regional co-location pattern discovery: A summary of results. In *Proceedings of the SIGSPATIAL GIS*, pages 2289–2297. ACM, November 2011.

[71] B. Mohl. Commonwealth politics, ideas and civic life in Massachusetts, 2015. Retrieved from: http://commonwealthmagazine.org/criminal-justice/018-67-favor-crime-prevention-rehab/ (accessed on 16 February 2015).

[72] A. Musdholifah and S. Hashim. KNN-kernel based clustering for spatio-temporal database. In *Proceedings of the IEEE International Conference on Computer and Communication Engineering (ICCCE)*, pages 1–6. Kuala Lumpur Malaysia, 2010.

[73] S. V. Nath. Crime pattern detection using data mining. In *Proceedings of the web intelligence and intelligent agent technology workshops*, pages 41–44. IEEE, 2006.

[74] NCPC. A manual for community based crime prevention-making South Africa safe, 2015. Retrieved from: National Crime Prevention Centre, Department of Safety and Security CSIR Pretoria (accessed on 10 April 2015).

[75] NCPS. National crime prevention strategy: Summary, 2015. Retrieved from: http://www.gov.za/documents/national-crime-prevention-strategy-summary (accessed on 12 January 2015).

[76] D. B. Neill and G. F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning*, 79(3):261–282, 2010.

[77] Eye Witness News. UCT suspects a serial rapist is operating near its campus, 2016. Retrieved from: http://ewn.co.za/2016/01/24/UCT-suspects-a-serial-rapist-is-operating-near-its-campus).

[78] International Association of Crime Analysts (IACA). *Crime Pattern Definitions for Tactical Analysis*. Standards, Methods, and Technology (SMT) Committee White Paper, 2011.

[79] Republic of South Africa. Western cape government report, 2013. Retrieved from: http://www..westerncape.gov.za/eng/pubs/ (accessed on 15 May 2013).

[80] F. Ozgul, C. Atzenbeck, A. Celik, and Z. Erdem. Incorporating data sources and methodologies for crime data mining. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 176–180. IEEE, 2011.

[81] J. Pattillo, N. Youssef, and S. Butenko. On clique relaxation models in network analysis. *European Journal of Operational Research*, 1(226):9–18, 2013.

[82] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning*, pages 727–734. ACM, 2000.

[83] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood. *Predictive Policing- The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation, 2013.

[84] P. Phillips and I. Lee. Mining co-distribution patterns for large crime datasets. *International Journal of Expert Systems with Applications*, 39(14):11556–11563, 2012.

[85] M. D. Porter. A statistical approach to crime linkage. arxiv e-prints. pp. 1-33, 2014.

[86] Predpol. Predpol predicts gun violence, 2013. Retrieved from: http://cortecs.org/wp-content/uploads/2014/10/predpol-gun-violence.pdf (accessed on 22 January, 2016).

[87] M. Price and P. Ball. The limits of observation for understanding mass violence. *Canadian Journal of Law and Society / Revue Canadienne Droit et Société*, pages 1–22, June 2015.

[88] V. Ramanathan and H. Wechsler. Phishing website detection using latent dirichlet allocation and adaboost. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 102–107. Washington, D.C. USA, 2012.

[89] M. Riesen and G. Serpen. A Bayesian belief network classifier for predicting victimization in national crime victimization survey. In *Proceedings of the IEEE International Conference on Artificial Intelligence*, pages 648–652. Las Vegas USA, 2009.

[90] South African Police Service. SAPS together squeezing crime to zero: Annual report, 2011. Retrieved from: http://www.saps.gov.za. (accessed on 12 February 2015).

[91] South African Police Service. Crime statistics overview:RSA, 2012. Retrieved from: http://www.saps.gov.za/statistics/reports/crimestats/2012/downloads (accessed on 15 May 2015).

[92] K. Shivsubramani and A. Ashok. M-Urgency: A next generation context-aware public safety application. In *Proceedings of the 13th International Conference on Human Computer Interaction With Mobile Devices and Services (MobileHCI)*, pages 647–652, 2011.

[93] South-Africa. Criminal law (sexual offences and related matters) amendment act, 2007. Retrieved from: http://www.justice.gov.za/legislation/acts/2007-032.pdf (accessed on August 2015).

[94] South Africa Info. South Africa two-thirds urbanised, 2013. Retrieved from: http://www.southafrica.info/news/urbanisation-240113.htm.VTY06uSdmPQ (accessed on 25 April 2015).

[95] S. Sumit, B. Fenye, L. Chang-Tien, and C. Ing-Ray. Crowdsafe: Crowd sourcing of crime incidents and safe routing on mobile devices. In *Proceedings of the 9th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 521–524. ACM, 2011.

[96] D. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi. Crime detection and criminal identification in india using data mining techniques. *AI and Society: International Journal of Knowledge, Culture and Communication*, 30(1):117–127, Springer London 2015.

[97] UN-Habitat. State of the world cities:prosperity of cities. *United Nations Human Settlements Programme (UN-HABITAT)*, 1(1):1–152, https://sustainabledevelopment.un.org/content/documents/745habitat.pdf 2012-2013.

[98] UNFPA. UNFPA State of the World Population Report, 2014. Retrieved from: http://www.unfpa.org/swop (accessed on 02 May 2015).

[99] A. Veremyev, O.A. Prokopyev, V. Boginski, and E. L. Pasiliao. Finding maximum subgraphs with relatively large vertex connectivity. *European Journal of Operational Research*, 2(239):349–362, 2014.

[100] N. Vincent, C. Stephen, L. Derek, and Y. Cheung. Incremental mining for temporal association rules for crime pattern discoveries. In *Proceedings of the ACM eighteenth conference on Australasian database*, pages 123–132. Australia, 2007.

[101] E. Vladimir and L. Ickjai. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Proceedings of the 6th International Conference on Geocomputation*, pages 1–11, 2001.

[102] C. Vlek, H. Prakken, S. Renooij, and B. Verheij. Modeling crime scenarios in a Bayesian network. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 150–159. ACM, 2013.

[103] B. Wang, H. Dong, A. Boedihardjo, C. Lu, H. Yu, I. Chen, and J. Dai. An integrated framework for spatio-temporal-textual search and mining. In *Proceedings of the ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL GIS)*, pages 570–573. ACM, 2012.

[104] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito. Crime hotspot mapping using the crime related factors: a spatial data mining approach. *Journal of Applied Intelligence*, 39(4):772–781, December 2013.

[105] T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Detecting patterns of crime with series finder. In *Proceedings of the European Conference on Machine Lerning and Principles and Practice of Knowledge Discovery in Databases*, pages 515–530. ECML-PKDD, 2013.

[106] T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Learning to detect patterns of crime. In *Machine Learning and Knowledge Discovery in Databases: Lecture Notes in Computer Science*, pages 515–530. Springer-Verlag Berlin Heidelberg, 2013.

[107] T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Finding patterns with a rotten core: Data mining for crime series with core sets. *Big Data*, 3(1):3–21, March 2015.

[108] White House Washington. Rape and Sexual Assault: A renewed call to action. *White House Council on Women and Girls*, 1(1):1–34, Jnauary 2014.

[109] D. Weisburd. Bringing social context back into the equation. *In Criminology and Public Policy*, 2(11):317–326, 2012.

[110] D. Weisburd and L.G. Mazerolle. Crime and disorder in drug hot spots: Implications for theory and practice in policing. *In Police Quarterly Journal*, 3(3):331–349, 2000.

[111] J. Woodhams, C. Hollin, and R. Bull. The psychology of linking crimes: A review of the evidence,. *Legal and Criminological Psychology*, 12:233–249, 2007.

[112] J. Woodhams and G. Labuschagne. A test of case linkage principles with solved and unsolved serial rapes. *Police and Criminal Psychology*, 27:8598, 2012.

[113] M. Yahia and M. El-taher. A new approach for evaluation of data mining techniques. *International Journal of Computer Science Issues (IJCSI)*, 7(5):181–186, 2010.

[114] Y. Yang and B. Padmanabhan. GHIC: A hierarchical pattern-based clustering algorithm for grouping web transactions. *In Transactions On Knowledge And Data Engineering*, 17(9):1300–1304, 2005.

[115] C.H. Yu, M. W. Ward, M. Morabito, and W. Ding. Crime forecasting using data mining techniques. In *Proceedings of the Eleventh International Conference on Data Mining Workshops*, pages 779–786. IEEE, 2011.

[116] C. Zhang and Z. Fang. An improved k-means clustering algorithm. *Journal of Information and Computational Science*, 10(1):193–199, 2013.

[117] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD international conference on Management of data SIGMOD'96*, pages 103–114. ACM, 1996.

[118] S. Ziauddin, K. Kammal, M. Khan, and Khan. Research on association rule mining. *Journal of Advances in Computational Mathematics and its Applications (ACMA)*, 2(1):226–236, July 2012.

[119] J. Zoete, M. Sjerps, D. Lagnado, and N. Fenton. Modelling crime linkage with Bayesian networks. *International Journal of Science and Justice: Elsevier Ireland*, 7(5):1–9, 2014.

Appendix A

Implementation Visualisations - CriClust System

A.1 CriClust System Login Interface

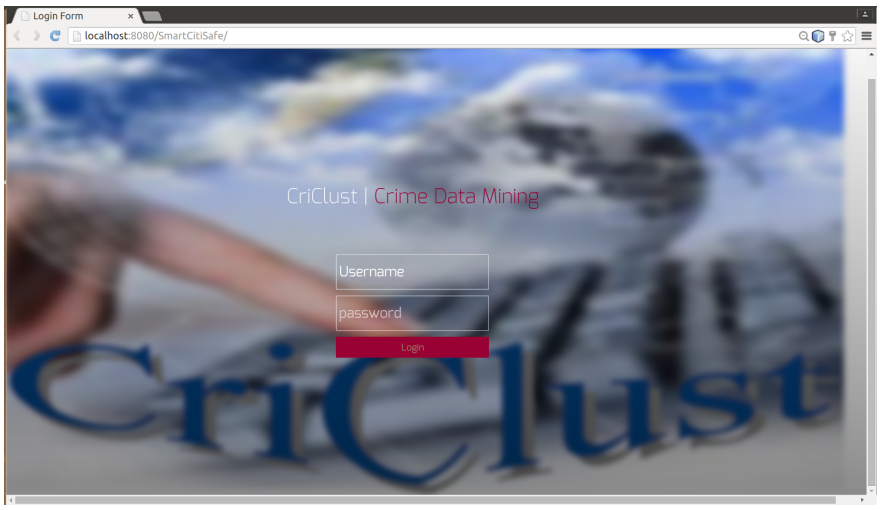


Figure A.1: CriClust system login interface for authentication, access and privacy control: By considering factors and concerns relating to the sensitivity and peculiarity of a crime mining system, the use of the system requires approved authentication by a public safety personnel or the user thereof.

A.2 Process Selection Interface for Data Processing Features

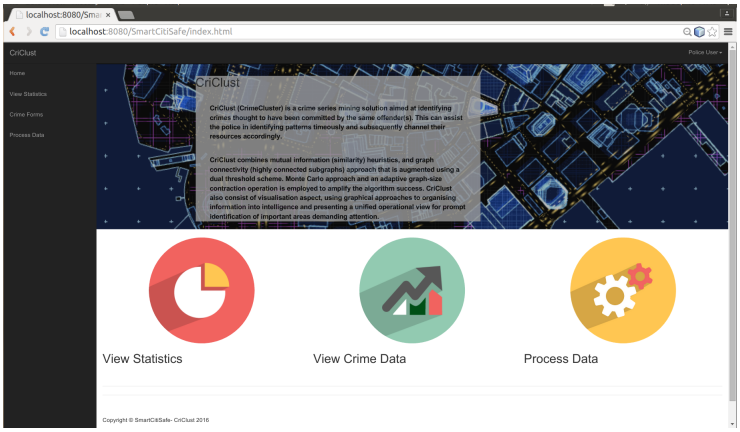


Figure A.2: CriClust selection interface for data processing features: Upon successful login, the functionalities of the system use this interface for flexible feature selection, based on the functional environment and the interesting facts required by public safety officers.

A.3 Cluster Processing Interface

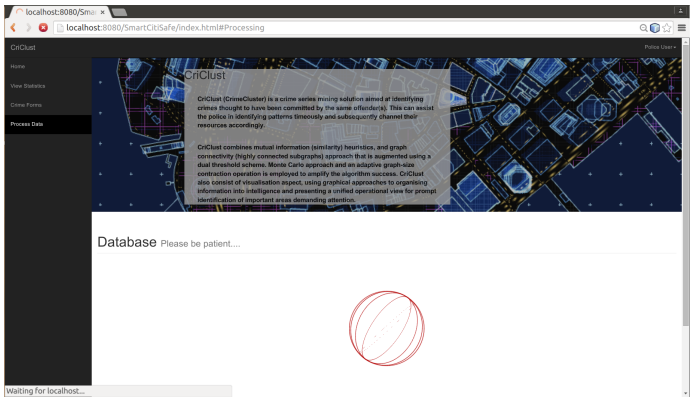


Figure A.3: CriClust processing interface: on selecting the "Process Data" feature, the system systematically accesses the database and process the required cluster. The completion of the process activates or enables the "cluster" tab on the left pane of the system to be able to view the map.

A.4 Interface for Graduated Colour Map of Cluster Information

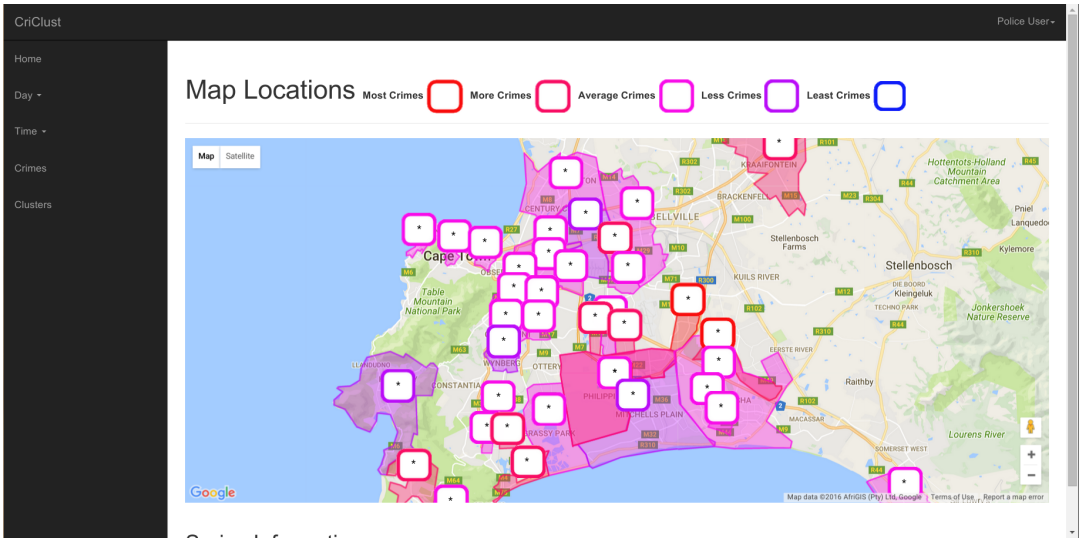


Figure A.4: Identified locations of crime: The system offers a high level view of information on the level of crime at a particular location. Each colour code is representative of the crime intensity at the highlighted location as captioned in the map (ranging from "more" crimes to "least" crimes). This is synonymous to hotspot-related identification.

A.5 Cluster Information Interface

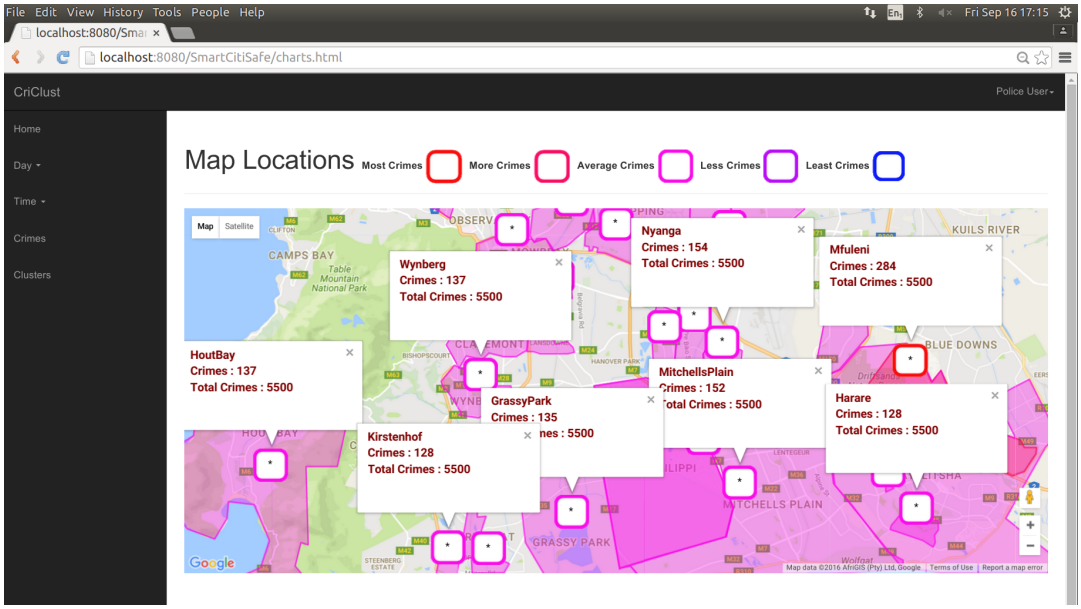


Figure A.5: The system reveals crimes per suburb: this example shows crime density in 8 locations based on information in 5500 records.

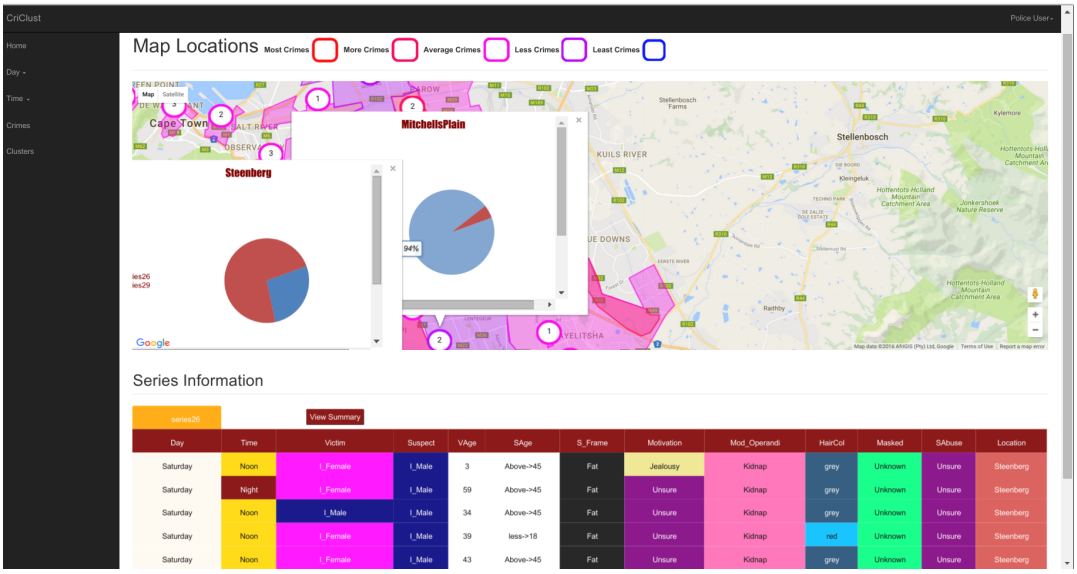


Figure A.6: An instance of PDE and raw data of the dominant series identified at Steenberg location, where crimes were committed by a fat Indian male who kidnaps victims on Saturdays, usually around noon.

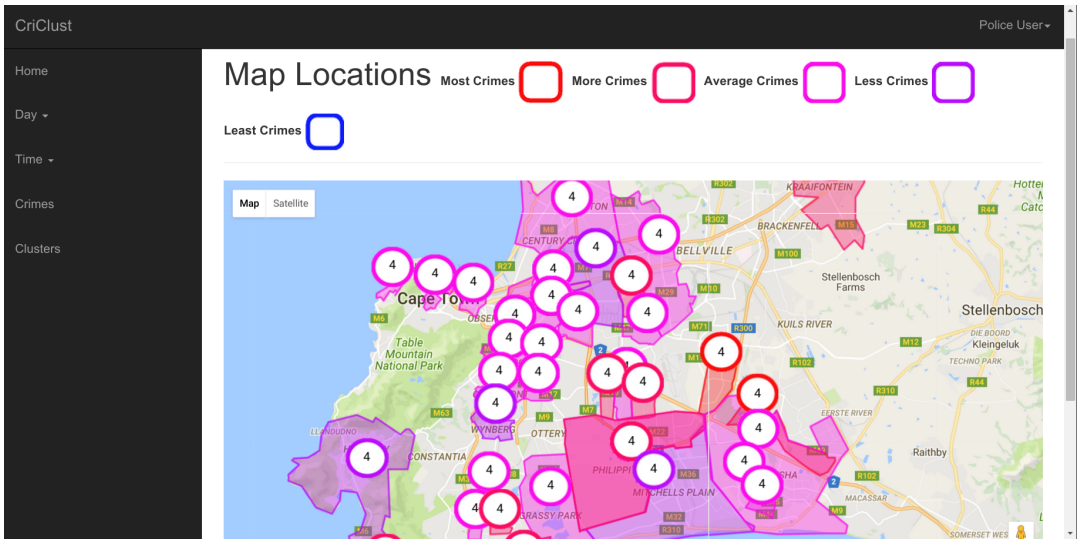


Figure A.7: CriClust performance on controlled experimental crime dataset where n=4: That is 4-series were generated per location showing these were correctly detected as 4 clusters.

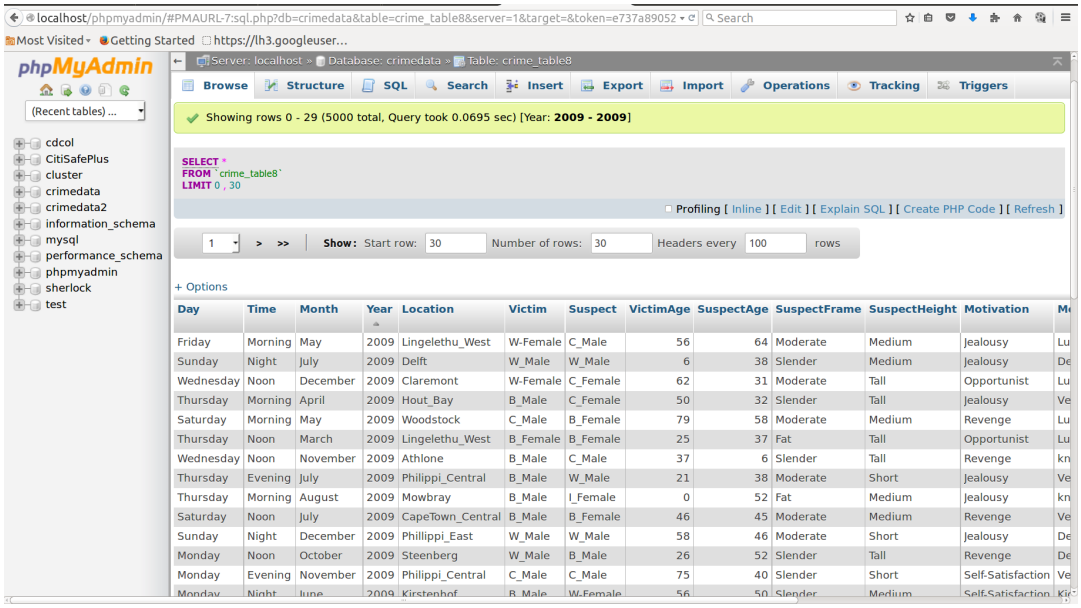


Figure A.8: Database technology for connecting to CriClust: the system relies on a MySQL database (XAMPP for linux, PHP Version 5.5.6) technology to run its features.

Appendix B

CriClust Sample Back-End Information

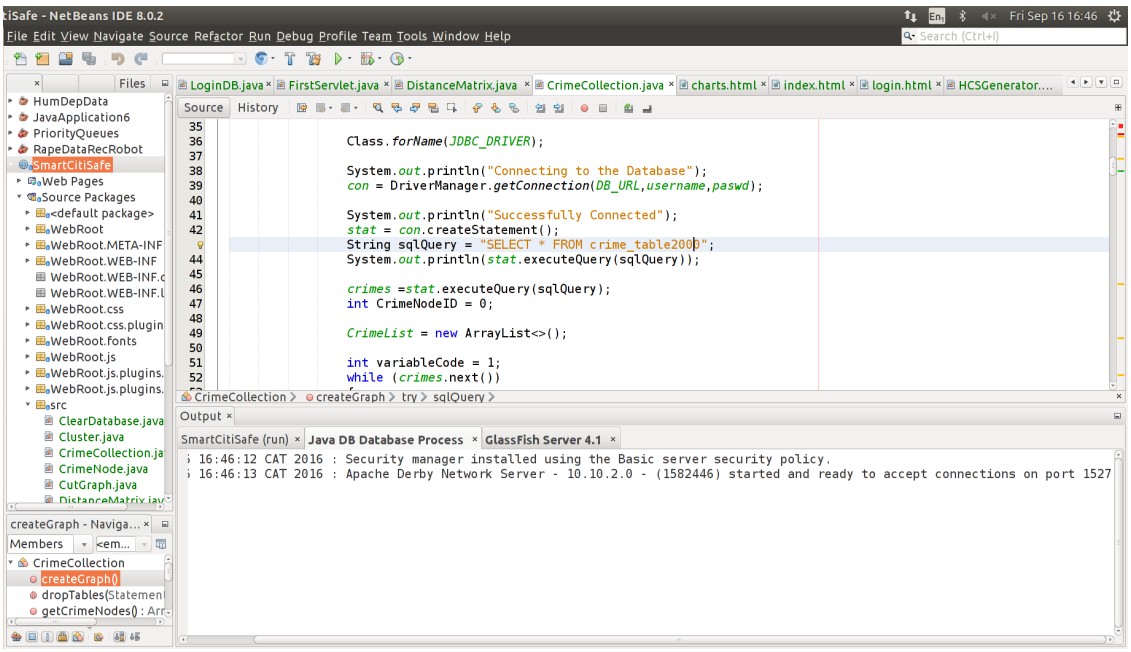


Figure B.1: Back-end interface: Connection to the database. The CriClust system depends on the MySQL database structure using the Java Database Connectivity (JDBC) API and XAMP for linux technology.

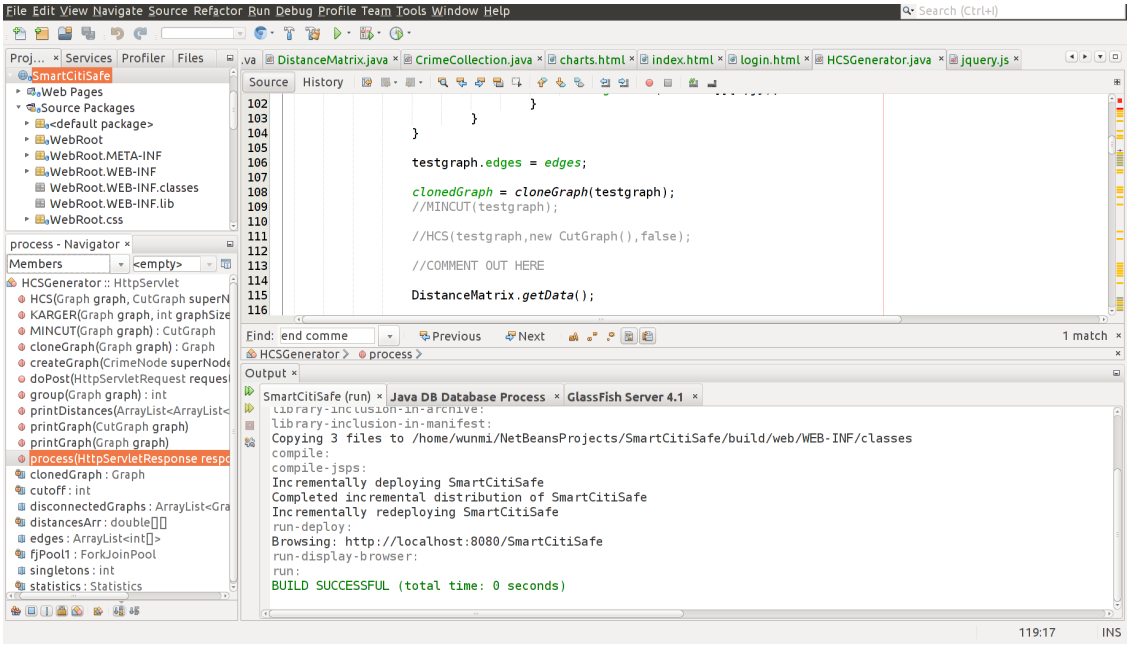


Figure B.2: Back-end interface: Deploying CriClust. When the CriClust system is launched, it establishes connection with the database and once connection to the database is successful, it deploys the application for further process feature selection and cluster processing.

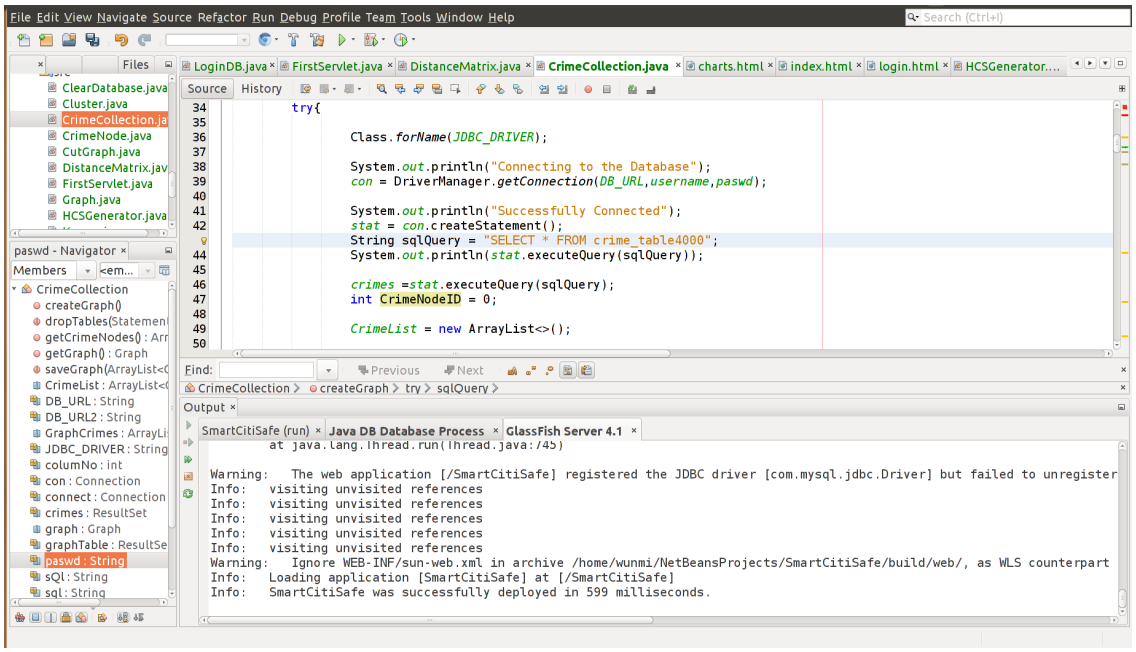


Figure B.3: Back-end interface: Deploying CriClust with 4000 records. The log information is steadily documented on the GlassFish server as the process executes.

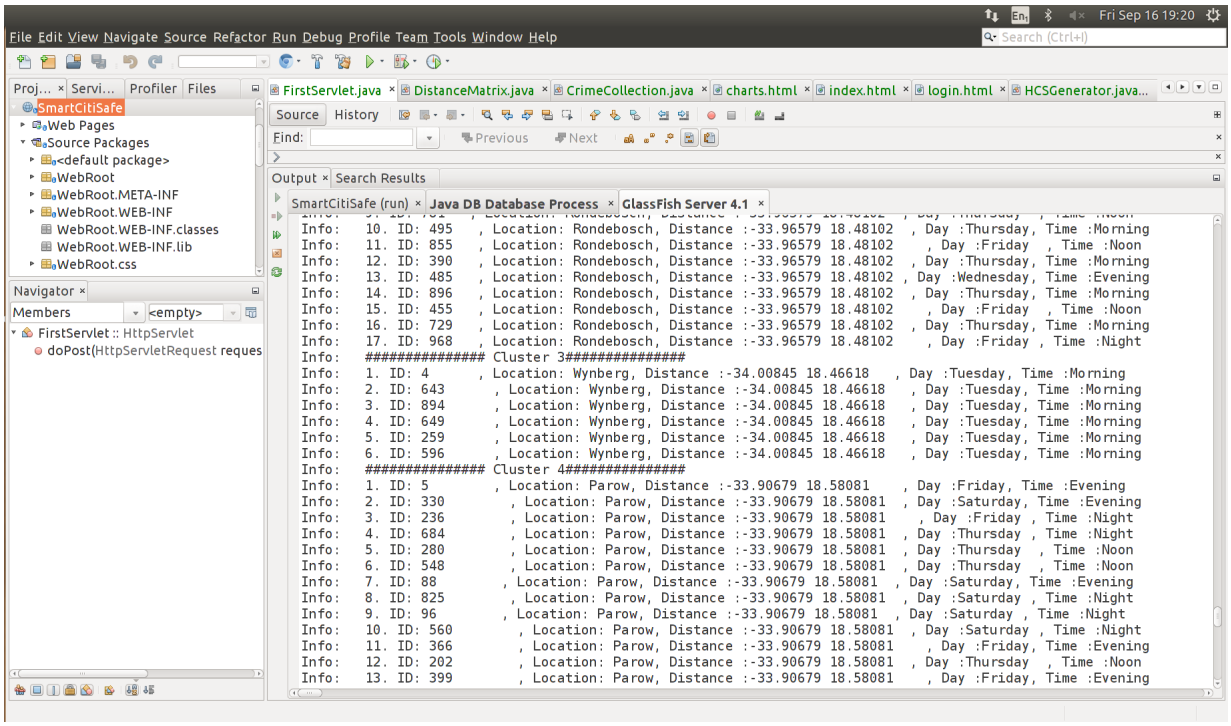


Figure B.4: Back-end interface: Information on cluster processing. The system progressively and systematically identifies and groups clusters based on the similarity condition and the full pipeline for the CriClust model.

B.1 GPS Coordinates of Some Locations

Coordinates.js ×

```
var coordinates = {Atlantis:{lat:{lat1:-33.471857, lat2:-33.608355},lng:{lng1:18.446434, lng2:18.520935}},
            Strand:{lat:{lat1:-34.095505, lat2:-34.136435},lng:{lng1:18.830722, lng2:18.862823}},
            NoordHoek:{lat:{lat1:-34.083236, lat2:-34.118416},lng:{lng1: 18.366030, lng2:18.399675}},
            FishHoek:{lat:{lat1:-34.115163, lat2:-34.145074},lng:{lng1:18.411717, lng2:18.425364}},
            Harare:{lat:{lat1:-34.052962, lat2:-34.060251},lng:{lng1:18.670295, lng2:18.677162}},
            Kraaifontein:{lat:{lat1:-33.822836, lat2:-33.877153},lng:{lng1:18.703874, lng2:18.752625}},
            LingelethuWest:{lat:{lat1:-34.041445, lat2:-34.042814},lng:{lng1:18.659357, lng2:18.661357}},
            Khayelitsha:{lat:{lat1:-33.986797, lat2:-34.054948},lng:{lng1:18.698977, lng2:18.644217}},
            Mfuleni:{lat:{lat1:-33.986032, lat2:-34.012218},lng:{lng1:18.656260, lng2:18.685356}},
            Delft:{lat:{lat1:-33.952950, lat2:-33.991956},lng:{lng1:18.631508, lng2:18.652966}},
            MitchellsPlain:{lat:{lat1:-34.023476, lat2:-34.070130},lng:{lng1:18.541656, lng2:18.638988}},
            PhillippiEast:{lat:{lat1:-34.001728, lat2:-34.056927},lng:{lng1:18.524193, lng2:18.622040}},
            Muizenberg:{lat:{lat1:-34.106879, lat2:-34.090248},lng:{lng1:18.484668, lng2:18.495311}},
            Steenberg:{lat:{lat1:-34.072638, lat2:-34.071945},lng:{lng1:18.471668, lng2:18.471732}},
            GrassyPark:{lat:{lat1:-34.024518, lat2:-34.091502},lng:{lng1:18.491509, lng2:18.530133}},
            Kirstenhof:{lat:{lat1:-34.065203, lat2:-34.079067,},lng:{lng1:18.447661, lng2:18.457317}},
            Parow:{lat:{lat1:-33.860693, lat2:-33.932507},lng:{lng1:18.574142, lng2:18.614655}},
            Nyanga:{lat:{lat1:-33.982970, lat2:-34.000655},lng:{lng1:18.573071, lng2:18.591267}},
            ElsiesRiver:{lat:{lat1:-33.91024, lat2:-33.936737},lng:{lng1: 18.564379, lng2:18.581889}},
            HoutBay:{lat:{lat1:-34.001239, lat2:-34.078337},lng:{lng1: 18.334708, lng2:18.402686}},
            PhilippiCentral:{lat:{lat1:-34.001728, lat2:-34.056927},lng:{lng1:18.524193, lng2:18.622040}},
            BishopLavis:{lat:{lat1:-33.946231, lat2:-33.947477},lng:{lng1:18.579889, lng2:18.591089}},
            Gugulethu:{lat:{lat1:-33.964809, lat2:-33.998118},lng:{lng1:18.560642, lng2:18.577035}},
            Goodwood:{lat:{lat1:-33.886026, lat2:-33.924066},lng:{lng1:18.522316, lng2:18.569351}},
            Dieprivier:{lat:{lat1:-34.029996, lat2:-34.066265,},lng:{lng1:18.461057, lng2:18.465691}},
            Manenberg:{lat:{lat1:-33.970337, lat2:-34.001438},lng:{lng1:18.548790, lng2: 18.560892}},
            Milnerton:{lat:{lat1:-33.871578, lat2:-33.874001},lng:{lng1:18.523148, lng2:18.530872}},
            Wynberg:{lat:{lat1:-34.005108, lat2:-34.004147,},lng:{lng1:18.458660, lng2:18.476942}},
            Langa:{lat:{lat1:-33.945101, lat2:-33.946419},lng:{lng1:18.531835, lng2:18.531706}},
            Lansdowne:{lat:{lat1:-33.984194, lat2:-33.985261},lng:{lng1:18.500577, lng2:18.501994}},
            Athlone:{lat:{lat1:-33.964546, lat2:-33.967785},lng:{lng1:18.503715, lng2:18.504101}},
            Mowbray:{lat:{lat1:-33.948813, lat2:-33.947069},lng:{lng1:18.478914, lng2:18.486811}},
            Claremont:{lat:{lat1:-33.986100, lat2:-33.982186},lng:{lng1:18.469279, lng2:18.472240}},
            Pinelands:{lat:{lat1:-33.940832, lat2:-33.930898},lng:{lng1:18.507435, lng2:18.514688}},
            Maitland:{lat:{lat1:-33.922763, lat2:-33.914999},lng:{lng1:18.497802, lng2:18.527413}},
            Rondebosch:{lat:{lat1:-33.957765, lat2:-33.969369},lng:{lng1:18.475219, lng2:18.481313}},
            SeaPoint:{lat:{lat1:-33.923388, lat2:-33.910781},lng:{lng1:18.390588, lng2:18.386168}},
```

Figure B.5: GPS coordinates of locations across Western Cape, South Africa.