An Application of Geometric Data Analysis Techniques to South African Crime Data

By

Benjamin William Gurr

Thesis presented in partial fulfilment of the requirements for the

Degree of Master of Commerce in the Faculty of Economics and Management Sciences

at

Stellenbosch University.

Supervisor: Prof. N.J. Le Roux

Faculty of Economics and Management Sciences

Department of Statistics and Actuarial Science

December 2016

Declaration

By submitting this thesis/dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: Benjamin Gurr

Date: December 2016

Abstract

Due to the high levels of violent crime in South Africa, improved methods of analysis are required in order to better scrutinize these statistics. This study diverges from traditional multivariate data analysis, and provides alternative methods for analysing crime data in South Africa. This study explores the applications of several types of geometric data analysis (GDA) methods to the study of crime in South Africa, these include: correspondence analysis, the correspondence analysis biplot, and the log-ratio biplot. Chapter 1 discusses the importance of data visualization in modern day statistics, as well as the geometric data analysis and its role as a multivariate analytical tool. Chapter 2 provides the motivation for the choice of subject matter to be explored in this study. As South Africa is recognised as having the eighth highest homicide rate in the world, along with a generally high level of violent crime, the analysis is conducted on reported violent crime statistics in South Africa. Additionally, the possible data collection challenges are also discussed in Chapter 2.

The study is conducted on the violent crime statistics in South Africa for the 2004-2013 reporting period, the structure and details of which are discussed in Chapter 3. In order for this study to be comparable, it is imperative that the definitions of all crimes included are well defined. Chapter 3 places a large emphasis on declaring the exact definition of the various crimes which are utilized in this study, as recorded by the South African Police Services.

The more common approaches to graphically representing crime data in South Africa are explored in Chapter 4. Chapter 4 also marks the beginning of the analysis of the South African crime data for the 2004-2013 reporting period. Univariate graphical techniques are used to analyse the data (line graphs and bar plots) for the 2004-2013 time period. However, as it is to be expected, they are hampered by serious limitations. In an attempt to improve on the analysis, focus is shifted to geometric data analysis techniques.

The general methodologies to correspondence analysis, biplots, and correspondence analysis biplots are discussed in Chapter 5. Both the algorithms and the construction of the associated figures are discussed for the aforementioned methods. The application of these methodologies are implemented in Chapter 6. The results of Chapter 6 suggest some improvement upon the results of Chapter 4. These techniques provided a geometric setting where both the crimes and provinces could be represented in a single diagram, and where the relationships between both sets of variables could be analysed. The correspondence analysis biplot proved to have some advantages in comparison to the correspondence analysis maps, as it can display numerous metrics, provide multiple calibrated axes, and allows for greater manipulation of the figure itself.

Chapter 7 introduced the concept of compositional data and the log-ratio biplot. The log-ratio biplot combined the functionality of the biplot, along with a comparability measure in terms of a ratio. The log-ratio biplot proved useful in the analysis of the South African crime data as it expressed differences on a ratio scale as multiplicative differences. Additionally, log-ratio analysis has the property of being sub-compositionally coherent.

Chapter 8 provides the summary and conclusions of this study. It was found that Gauteng categorically has the largest number of reported violent crimes over the reported period (2004-2013). However, the Western Cape proved to have the highest violent crime rates per capita of all the South African provinces. It was noted that over the past decade South Africa has experienced a downward trend in the number of reported murders. However, there has been a spike in the number of reported cases of murder in more recent year. This is spike is mostly driven by the large increases in reported murder cases in the Western Cape, Gauteng and KwaZulu-Natal. The most notable trend seen in the South African crime data is the rapid increase in the number of reported cases of drug-related crimes over the reported period across all provinces, but more noticeably in the Western Cape and Gauteng. On a whole, a majority of the South African provinces share similar violent crime profiles, however, Gauteng and the Western Cape deviate away from other provinces. This is due to Gauteng's high association to robbery with aggravating circumstances and the Western Cape's high association to drug-related crime.

This study presents some evidence that the use of geometric data analysis techniques provides an improvement upon traditional reporting methods for the South African crime data. Geometric data analysis and its related methods should thus form an integral part of any study conducted into the topic at hand.

Opsomming

Die besonder hoë vlakke van misdaad in Suid-Afrika noodsaak verbeterde metodes van analise om hierdie statistieke te ondersoek. Hierdie ondersoek wyk af van tradisionele meerveranderlike metodes en verskaf alternatiewe metodes om Suid-Afrikaanse misdaadsyfers te analiseer. Die toepassing van verskillende vorme van geometriese data analise (GDA) tegnieke om misdaadsyfers in Suid-Afrika te ondersoek, vorm die fokus van hierdie studie. Die volgende GDA tegnieke word onder meer beskou: ooreenstemmingsanalise, die ooreenstemmingsanalise bistipping en die log-ratio bistipping. Hoofstuk 1 bespreek die belangrikheid van data-visualisering in hedendaagse statistiek sowel as GDA en die rol daarvan as ’n meerveranderlike statistiese tegniek. Hoofstuk 2 verskaf die motivering vir die onderwerp van studie in hierdie ondersoek. Aangesien Suid-Afrika algemeen erken word as die land met die agste hoogste vlak van ernstige misdaad in die wêreld te same met ’n algemeen hoë vlak van misdaad, word hierdie ondersoek uitgevoer op gerapporteerde ernstige misdaad statistieke in Suid-Afrika. Verder word die uitdagings om sodanige data in te samel ook in Hoofstuk 2 aangespreek.

Die studie word uitgevoer op gewelddadige misdaadsyfers in Suid-Afrika vir die 2004-2013 periode van rapportering. Die struktuur en besonderhede van hierdie syfers word bespreek in Hoofstuk 3. Ten einde te verseker dat hierdie studie vergelykbaar moet wees, is dit van die uiterste belang dat al die soorte misdaad wat beskou word volledig gedefinieer moet wees. Hoofstuk 3 plaas dan ook ’n hoë premie op duidelike en volledige definisies van al die soorte van misdaad wat in hierdie studie beskou word. Hierdie definisies is in ooreenstemming met hoe dit deur die Suid-Afrikaanse Polisiediens omskryf word by die rapportering van misdaad.

Die tradisionele benaderings wat gevolg word met die grafiese voorstelling van Suid-Afrikaanse misdaadgegewens word ondersoek in Hoofstuk 4. Hierdie hoofstuk vorm ook die begin van die analise van die Suid-Afrikaanse misdaadgegewens soos amptelik gerapporteer vir die tydperk 2004-2013. Eenveranderlike grafiese tegnieke (lyngrafieke en staafdiagramme) word gebruik om die data te analiseer vir die tydperk 2004-2013. Soos egter te verwagte, gaan hierdie tegnieke mank aan ernstige tekortkomings. In ’n poging om hierdie tekortkomings aan te spreek, verskuif die fokus dan na GDA tegnieke.

Die algemene metodologie onderliggend aan ooreenstemmingsanalise, bistippings en ooreenstemmingsanalise bistippings word in Hoofstuk 5 bespreek. Beide die algoritmes en die konstruksie van die geassosieerde grafiese voorstellings word bespreek vir die voorafgaande tegnieke. Die toepassing van die metodologie vind neerslag in Hoofstuk 6. Die resultate van Hoofstuk 6 dui dan ook op ’n verbetering op die resultate soos gerapporteer in Hoofstuk 4. Die tegnieke wat in Hoofstukke 5 en 6 aan die hand gedoen word, verskaf die geometriese grondslag waarop beide die misdaadtipes en die provinsies gesamentlik in ’n enkele grafiek voorgestel kan word. Sodoende word dit moontlik om die verwantskappe tussen hierdie twee stelle veranderlikes te analiseer. Dit word aangetoon dat die gebruik van ooreenstemmingsanalise bistippings bepaalde voordele inhou bo die gebruik van konvensionele ooreenstemmingsanalise diagramme aangesien dit die gebruik van verskeie metrieke, veelvuldige gekalibreerde asse sowel as ’n groter mate van manipulasie van die figuur self toelaat.

Hoofstuk 7 stel aan die orde die konsep van komposisie-data en die log-verhouding bistipping. Die log-verhouding bistipping kombineer die funksionaliteit van die konvensionele bistipping met ’n vergelykbare maatstaf in terme van ’n verhouding (ratio). Die log-verhouding bistipping blyk van waarde te wees by die analise van Suid-Afrikaanse misdaadsyfers aangesien dit verskille op ’n ratio-skaal as multiplikatiewe verskille uitdruk. Verder het die log-verhouding analise die eienskap om sub-komposisie koherent te wees.

Hoofstuk 8 bevat die opsomming en gevolgtrekkings van hierdie studie. Dit is aangetoon dat vir Gauteng per kategorie die grootste getal gewelddadige misdade gerapporteer is oor die tydperk van ondersoek (2004-2013). Dit het egter geblyk dat die Wes-Kaap die grootste gewelddadige misdaadkoerse per capita van al die Suid-Afrikaanse provinsies het. Dit is verder ook aangetoon dat gedurende die afgelope dekade daar ’n dalende tendens was in die aantal moorde wat gerapporteer is. In teenstelling met hierdie algemene tendens toon die gegewens vir die jongste jare ’n toename in gerapporteerde moorde. Hierdie toename word hoofsaaklik gedryf deur ’n toename in die aantal gerapporteerde moorde in die Wes-Kaap, Gauteng en KwaZulu-Natal. Die mees uitstaande tendens in die Suid-Afrikaanse misdaadsyfers is die snelle toename in die aantal dwelm-verwante misdade. Hoewel hierdie tendens by al die provinsies voorkom, is dit veral die geval in die Wes-Kaap en Gauteng. Oor die algemeen deel ’n meerderheid van Suid-Afrikaanse provinsies gelyksoortige gewelddadige misdaadprofiele, maar Gauteng en die Wes-Kaap toon opmerkliklike verskille hiermee. Dit is veral as gevolg van die hoë voorkoms in Gauteng van roof met verswarende omstandighede en die hoë voorkoms van dwelm-verwante misdaad in die Wes-Kaap.

Hierdie ondersoek verskaf getuienis dat die gebruik van GDA tegnieke tot ’n verbetering kan lei in die wyse waarop Suid-Afrikaanse misdaadsyfers gerapporteer word wanneer dit met die tradisionele metodes vergelyk word. GDA en verwante tegnieke behoort dus ’n integrale deel uit te maak van ’n studie van Suid-Afrikaanse misdaadsyfers.

Acknowledgements

I hereby wish to express my sincere gratitude to Prof N.J. Le Roux from the Department of Statistics and Actuarial Science at Stellenbosch University for his mentorship and guidance during this study. I would like to thank my family, especially my parents, for all the help and support throughout my academic career. I would also like to acknowledge my fiancée Sara, who provided much needed support and assistance through many long days and late nights while completing this project.

Contents

List of Figure

List of Tables

Chapter 1: Introduction

Data visualization provides an excellent approach for exploratory data analysis and is essential in presenting results in an effective manner (Chen, Hardle & Unwin, 2008). Although graphics have been used extensively in statistics for a considerable period, there is not a substantial amount of literature on the topic. The aforementioned can be attributed to what Chen *et al.* (2008:37) termed as the "modern dark ages", a period in the 20th century where advances in data visualization came to a standstill. However, after this period data visualization experienced a rebirth and subsequently in the last quarter of the 20th century, large advances were made in the topics of high-dimensional, interactive, and dynamic data visualization. Of particular interest to this study is the period from early 1970s to mid-1980s. During this period, many of the advances in data visualization focused on static graphs for multidimensional quantitative data, which allowed analysts to see relations in progressively higher dimensions (Chen *et al.*, 2008:41). At the same time, numerous advances were made in the methods for studying multidimensional contingency tables. One of the most recent contributions to the study of such tables was made by Friendly and Meyer in their book *Discrete Data Analysis with R* (Friendly & Meyer, 2016). This text aims to provide the reader with an introduction to modern methods of categorical data analysis, both discrete response data and frequency data. Additionally, there is an emphasis on the use of graphical methods for exploring data, identifying noteworthy features, visualizing fitted models, and presenting the results.

Although data visualization was previously a neglected topic in statistical texts, it has a much larger role to play in recent years. The American Statistical Association has published a quarterly journal since 1992, which focuses on computational and graphical statistics. From the back issues of The American Statistician, there is a large amount of readily available literature on the topic of data visualization and graphical statistics. Additionally, The American Statistician recently published an issue (Volume 69, Issue 4, 2015) with a specific focus on statistics and the undergraduate curriculum. Data visualisation, and its role in the study of statistics, was a highlighted topic in many of the articles within this issue.

One noteworthy assertion regarding data visualization was made by Baumer (2015:337), wherein he stated that complex statistical analysis has little value unless it can be communicated clearly, especially to the non-statistician. He further contends that often complex ideas can be lost by the manner in which they are presented, and thus great importance should be placed on how results are represented graphically. However, communicating results to the reader is not the only motivation for the recent emphasis on data visualization by statisticians. Leman, House, and Hoegh (2015:398) believe that with the increase of massive, multifaceted, and complex data, the role of data visualization far outweighs the method of classical hypothesis testing, since they are able to provide much greater insight into the problem at hand. The role of data visualization in exploratory data analysis is also emphasized, as it can uncover important aspects of a problem without needing knowledge of advanced statistical methods (Nolan & Temple Lang, 2015:297).

Graphical representations of data are no longer confined to the academic fields of science and technology, and are used in a broad spectrum of disciplines. In this particular study, the focus is turned to data visualization of violent crime in South Africa and how the application of geometric data analysis can improve upon the standard reporting procedure.

1.1 Geometric Data Analysis

A particular focal point of this study is on geometric data analysis. Geometric data analysis (GDA) is a proposed name by Le Roux and Rouanet (2004:1) for the approach of multivariate statistics that represent multivariate data sets as clouds of points, and bases the interpretation of the data on these clouds. Multivariate statistics can be classified as a branch of statistics which involves the study of several variables. The basic structure of data sets in multivariate statistics are tables with rows representing the observations/individuals, and columns representing the variables (Le Roux & Rouanet, 2004:1).

In a geometric setting, objects (points, lines, planes, geometric figures) may be described by numbers, but cannot simply be reduced to numbers (Le Roux & Rouanet, 2004:9). Geometric displays allow a considerable amount of information to be conveyed in a relatively simplistic form, which can be far more efficient than representing data by a multitude of numbers. Additionally, with the assistance of modern computing, today it is possible to represent multivariate data in a geometric manner, rather than the traditional qualitative approach.

GDA provides a mechanism for coping with multivariate data by modelling the data sets as clouds of points in multidimensional Euclidean spaces. These clouds of points are constructed from multivariate data tables, and the construction is based on the mathematical structures of abstract linear algebra (Le Roux & Rouanet, 2004:419-449). Creating these cloud structures is an imperative part of GDA. As stated by Le Roux and Rouanet, GDA is the formal-geometric approach of multivariate data analysis (Le Roux & Rouanet, 2004:8).

This study proposes the application of GDA and its related methods to the South African crime data. It is hoped that such techniques will provide a much needed improvement on the traditional univariate visualization methods which have become the standard in reporting crime statistics. The GDA methods which will be highlighted in this text are correspondence analysis, biplots, correspondence analysis biplots, and the log-ratio biplot. A detailed explanation of the aforementioned methods will be provided in later chapters.

1.2 Scope of the Thesis

This thesis follows a logical progression from an overview of the problem at hand, to the actual analysis of the data. The problem of violent crime and its effects on society are discussed in Chapter

2. Violent crime statistics are investigated on both a global and local (South African) level. Additionally, the challenges around data collection are also discussed. Chapter 3 then focuses specifically on the South African crime data sets, and discusses both the structure and definitions of the variables of interest which comprise these tables.

A univariate analysis of the data is conducted in Chapter 4. Such analysis will provide a basic understanding, as well as reveal any trends within the data. The methods utilised in Chapter 4 are regarded as the standard reporting techniques and are thus a logical starting point for the analysis. Before the application of bivariate methods can be applied, an overview of said techniques is required. Thus, Chapter 5 provides the basic methodologies for the bivariate methods utilised in Chapter 6.

Both correspondence analysis and the correspondence analysis biplots are used to analyse the South African data in Chapter 6. Such methods provide the ability to visually analyse the relationships between numerous observations and variables within a single space, thus improving upon the methods utilized in Chapter 4.

The analysis of compositional data is discussed in Chapter 7. Due to the structure of the South African crime data, it is possible to treat it as compositional data. The analysis of compositions requires special attention as the chosen method of analysis needs to be sub-compositionally coherent. Thus, the log-ratio analysis and its application is discussed in Chapter 7.

The final chapter provides a short summary of the findings and conclusions from the applications of the aforementioned methods.

Chapter 2: Crime

In 2014 the United Nations Office on Drugs and Crime (UNODC) released their *"Global Study on Homicide 2013*" report. According to the UNODC, the purpose of this report was to *"shed light on the worst of crimes - the intentional killing of one human being by another"* (UNODC, 2013:5). This report is thorough and provides clear statistics on the types of homicide committed in various countries across the world.

In 2013, intentional homicide was the cause of death of almost half a million (437 000) people across the world (UNODC, 2013:11). Of these 437 000 homicides, 36 per cent occurred in the Americas, 31 per cent in Africa, 28 per cent in Asia, while Europe and Oceania only accounted for 5 and 0.3 per cent respectively (UNODC, 2013:11). Although intentional homicide trends differ across regions, it is a global concern as almost 750 million people live in countries with high homicide levels (UNODC, 2013:12). Additionally, personal security is still regarded a major concern for more than 1 in every 10 people across the globe. Thus, the study of intentional homicide is a well-motivated area of research.

The study of intentional homicide does not only provide insight into the crime itself, but also acts as an accurate and comparable indicator for measuring violence (UNODC, 2013:11). Since the impact of homicide extends beyond the loss of human life and creates a climate of fear and uncertainty, intentional homicide (and violent crime) is a threat to the entire population (UNODC, 2013:11). It then follows that a study on intentional homicide can be extended to the study of violent crime as a whole. The UNODC's 2013 global study on homicide achieved just this (UNODC, 2013).

The UNODC aimed to provide insight into violent crimes on a global scale through the study of intentional homicide rates across the globe (UNODC, 2013:11). The UNODC created an analytical framework to help governments to develop strategies and policies in order to protect those most at risk and to address those likely to offend (UNODC, 2013:11). The onus is then on the individual governments to utilize the UNODC's framework in order to create effective strategies to combat violent crimes within their own countries.

By conducting a global study, the UNODC helped identify homicide "hot spots" which warrant further monitoring and investigation. The global average homicide rate is stated as 6.2 per 100 000 population by the UNODC. Southern Africa and Central America have rates over four times higher than the global average, making these the sub-regions with the highest homicide rates in the world (UNODC, 2013:12). South Africa, in particular, is highlighted within these "hot spots", and is regarded as having the eighth highest intentional homicide rate (30.7 per 100 000 population) in the world ("Is South Africa's murder rate in the top three globally?", 2015).

The high level of homicide rates in South Africa acts as an indicator of the high levels of violence within the country. It is then of interest that a study into the violent crimes in South Africa is conducted. This thesis aims to provide insight into the levels and trends of violent crime in South Africa over the past decade.

2.1 The South African Crime Problem

At the time of writing, the most recent report from the Institute for Security Studies (ISS) highlights South Africa's 2013 homicide rate as being 32.2 per 100 000 population ("Explaining the official crime statistics for 2013 /14", 2014:2). A rate which is approximately five times larger than the global average of 6 per 100 000 population ("Explaining the official crime statistics for 2013 /14", 2014:2). Despite the fact that South Africa has reported a slow and steady decline of the homicide rate in the past decade, there have been consecutive increases in the last two years ("Explaining the official crime statistics for 2013 /14", 2014:2).

However, it is not only murder rates which are a concern to South Africa, but rather all forms of violent crime. The fact sheet provided by the ISS touches briefly on how the sustained levels of violent crime in South Africa have had a cumulative and negative affect on South Africa's growth and development. Not only can businesses suffer a loss in revenue and productivity, but investor confidence is also negatively affected. Violent crime also affects individuals by having long term effects on their health and the ability to work, but also increases fear ("Explaining the official crime statistics for 2013 /14", 2014:1). In recent times, even the South African government has made public statements about the negative effects of violent crime, stating that both productivity and foreign investment are negatively affected by the country's high levels of violent crime ("Violent crime damaging SA economy – govt", 2013).

In order to improve the living conditions in South Africa, it is imperative that the high levels of violent crime are immediately addressed. In a report by The Centre for the Study of Violence and Reconciliation (CSVR), it is stated that *"there is not one single factor which explains the high levels of violence or violent crime in South Africa. Violent crime in South Africa, as in other countries, is therefore the product of a variety of factors*" (CSVR, 2009:10). What the specific factors are which influence crime and homicide rates in various countries around the world has been debated extensively, with academics rarely agreeing with one another, and often having opposing opinions. It is therefore an ongoing investigation into what factors explicitly affect violent crime rates in countries all across the world. However, from the literature it appears that there is a general consensus that crime is multifaceted in nature.

At the time of writing, the most recent papers on South African crime were published in the 12th edition of the Journal of Investigative Psychology and Offender Profiling. This issue was commissioned by the Federal Bureau of Investigation's National Centre for the Analysis of Violent Crime (Salfati & Labuschagne, 2015a:1). This was a special edition of the journal which had a specific focus on serial homicide in South Africa. The publication of this special issue draws attention to the fact that homicide is a prevailing issue within South Africa. Additionally, this journal highlights the fact that both South African and foreign academics have sought to identify the cause of South Africa's unusually high homicide rates, see (Horning, Salfati & Labuschagne, 2015; Salfati & Labuschagne, 2015b; Salfati, Horning, Sorochinski & Labuschagne, 2015; Salfati, Labuschagne, Horning, Sorochinski & De Wet, 2015; Sorochinski, Salfati & Labuschagne, 2015).

In summary, the violent crime and homicide rates of South Africa are a growing concern. Not only do these rates affect the health and safety of South Africans, they are also detrimental to the economy itself. It is of the utmost importance that researchers, and particularly South Africans, take all available steps to better understand violent crime and homicide in South Africa. Only by achieving a better understanding of the behaviour of violent crime in South Africa can it be effectively managed.

2.2 The UNODC and SAPS Reports

The motivation for this thesis was largely drawn from the 2013 United Nations Office on Drugs and Crime (UNODC) Global Study on Homicide, and the South African Police Service's Analysis of the National Crime Statistics (UNODC, 2013; SAPS, 2014). The Global Study on Homicide report was a study conducted across a number of countries by the UNODC in order to better understand homicide rates, and the various methods of killing across the globe. Due to the sheer scale and magnitude of the UNODC report, it is used to provide guidelines to the collection and analysis of data within this study. On the other hand, the SAPS report is strictly confined to the subject of South African crime, and provides a local perspective to the subject matter.

Both the above reports follow similar structures and use similar techniques to present the data. Both reports make extensive use of univariate data visualization methods such as line graphs, bar plots and pie charts. Such methods can be regarded as the traditional approaches to representing crime statistics. As was stated in the introduction, crime is not subject to one single factor, and can therefore be regarded as multivariate in nature. It is due to this fact that the use of GDA and its related techniques is suggested to provide a better method of visually representing the data.

One of the greatest advantages to utilizing GDA methods is that the distances between points in the displays have a meaningful interpretation. Relationships between multiple variables and observations can be detected without the explicit need for quantitative data analysis, thus providing insight into the data by perusal of a well-constructed graph. Additionally, the required number of figures is greatly reduced as multiple variables and observations can be observed in one single figure. This then provides a substantial amount of data which are communicated in one relatively easy-to-read figure.

In both the UNODC and SAPS reports, the collection of crime statistics played a critical role in their construction. The following two subsections provide an in depth overview of data collection for crime statistics, both internationally and nationally.

2.2.1 Data Collection and Challenges

Crime rates and statistics have an inherent data collection and analysis problem. Chapter 6 of the UNODC Global Study on Homicide is dedicated solely to addressing these problems (UNODC, 2013:99). A review of this chapter is a fitting start to the overview of possible data challenges which may be encountered in research utilizing homicide rates and statistics.

The Global Study on Homicide analyses the reported homicide rates of over 200 countries, 193 of which are United Nation member states. A majority of the reported statistics are gathered from national data repositories, such as the criminal justice or the public health system of the respective countries. The report states that these two sources are used because of the great importance national authorities typically devote to recording and investigating deaths resulting from violent acts. For this reason, either of the aforementioned sources is regarded as the best option available for the collection of homicide statistics (UNODC, 2013:99).

Creating the infrastructure for recording and maintaining accurate records of such statistics is a lengthy and expensive process. Some countries do not have suitable systems in place to facilitate the accurate collection and storage of such data (UNODC, 2013:99). This leads to a lack of national homicide data for some countries. The UNODC overcame this problem by using homicide data derived from the World Health Organisation (WHO) estimates. These derived estimates are not regarded as accurate measures, thus affecting the overall quality of the report. However, the UNODC does report an incremental improvement in administrative records in Africa, the Americas and Oceania between its 2011 and 2013 reports (UNODC, 2013:99).

The collection and availability of data are not the only problems facing homicide data. The issue of data quality is crucial from a statistical perspective. The quality of the data can be assessed by the accuracy and the international comparability of the data. The UNODC assessed the accuracy of the data by comparing the homicide rates reported by a country’s criminal justice and public health systems. If there are large discrepancies between the two sources it would highlight any inaccuracies in the data, whereas if both systems reported similar results one would perceive the data as being accurate. The international comparability of the data are of great importance when a study is conducted between countries. International comparability depends largely on the definition of intentional homicide used when national systems record such statistics. The UNODC defines intentional homicide at an international level as the *“unlawful death purposefully inflicted on a person by another person”* (UNODC, 2013:102). This definition is often very close to that of intentional homicide used by various countries, which results in national homicide statistics which are highly comparable at the international level.

There are a number of published academic papers where the data challenges facing homicide statistics are discussed. In contrast to the viewpoint of the UNODC, Mccall and Niewbeerta (2007:168) suggest that the varying definitions of homicide across nations makes an internationally comparative study of homicide statistics challenging. Messner (1989:599) on the other hand believed that there is in fact a distinct advantage to a cross-national design as it accounts for a large amount of variation in the variable of interest. Underreporting is also highlighted as an issue in the literature. In particular, Fajnzylber, Lederman and Loayza (2002:8) find that underreporting of homicides is an inherent problem in countries with low quality police and judicial systems, or countries with a poorly educated population.

In the paper released by Mccall & Niewbeerta (2007) they discuss both the limitations of data availability and comparisons across countries. These limitations, which have inhibited cross-national research in the past can now be eradicated. Mccall & Niewbeerta use these issues to motivate sourcing the data of the European Commission's Eurostat. Eurostat is a statistical office which has *'compiled subnational level data from many EU member countries and has worked carefully to ensure comparability of social indicators across countries'* (Mccall & Nieuwbeerta, 2007:169). In such economic or regional groups of countries there are often standardized reporting techniques and definitions, making the collection and analysis of data far easier and more reliable.

In summary, it would appear then that data sourced from affluent first world countries would be the most reliable and accurate. To ensure the best international comparability of the data one should conduct a study on a group of countries, such as the European Union, where reporting methods are standardized across all the countries. Unfortunately, research cannot be limited to the few countries which do have adequate data available. The problem of high homicide rates is prudent in poorer and developing countries which often lack good police, judicial, and education systems. In such cases having multiple sources of data would prove beneficial as then the accuracy of the data could be assessed by comparing the multiple reported statistics. If multiple sources of data are not available, then it would be impossible to assess the accuracy of the recorded statistics.

2.2.2 Data Collection and Challenges in South Africa

A detailed explanation of the data collection issues for homicide data was provided for in the previous subsection. However, this study does not limit itself to homicide statistics, and instead branches out into several types of violent crime. There is a shortage of texts available on the data collection of various types of violent crime, as the majority of the literature is dedicated to homicide statistics. In this section we apply the previous findings surrounding homicide statistics to the South African case of data collection, and further extrapolate the same principles to several other violent crimes in South Africa.

The first and most prominent issue when collecting homicide statistics was ensuring a uniform definition of the term 'homicide'. As mentioned previously, this problem arises when dealing with cross national homicide rates. This study has chosen to limit the research to South Africa to eliminate this issue. To ensure a uniform definition of homicide, only the reported homicide rates from the nine South African provinces are used. The advantages and disadvantages of limiting the research to a national level have been debated in other papers (Messner, 1989; Mccall & Nieuwbeerta, 2007). Similarly, the definitions of all violent crimes in this report will also be uniform as their definitions are derived from the same judicial system. However, definitions of crimes do change within a country over time. Often reporting standards change, or there are amendments to legislature which will change the definitions of crimes. It is therefore crucial that the definitions of the crimes in a study are clearly defined so that later comparisons of studies and results can be conducted accurately (see Subsection 3.2).

In South Africa there are three potential sources of crime data, all of which are credible. The first and foremost source would be that of the South African Police Service (SAPS). The SAPS releases yearly crime reports and the data are made freely available on the SAPS website. The SAPS faces crime in South Africa at ground level, and is the main source of recording crime statistics in South Africa. Additionally, the national crime statistics provided by the SAPS have been subject to the audit process of the Auditor-General (SAPS, 2014:1). At the time of writing, the national crime statistics are also in the advanced stages of meeting the eight South African Quality Assessment Framework (SASQAF) quality dimensions of the Statistician-General (SAPS, 2014:1). Due to the aforementioned fact, the SAPS crime statistics can therefore be regarded as highly reliable and accurate.

Other sources which are available to use are StatsSA, along with the National Injury Mortality Surveillance System (NIMSS). StatsSA is the national statistical service of South Africa, a reliable source with data readily available on the StatsSA website. Although relatively new, NIMSS has been highlighted as being a very reliable source of data on fatal injuries. The aforementioned is confirmed by Thomson, who regarded NIMSS as the most accurate available data source for such data (Thomson, 2004:10). NIMMS data are not as accessible as SAPS or StatsSA, requiring one to apply for the data from UNISA. Additionally, the NIMMS data base only provides data on causes of death, rather than crime statistics. The NIMMS data base would provide reliable data for homicide or mortality rates, however, it does not provide substantial information on crime statistics for this current study and has therefore been excluded.

It can be concluded that for the study of violent crimes in South Africa that there are two feasible options, the SAPS and StatsSA data bases, both of which are readily available to all researchers. The debate as to which data source is the most accurate is still ongoing. However, a choice of either data base would be regarded as an accurate source.

Chapter 2 has introduced the violent crime problem around the globe, and more specifically in South Africa. Additionally, the issues regarding data collection has been discussed, both on a global and local level. Chapter 3 aims to provide greater insight into the data sourced for the analysis within this study, as well as a detailed review of the variables comprising the data.

Chapter 3: The South African Crime Data

This chapter aims to provide insight into the South African crime data, its structure, form, and the definition of the variable names. The data used in this study are provided for by the South African Police Service (SAPS). Crime statistics are provided for on a yearly basis by the SAPS, and are readily available from www.saps.gov.za. The reported statistics follow the SAPS financial year as opposed to a calendar year, and run from April to March. That is to say that the 2004 reported statistics run from April 2004-March 2005. However, for simplicity throughout this study, the year 2004 will refer to the crime data for the 2004/2005 period. Similar abbreviation will apply to all other financial years.

3.1 Structure of the South African Crime Data

The South African crime data for the period 2004-2013 are presented in the form of ten contingency tables of size 12 × 9 (see Table 3-3). The rows depict the twelve violent crimes of interest to this study, and the columns represent the nine South African provinces. As the focus of this study is an application of biplots on the South African crime data, often abbreviations of the variable names are required in order to preserve the interpretability of the data when displayed graphically. The nine provinces and their respective abbreviations are presented in Table 3-1. Similarly, the twelve crime types and their respective abbreviations to be used throughout this study are presented in Table 3-2. An example of the South African crime data are provided in Table 3-3. The values which appear in Table 3-3 are frequencies of the various crimes which occur in any one of the nine provinces for the 2004/2005 reporting year. For instance, the very first cell in Table 3-3 displays the value 3 352, which represents the total number of reported cases of murder in the Eastern Cape for the April 2004-March 2005 reporting period. The remaining nine contingency tables that span the 2005-2013 period are presented in the appendix for the reader’s reference.

***Table 3-1**: South African province names and their respective abbreviations.*

Province	Abbreviation
Eastern Cape	EC
Free State	FS
Gauteng	GT
KwaZulu-Natal	KZ
Limpopo	LP
Mpumalanga	MP
North West	NW
Northern Cape	NC
Western Cape	WC

***Table 3-2**: The names of the 12 crimes of interest in this study, as well as their respective abbreviations.*

Crime	Abbreviation
Arson	Ars
Assault with the intent to cause grievous bodily harm	Agb
Attempted murder	Atm
Burglary at non-residential premises	Bnr
Burglary at residential premises	Brp
Carjacking	Crj
Common assault	Ast
Drug-related crime	Drg
Illegal possession of firearms and ammunition	Ilf
Murder	Mdr
Robbery with aggravating circumstances	Rac
Total sexual offences	Tso

***Table 3-3**: The 2004/2005 South African crime data as provided for by the SAPS.*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3352	902	3818	5001	733	1099	800	408	2680
Tso	8626	4972	16333	12122	5070	4674	4610	2212	10498
Atm	3046	1324	6661	5979	1032	1568	1065	1351	2490
Agb	40447	17998	54138	33898	17756	19978	18097	13188	33869
Ast	25763	25197	72484	37852	18148	15736	14612	9326	48739
Rac	9576	4532	57628	25207	3285	6947	5376	1095	13143
Ars	1417	503	1985	1470	762	579	506	242	720
Bnr	6415	4063	12986	8990	4994	2992	4288	2370	8950
Brp	33364	17802	77383	43122	13569	21216	15378	7353	46977
Ilf	1938	432	3974	4880	563	792	549	122	2247
Drg	9061	4063	10722	19290	1786	1714	4383	2550	30432
Crj	529	156	7230	2703	203	485	221	6	901

3.2 Definition of Variable Names

The twelve crimes which are presented in Table 3-2 were chosen to match as closely as possible, the South African crime data studied in Chapter 7 of Understanding Biplots (Gower, Gardner-Lubbe & Le Roux, 2011:312). As there are no other cited texts which make use of biplots on the study of crime in South Africa, the Understanding Biplots application is used as a comparative study, and hence the use of similar crime data. However, it must be stressed that the definitions of crimes plays an imperative role when studying such data. Reporting standards change, and so do the definition of crime types, which can lead to historical data that are not directly comparable to current statistics. For instance, the definition of total sexual offences has experienced changes in its legal definition over the reported period, which makes direct comparison to past data difficult (Republic of South Africa, 2007). Additionally, crimes which may have previously been reported as a separate statistic have been amalgamated into a broader definitions by the SAPS. It is therefore imperative that the definitions of the various crime types are clearly defined in the context of South African legislature and the reporting standards of the SAPS.

The formal definitions of the crime types studied in this thesis are provided for below. Their definitions are taken verbatim from the 2013/2014 SAPS National Crime Statistics-Addendum to the Annual Report, pages 78-83 (SAPS, 2014).

3.2.1 Assault with the Intent to Cause Grievous Bodily Harm

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, assault with the intent to cause grievous bodily harm is defined as:

"*… the unlawful and intentional direct or indirect application of force to the body of another person with the intention of causing grievous bodily harm to that person*" (SAPS, 2014:79).

3.2.2 Arson

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, arson is defined as:

"…*the unlawful and intentional setting of fire to immovable property belonging to another or to one's own immovable insured property, in order to claim the value of the property from the insurer"* (SAPS, 2014:80).

3.2.3 Common Assault

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, common assault is defined as the unlawful and intentional:

i. *"direct or indirect application of force to the body of another person*", or
ii. *"threat of application of immediate personal violence to another, in circumstances in which the threatened person is prevailed upon to believe that the person who is threatening him or her has the intention and power to carry out his threat"* (SAPS, 2014:79).

3.2.4 Attempted Murder

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, attempted murder is defined as:

"*…the commission of an unlawful act with the intention of killing another human being but which does not result in the death of that human being*" (SAPS, 2014:78).

3.2.5 Burglary at Non-Residential Premises

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, burglary at non-residential premises is defined as:

"*… a person who unlawfully and intentionally breaks into a building or similar structure which is not used for human habitation and does not form part of residential premises, and enters or penetrates it with part of his body or with an instrument with which he intends to control something on the premises, with the intention to commit a crime on the premises*" (SAPS, 2014:81).

3.2.6 Burglary at Residential Premises

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, burglary at residential premises is defined as:

"…a person who unlawfully and intentionally breaks into a building or similar structure, used for human habitation, and enters or penetrates it with part of his or her body or with an instrument with which he or she intends to control something on the premises, with the intention to commit a crime on the premises" (SAPS, 2014:81).

3.2.7 Carjacking

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, carjacking is defined as:

"…the unlawful and intentional forceful removal and appropriation of a motor vehicle (excluding a truck) belonging to another" (SAPS, 2014:79).

3.2.8 Drug-Related Crime

Drug-related crime is compiled of two sub-categories. The national crime statistics report use the definitions of such crimes provided for by the 1992 Drugs and Drug Trafficking Act No. 140. These sub-categories, and their respective definitions are presented below.

3.2.8.1 *Unlawful Use or Possession of Drugs*

As stated in Section 4 of the Drugs and Drug Trafficking Act No. 140 of 1992, the unlawful use or possession of drugs is defined as any person who uses or has in his possession:

i. "*any dependence-producing substance*", or

ii. *"any dangerous dependence-producing substance or any undesirable dependence-producing substance"* (Republic of South Africa, 1992:9).

3.2.8.2 *Unlawful Dealing in Drugs*

As stated in Section 5 of the Drugs and Drug Trafficking Act No. 140 of 1992, the unlawful dealing in drugs pertains to any person who shall deal in:

i. *"any dependence-producing substance"*, or

ii. *"any dangerous dependence-producing substance or any undesirable dependence-producing substance"* (Republic of South Africa, 1992:11).

3.2.9 Illegal Possession of Firearms and Ammunition

Illegal possession of firearms and ammunition is compiled of two sub-categories. The national crime statistics report use the definitions of such crimes provided for by the Firearms Control Act No. 60 of 2000. These sub-categories, and their respective definitions are presented below.

3.2.9.1 *Unlawful Possession of a Firearm*

Section 3 of the Firearms Control Act No. 60 of 2000, states that:

"No person may possess a firearm unless he or she holds a licence, permit or authorisation issued in terms of this Act for that firearm" (Republic of South Africa, 2001:9).

3.2.9.2 *Unlawful Possession of Ammunition*

Section 90 of the Firearms Control Act No. 60 of 2000, states that no person may possess any ammunition unless he or she:

i. *"holds a licence in respect of a firearm capable of discharging that ammunition*";

ii. *"holds a permit to possess ammunition"*;

iii. *"holds a dealer's licence, manufacturer's licence, gunsmith's licence, export or in-transit permit or transporter's permit issued in terms of this Act"* ; or

iv. *"is otherwise authorised to do so"* (Republic of South Africa, 2001:29).

3.2.10 Murder

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, murder is defined as:

"…the unlawful and intentional killing of another human being" (SAPS, 2014:78).

3.2.11 Robbery with Aggravating Circumstances

Similarly to drug-related crime, robbery with aggravating circumstances consists of several sub-categories. However, a broad definition of this category is provided for by the national crime statistics report as:

"…the unlawful and intentional forceful removal and appropriation in aggravating circumstances of movable tangible property belonging to another" (SAPS, 2014:79).

The respective definitions of the sub-categories of robbery with aggravating circumstances are presented below.

3.2.11.1 *Carjacking*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, carjacking is defined as:

"…the unlawful and intentional forceful removal and appropriation of a motor vehicle (excluding a truck) belonging to another" (SAPS, 2014:79).

3.2.11.2 *Robbery of Truck*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, robbery of truck is defined as:

"…the unlawful and intentional forceful removal and appropriation of a truck (excluding a light delivery vehicle) belonging to another" (SAPS, 2014:80).

3.2.11.3 *Cash-In-Transit Robbery*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, cash-in-transit robbery is defined as:

"…the unlawful and intentional forceful removal and appropriation of money or containers for the conveyance of money, belonging to another while such money or containers for the conveyance of money are being transported by a security company on behalf of the owner thereof" (SAPS, 2014:80).

3.2.11.4 *Bank Robbery*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, bank robbery is defined as:

"…the unlawful and intentional forceful removal and appropriation of money which belongs to a bank from the bank during the office hours of that bank" (SAPS, 2014:80).

3.2.11.5 *Robbery at a Residential Premises*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, robbery at a residential premises is defined as:

"… the unlawful and intentional forceful removal and appropriation of property from residential premises of another person" (SAPS, 2014:80).

3.2.11.6 *Robbery at Non-Residential Premises*

As stated in the addendum of An Analysis of the National Crime Statistics 2013/14, robbery at a non-residential premises is defined as:

"…the unlawful and intentional forceful removal and appropriation of property from the business of another person" (SAPS, 2014:80).

3.2.12 Total Sexual Offenses

Total sexual offenses is compiled of several sub-categories. The national crime statistics report use the definitions of such crimes provided for by the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007. These sub-categories, and their respective definitions are presented below.

3.2.12.1 *Rape*

As stated in Section 3 of the Criminal Law (Sexual Offences and Related Matters) amendment Act No. 32 of 2007, rape is defined as:

"Any person ('A') who unlawfully and intentionally commits an act of sexual penetration with a complainant ('B'), without the consent of B, is guilty of the offence of rape." (Republic of South Africa, 2007:10).

3.2.12.2 *Compelled Rape*

As stated in Section 4 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007, compelled rape is defined as:

"Any person ('A') who unlawfully and intentionally compels a third person ('C'), without the consent of C, to commit an act of sexual penetration with a complainant ('B'), without the consent of B" (Republic of South Africa, 2007:10).

3.2.12.3 *Sexual Assault*

As stated in Section 5 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007, sexual assault is defined as a person who unlawfully and intentionally:

i. *"sexually violates a complainant ('B') without the consent of B",* or
ii. *"inspires the belief in a complainant ('B') that B will be sexually violated"* (Republic of South Africa, 2007:11).

3.2.12.4 *Compelled Sexual Assault*

As stated in Section 6 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007, compelled sexual assault is defined as:

"A person ('A') who unlawfully and intentionally compels a third person ('C'), without the consent of C, to commit an act of sexual violation with a complainant ('B'), without the consent of B" (Republic of South Africa, 2007:11).

3.2.12.5 *Acts of Consensual Sexual Penetration with Certain Children (Statutory Rape)*

As stated in Section 15 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007, a person ('A') who is guilty of statutory rape:

"…commits an act of sexual penetration with a child ('B'), despite the consent of B to the commission of such an act" (Republic of South Africa, 2007:13)*.*

3.2.12.6 *Acts of Consensual Sexual Violation with Certain Children (Statutory Sexual Assault)*

As stated in Section 16 of the Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007, a person ('A') who is guilty of statutory sexual assault:

"…commits an act of sexual violation with a child ('B'), despite the consent of B to the commission of such an act" (Republic of South Africa, 2007:13)*.*

3.3 Conclusion

The reader has now been provided with an overview of the structure of the South African crime data. As previously stated, the definitions of the crime types play a crucial role in understanding reported crime statistics. It is therefore imperative that the reader has a clear understanding of the aforementioned definitions before commencing to further chapters.

Chapter 4 conducts a univariate data analysis on the South African crime data discussed above. This analysis acts as the starting point of the investigation into violent crimes in South Africa.

Chapter 4: Univariate Data Analysis

This chapter presents an application of univariate graphical techniques to the South African crime data. The univariate techniques are by far the most rudimentary techniques to be utilized in this study. Due to their simple nature, no theoretical explanation is required for the univariate methods. In univariate data analysis there is only a single variable of interest displayed at a time in the diagrams. Although the univariate analysis is elementary, it still proves beneficial.

The figures which follow will provide basic information about the available data, such as a measure of central tendency and variability in the data. It is possible to manipulate the univariate methods in order to extract a considerable amount of information from simple graphical techniques. The greatest pitfall to the univariate analysis is the number of displays required to conduct an in depth analysis. Due to this, only the most relevant and beneficial figures will appear below. The results from the aforementioned figures will then provide information on trends and focal points in the data for later chapters.

The analysis develops from basic line graphs and progresses to stacked bar plots for further analysis of the data.

4.1 Line Graphs

Figure 4-1 provides a representation of the total crime frequency per province over the ten year reporting period (2004-2013). Henceforth, “total crime frequency” will refer to the total frequency of the twelve crimes of interest specific to this study. One of the most notable aspects of Figure 4-1 is that it is split between the lower cluster of provinces and the three provinces (Gauteng, Western Cape, & KwaZulu-Natal) positioned at the top of the graph. The Eastern Cape appears to be dissecting the aforementioned two groups, and the Northern Cape is consistently the bottom runner.

It would be expected that the total crime frequency of a province would be directly related to the population size of the respective province. Contemplating this, it would be expected that the greater the population size, the higher the respective province would be located in Figure 4-1. The mid-year population estimates per province for the 2004-2013 period are provided for in Table 4-1. The provinces have been ordered from largest population to smallest in Table 4-1. Given that Gauteng and KwaZulu-Natal have categorically the largest total populations, it would be expected that they would have the largest total crime frequencies. Figure 4-1 revealed that this is in fact not the case. Gauteng consistently has the highest total crime frequency followed by the Western Cape. Although KwaZulu-Natal and the Western Cape follow almost identical trends, the Western Cape surpasses that of KwaZulu-Natal in later years (2010-2013) and is therefore regarded as having the second highest crime rate.

If the provinces all had similar crime rates per capita, then it would be reasonable to expect the province lines in Figure 4-1 to fall in order of which they appear in Table 4-1. Figure 4-1 shows this premise to be false. The Western Cape has previously been noted for its high position in the figure despite its relatively small population size. The Free State, which is positioned at the top of the lower cluster of provinces, is positioned higher than expected as it has the second lowest total population. Whereas, Limpopo is positioned fairly low on the graph when considering its population size.

There is then evidence to suggest that the total crime frequency is not the best suited measure for such comparisons. Crime rates per capita provide a more comparable measure of crime frequencies between provinces. The rates per capita can be derived by dividing the total crime frequencies of Figure 4-1 by their respective population estimates of Table 4-1.

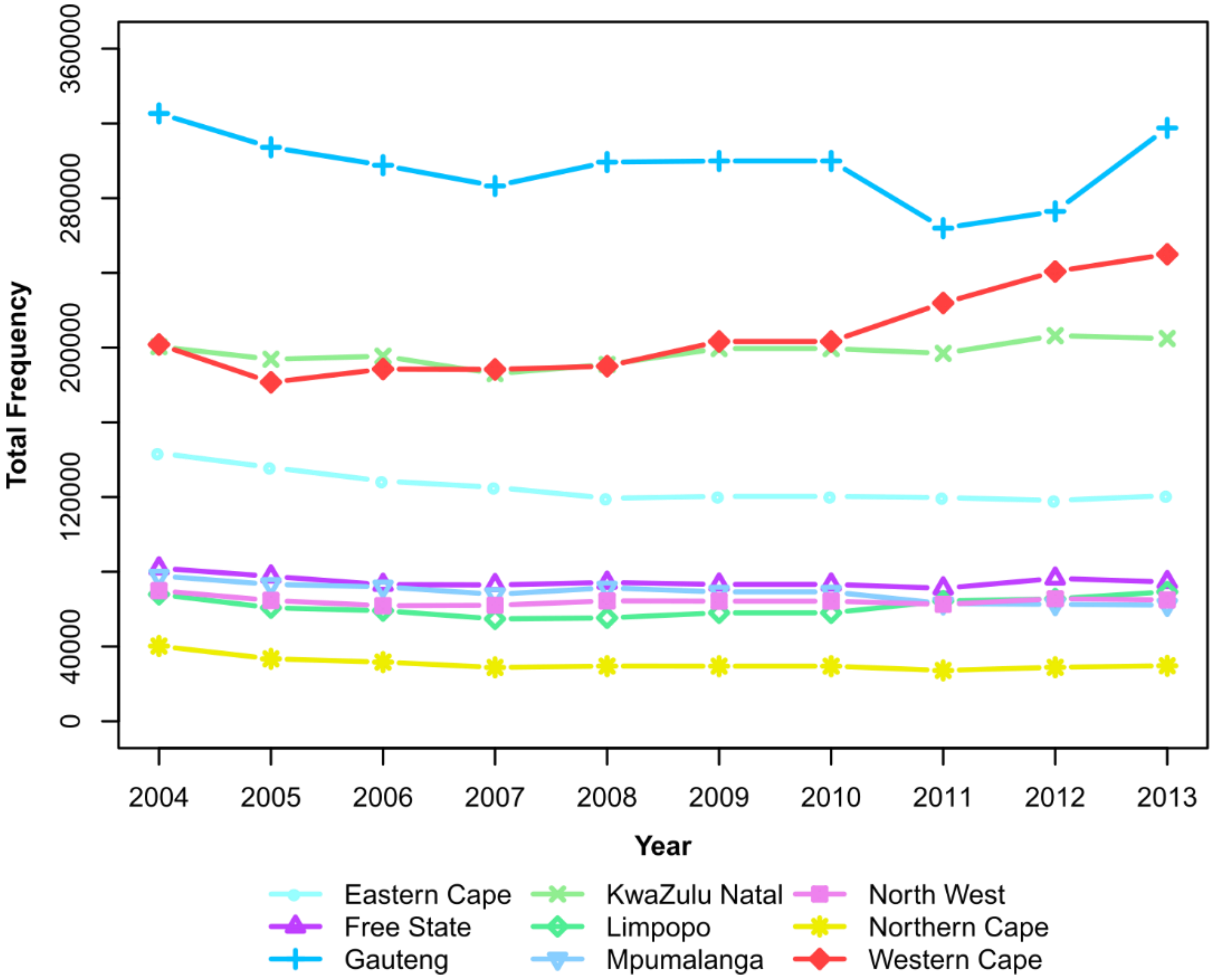

***Figure 4-1**: Total violent crime frequency per province 2004-2013.*

Table 4-1*: Mid-year population estimates (in millions) adapted from www.easydata.co.za.*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
GT	10.71	10.90	11.11	11.32	11.53	11.75	11.97	12.21	12.45	12.69
KZ	9.27	9.39	9.51	9.64	9.78	9.92	10.07	10.22	10.38	10.55
EC	6.06	6.12	6.18	6.24	6.31	6.38	6.46	6.54	6.62	6.72
WC	5.16	5.25	5.33	5.42	5.51	5.61	5.70	5.80	5.89	5.99
LP	4.92	4.98	5.04	5.10	5.16	5.23	5.31	5.38	5.46	5.55
MP	3.63	3.68	3.73	3.79	3.84	3.90	3.96	4.02	4.09	4.15
NW	3.16	3.21	3.25	3.30	3.35	3.40	3.45	3.50	3.55	3.60
FS	2.65	2.66	2.67	2.68	2.70	2.71	2.73	2.74	2.76	2.78
NC	1.08	1.09	1.10	1.10	1.11	1.12	1.13	1.14	1.15	1.16

The calculated per capita crime rates are displayed in Figure 4-2. There appears to be a considerable difference between the displays of Figure 4-1 and that of Figure 4-2, first of which is the respective ordering of the provinces. The Western Cape is now clearly positioned at the top, followed by the Northern Cape and the Free State. Notably, the Northern Cape has migrated from the bottom of the graph to near the top. Gauteng has fallen and so has KwaZulu-Natal, both of which were previously at the top of Figure 4-1. Interestingly, the slopes of the respective province lines have remained fairly similar to that of Figure 4-1.

It is evident that the change in crime rates per capita follow similar trends to that of the total crime frequencies for the respective provinces. All provinces, barring the Western Cape and Limpopo, are positioned fairly close together, however two smaller groupings do exist. These are the grouping of the Northern Cape, Free State, and Gauteng, as well as the grouping of KwaZulu-Natal, the North West Province, the Eastern Cape, and Mpumalanga. It is intriguing to note that these groupings appears to dissipate over the years and by 2013 there is a much more uniform pattern to the spacing between provinces. The Western Cape is again the exception to this trend, with the distance between itself and the other provinces increasing steadily. Barring the Western Cape, these patterns suggests a more even distribution of the crime rates per capita for the South African provinces by 2013. It then follows that there is a considerable difference between analysing the total frequencies and the relative frequencies for the South African crime data. In order to provide an accurate analysis for such data, techniques which account for both the total frequencies and relative frequencies should be utilized in the study.

There exists a single constant between Figure 4-1 and Figure 4-2, in that the directions and movements of the individual province lines are very similar between the two figures, despite their relative positions changing. The analysis of the trends in Figure 4-1 then carry forward to that of Figure 4-2. As previously mentioned, South Africa has been experiencing consistent decreases in crime over the last decade. However, in Figure 4-1 it can easily be seen that both Gauteng and the Western Cape have experienced considerable increases in their total crime frequencies in more recent years. Limpopo and KwaZulu-Natal have experienced a marginal increase over the last ten years and the Eastern Cape experienced a decrease in earlier years.

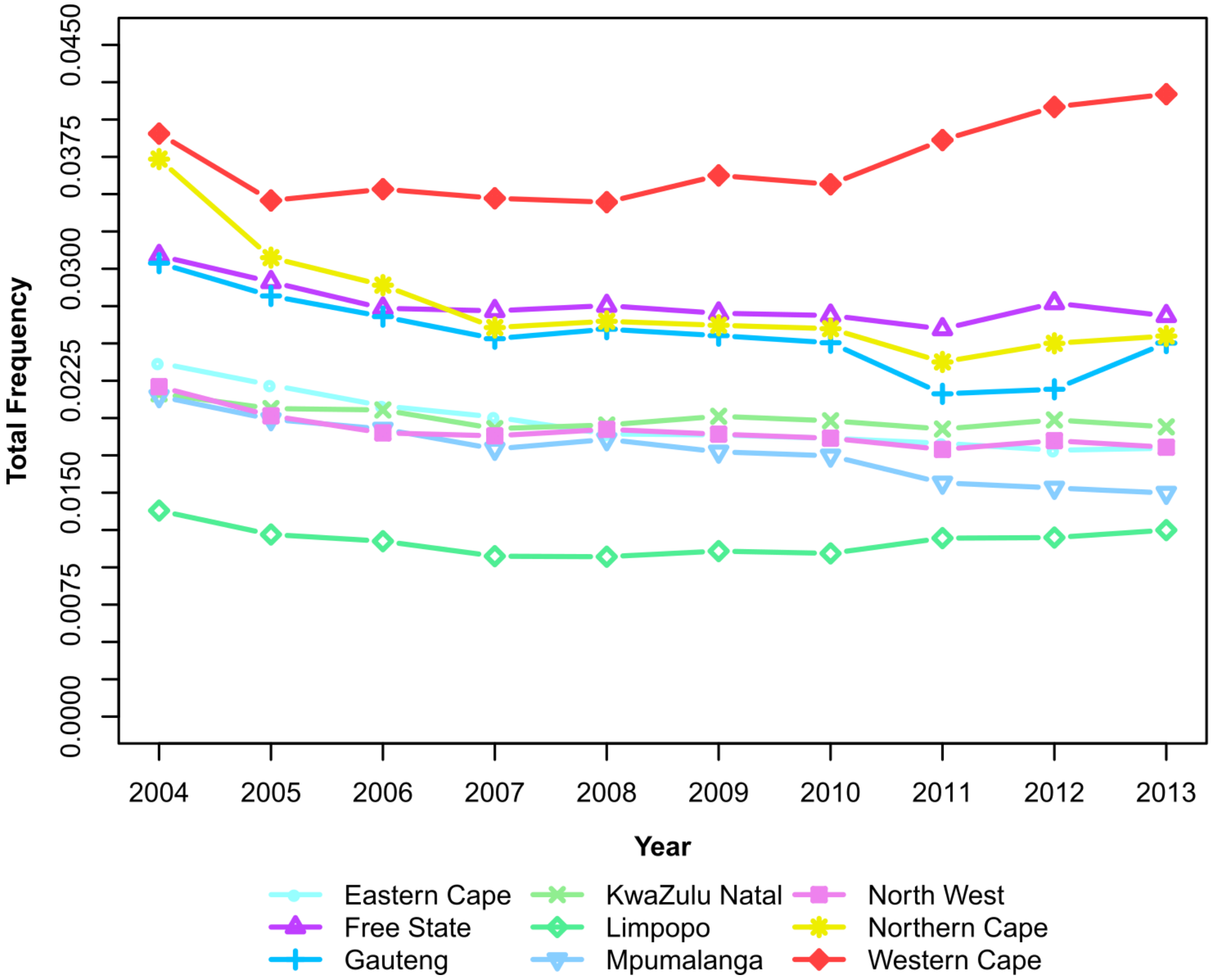

***Figure 4-2**: Violent crime rates per capita by province 2004-2013.*

In Figure 4-1 it appears that all the provinces positioned at the bottom of the graph have not experienced considerable changes over last ten years. The exceptions being Mpumalanga and Limpopo which experienced a decrease and an increase of reported crimes in recent years respectively. The variation in the provinces at the bottom of the graph are somewhat concealed by the large scale on the Y-axis. To alleviate the problem in Figure 4-1, Gauteng, KwaZulu-Natal, and the Western Cape have been removed, to produce Figure 4-3.

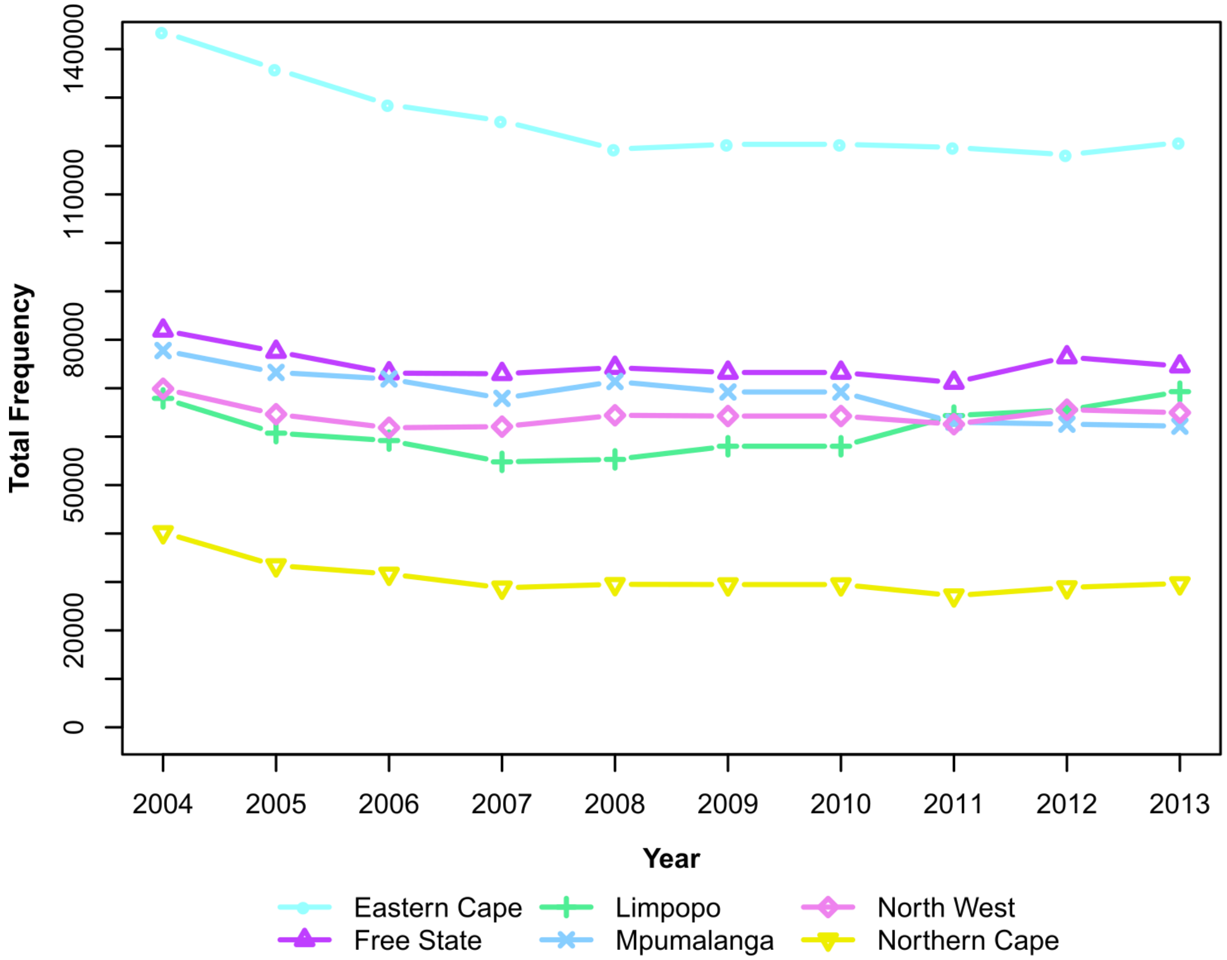

***Figure 4-3**: Violent crime rates per capita by province 2004-2013, with Gauteng, KwaZulu-Natal, and the Western Cape removed.*

In Figure 4-3 trends which were not previously discernible can now be easily detected, most notable of which is the decrease in the total crime frequency in the Eastern Cape. The decrease was evident in Figure 4-1 but the magnitude of the decrease was masked due to the scale of the Y-axis. In addition, the decrease and increase for Mpumalanga and Limpopo respectively is made clearer and more evident in Figure 4-3. All other provinces experience only marginal changes as in Figure 4-1, indicating a minimal amount of variation over the years.

4.1.1 The Composition of the Total Crime Frequency Plot

Having analysed the raw frequencies of the South African provinces, it is then of interest to analyse the decomposition of the total crime frequency. Due to the nature of the available data it is possible to create individual line graphs which depict the composition of the total crime frequencies for the individual provinces. For simplicity, a single figure has been produced which contains all the individual province plots (Figure 4-4). There are no tick marks provided on the Y-axes in Figure 4-4 as the actual frequencies are not of great interest here. Instead, the plots are developed in order to provide insight into the trends and movements of the individual crimes.

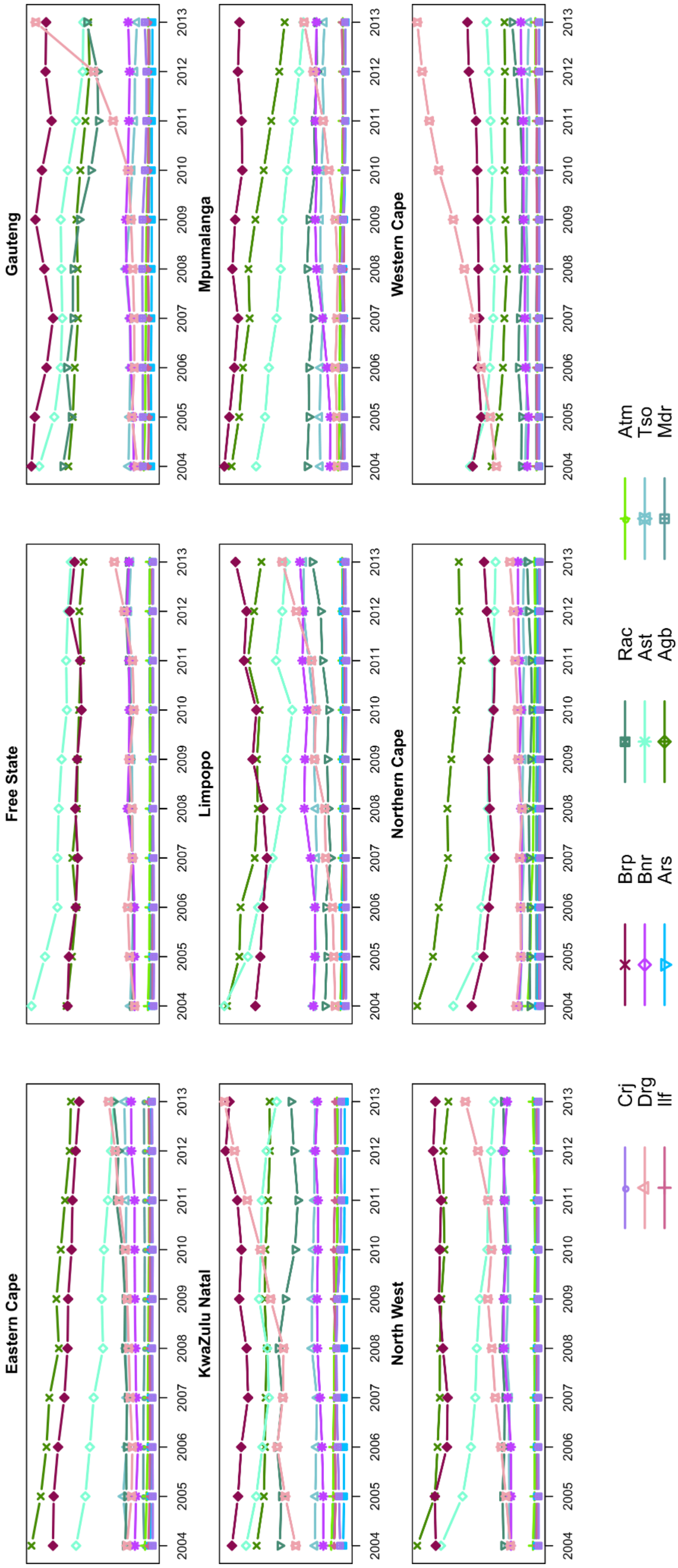

***Figure 4-4**: Trellis chart depicting the compositions to the total crime frequencies for the individual provinces 2004-2013.*

In Figure 4-4 it is apparent that there are some very clear trends over the ten year period. For most provinces there are three to four high frequency crimes, whereas the remaining crimes are grouped together near the bottom of the graphs. Although the majority of the crimes at the lower end of the graphs appear to be marginally increasing, the high frequency crimes appear to have a general downward trend. That is to say that in most of the provinces burglary at residential premises, common assault, and assault with the intent to cause grievous bodily harm have followed a downward trend over the ten year period. The exceptions of which appear to be KwaZulu-Natal and Limpopo where burglary at residential premises appears to have followed a slight positive trend. Additionally, it is of interest to note that in KwaZulu-Natal and Gauteng, robbery with aggravating circumstances is positioned much higher in their respective graphs than in the other provinces. Gauteng and KwaZulu-Natal appear to share a weak association with carjacking, as the respective lines are not located at the very bottom as in other provinces. Robbery with aggravating circumstances also appears to be following an upward trend in the Eastern Cape, Limpopo, and the Western Cape.

The clearest trend is the universal increase in drug-related crime throughout all the provinces. The Western Cape is dissected by the diagonal line, representing drug-related crime, running across its respective graph. KwaZulu-Natal experiences a similar trend, indicating a concerning increase in drug-related crime over the ten year period in both of these provinces. Gauteng experiences a very sharp increase in drug-related crime between 2010 and 2013, where the reported number of cases increased from 16 457 to 74 713, a truly astounding increase. From Figure 4-4, it is abundantly clear that drug-related crime has become a prominent concern throughout South African provinces.

As was highlighted by the UNODC report, South Africa has one of the highest homicide rates in the world. In order to provide some insight about the crime of murder a separate plot is provided. Figure 4-5 below displays the total frequency of murders per province over the ten year period. It appears that in most of the provinces, murder remains fairly unchanged over the years. Gauteng, Limpopo, and the Western Cape experience the largest changes in murder frequencies. However, the aforementioned provinces appear to be converging towards the Eastern Cape in later years.

As previously mentioned most provinces experienced large decreases in common assault, and assault with the intent to cause grievous bodily harm over the ten year period. As both of these crimes can be interpreted as precursors to committing homicide, there would be an expected decrease in homicide following a decrease in either of the aforementioned crimes. Figure 4-5 shows this premise to be false. KwaZulu-Natal, which experiences the largest decrease in murder in Figure 4-5, only experienced marginal decreases in both common assault, and assault with the intent to cause grievous bodily harm. Other provinces which experienced much greater reductions in the aforementioned crimes experienced much smaller decreases in murder or remained unchanged. It then follows that a decrease in either types of assault is not necessarily followed by a decrease in murder.

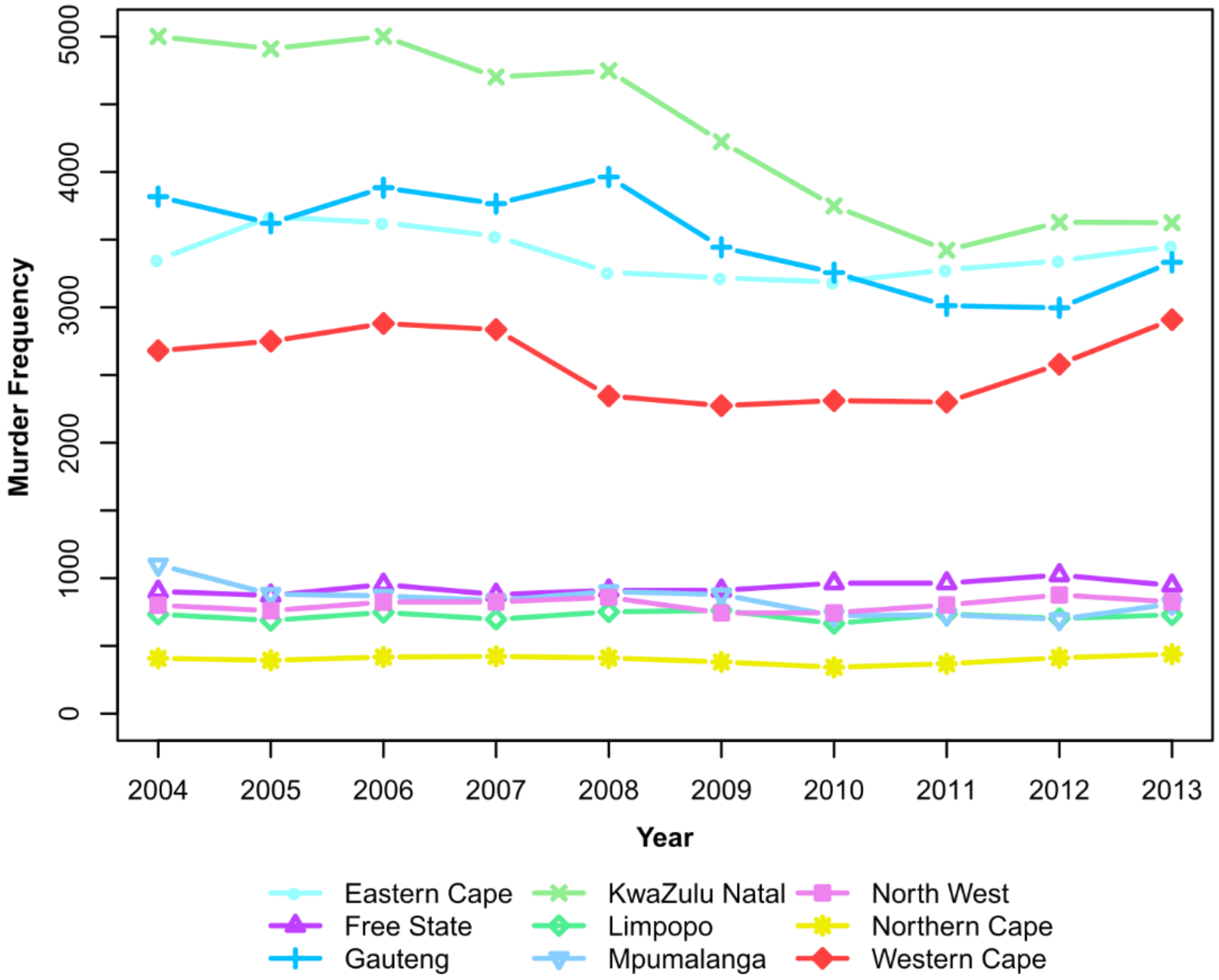

***Figure 4-5**: Total frequency of murder per province for the 2004-2013 period.*

Similarly to Figure 4-1, the progression of the lower frequency provinces is somewhat concealed by the scale of the Y-axis in Figure 4-5. To remedy this, Figure 4-5 has been reproduced as Figure 4-6 which excludes the Eastern Cape, Gauteng, KwaZulu-Natal, and the Western Cape. Immediately the variation in the province lines is much clearer. There remains a very horizontal pattern to the provinces in Figure 4-6, however, a slight downward trend for Mpumalanga is now detectable. The Free State experiences a marginal increase in murder despite experiencing a small decrease in both types of assault. All other provinces experience downward trends in both types of assault which are accompanied by relatively constant murder rates.

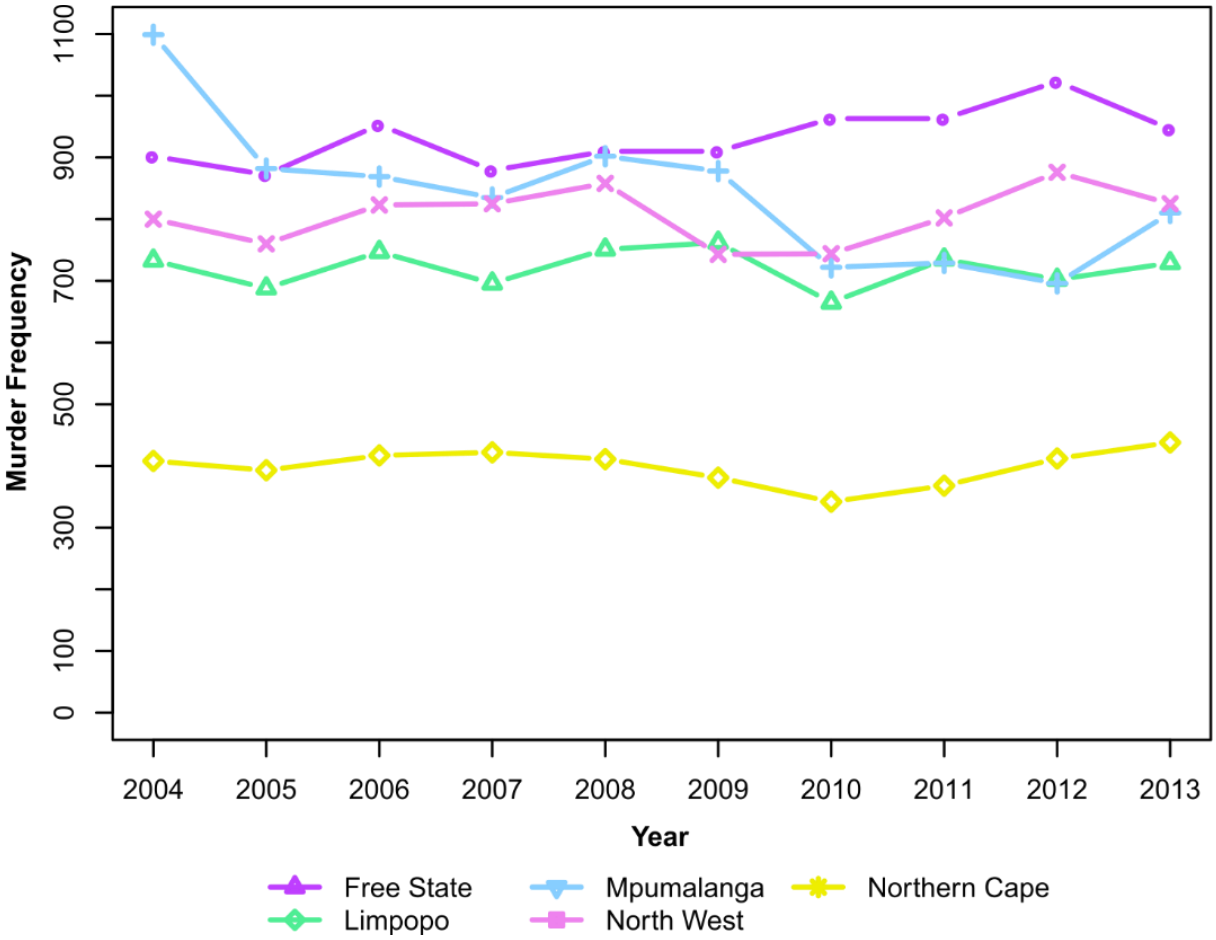

Figure 4-6*: Total frequency of murder per province for the 2004-2013 period, with Eastern Cape, Gauteng, KwaZulu-Natal, and the Western Cape removed.*

A general idea of the behaviours of the South African crime data has now been established. From Figure 4-1 and Figure 4-2 it was determined that it proves beneficial to not only analyse the absolute frequencies but the relative frequencies as well. Following on from this, it may then prove beneficial to view the proportional crime figures instead. In the following section bar plots are utilized to display the absolute and proportional crime frequencies for the individual provinces.

4.2 Bar Plots

The second univariate method to be utilized in this study is the bar plot. Unlike the previous methods, the bar plot focuses less on the trends in the data and more so on the composition of the total frequency. The following bar plots are produced with the aim of providing insight into the composition of the total crime frequencies of the individual provinces.

4.2.1 Stacked Bar Plots

The total crime frequencies of the 2004 South African crime data are displayed in the stacked bar plot below (Figure 4-7). In Figure 4-7, each bar represents the total crime frequencies of the individual provinces for 2004. In addition to the total frequencies being displayed, the compositions of the total frequencies are also displayed in the plot. The individual bars are comprised of stacked segments, each of which displays the contribution of the respective crimes for each province’s total frequency. The individual crimes frequencies are stacked upon one another to display their respective contributions to the total crime frequency for that province, hence the *stacked* bar plot.

When exclusively looking at the sizes of the individual bars Gauteng, KwaZulu-Natal, and the Western Cape are highlighted as having the largest number of total crime frequencies. Conversely, the Northern Cape appears to have the lowest total crime frequencies. The visual representation of the total crime frequencies is a shared characteristic between the line graphs and bar plots. However, the bar plots provide a much clearer representation of the total frequency composition. Despite this, information is lost due to the fact that there is no longer a progression over time. In order to display trends in the compositions over time, ten individual bar plots would be required. Figure 4-8 displays the stacked bar plot of crime frequencies over the ten year period 2004-2013.

Although trends are difficult to detect in Figure 4-8, it does display a different aspect of the data not easily seen before. The composition of the total crime frequency per province are now displayed for the reporting period. It is now possible to see which crimes contribute the most to each province’s total crime frequency over the ten year period. In 2004, burglary at residential premises appear to be the largest contributor to the total crime frequency in the majority of the provinces, followed by assault with the intent to cause grievous bodily harm, and common assault. Similar results are visible in Figure 4-4.

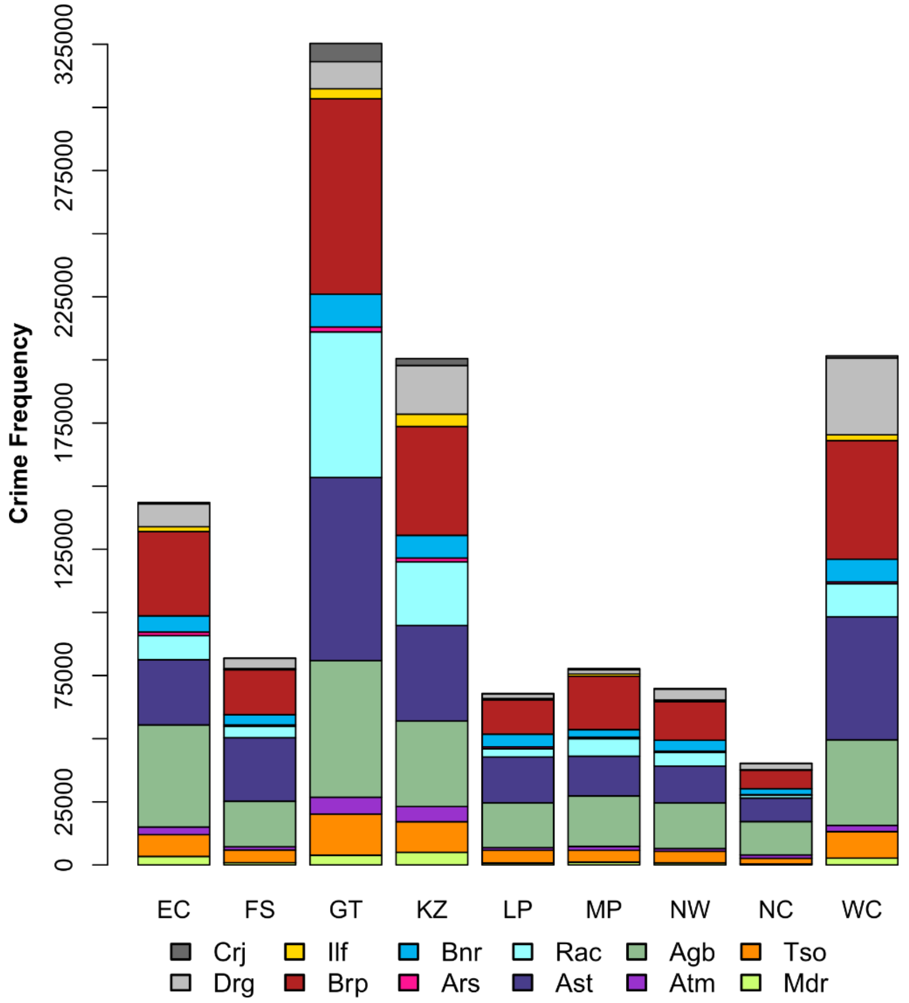

***Figure 4-7**: Stacked bar plot of total crime frequencies per province for 2004-2005.*

Over the ten year period Gauteng consistently remains the leader with regards to the total crime frequency. In 2004 both KwaZulu-Natal and the Western Cape have a similar total crime frequency and a similar composition to their total frequencies. In later years the Western Cape surpasses KwaZulu-Natal quite substantially. This is most likely due to the considerable increase in drug-related crime in the Western Cape, seen by the ever growing pink segment in the Western Cape stacked bar. This increase in drug-related crime in the Western Cape brings its total crime frequency ever closer to that of Gauteng, despite having a population of approximately half the size. There is a fair amount of movement with regards to Limpopo, Mpumalanga, and the North West Province. In earlier years Mpumalanga had a marginally higher total crime frequency than the remaining two provinces. This continued up until 2011, where, due to a decrease and an increase in Mpumalanga’s and Limpopo’s respective total crime frequencies, their positions changed. Mpumalanga is now the lowest of the three, and Limpopo being the highest.

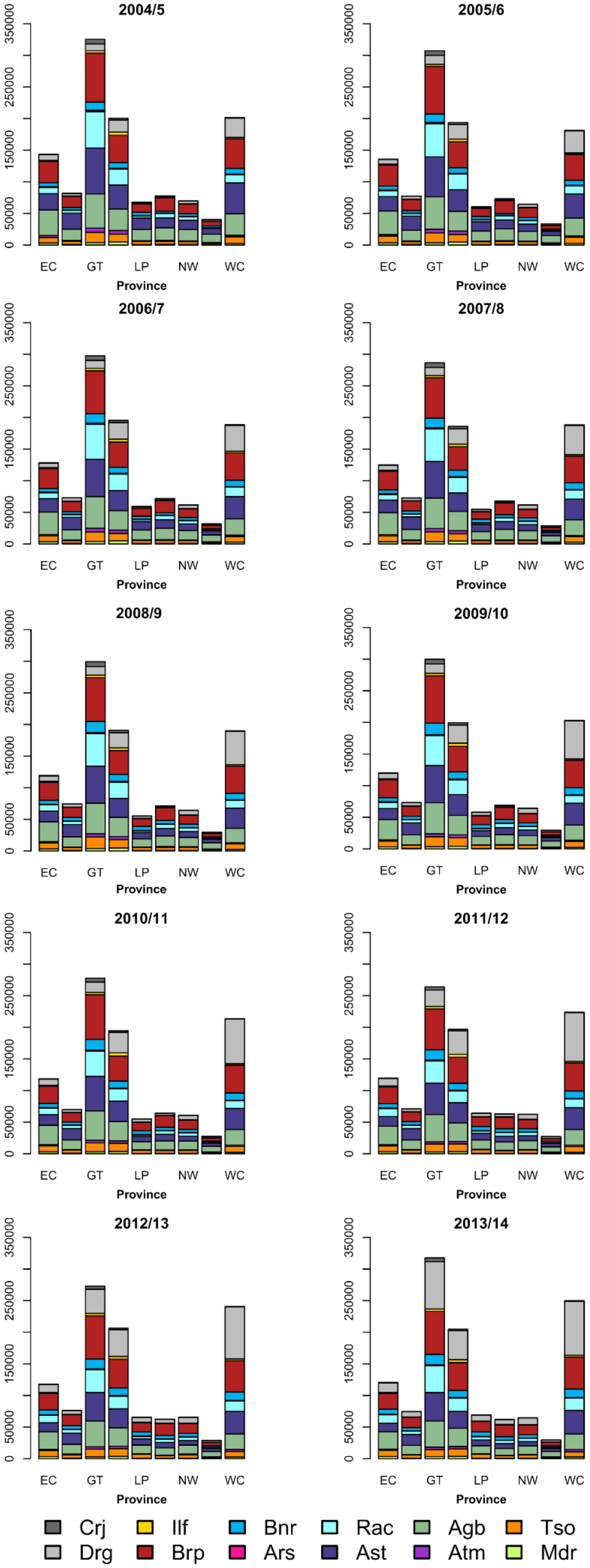

***Figure 4-8**: Trellis chart displaying the stacked bar plot of total crime frequencies per province for 2004-2013 period.*

4.2.2 Proportional Contribution Bar Plots

As seen in Section 4.1, it proved beneficial to view relative frequencies rather than just absolute frequencies. Focus now shifts to proportional compositions of the total crime frequency.

Although Figure 4-7 and Figure 4-8 display a new aspect of the data, the problem of comparing absolute frequencies still exists. Figure 4-7 is difficult to interpret as each province has a different total crime frequency and thus direct comparisons between the bars is difficult. A logical progression is then to analyse the proportional contributions of the individual crimes to the total crime frequency per province. In order to investigate if provinces have similar proportional contributions to their total crime frequencies, each crime frequency for each province is divided by its respective total crime frequency. Figure 4-9 is an example of such a graph when applied to the Eastern Cape over the 2004-2013 period.

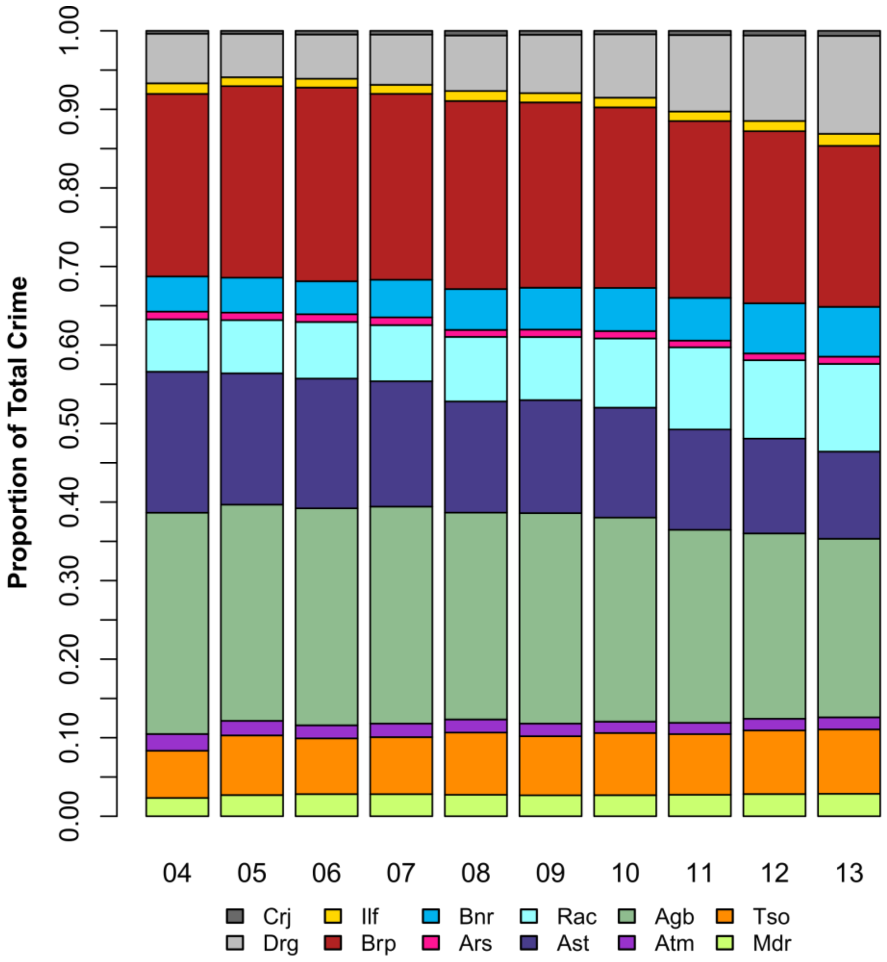

***Figure 4-9**: Proportional contribution stacked bar plot for the Eastern Cape total crime frequency for the 2004-2013 period.*

Each bar in Figure 4-9 is, similarly to before, made up of a number of component parts which are stacked upon one another. The colours represent the same crimes as in Figure 4-7, but the size of each portion of the bars represents each individual crime's proportional contribution to the total crime frequency. Figure 4-9 clearly represents which crimes contribute most to the total crime frequency in the Eastern Cape over the past decade.

Burglary at residential premises and assault with the intent to cause grievous bodily harm are clearly the largest contributors to the total frequencies. Arson and carjacking are barely visible in the graph, representing a minimal contribution by these two crimes. Additionally, Figure 4-9 provides information on the variations in the contributions over time. There is a diminishing trend in the crimes of common assault, burglary at residential premises, and assault with the intent to cause grievous bodily harm. This trend, more noticeable in common assault, represents a decrease in the proportional contribution to the total crime frequency over the years. However, this does not necessarily mean that the number of reported cases of these three crimes has decreased, rather that only its contribution to the total crime frequency has decreased. A large increase or decrease in the frequency of one crime could result in a decreased or increased proportion for a crime of which the frequency remained unchanged.

Accompanying the aforementioned decreasing trends is an increase in drug-related crimes and robbery with aggravating circumstances. Furthermore, there appears to be a weaker, but still relevant, increase in robbery at non-residential premises. Although there is a change in the contributions to the total crime frequency in the Eastern Cape, the bars appear fairly consistent over time with only marginal changes, suggesting that the Eastern Cape’s profile has not experienced considerable variation over the years.

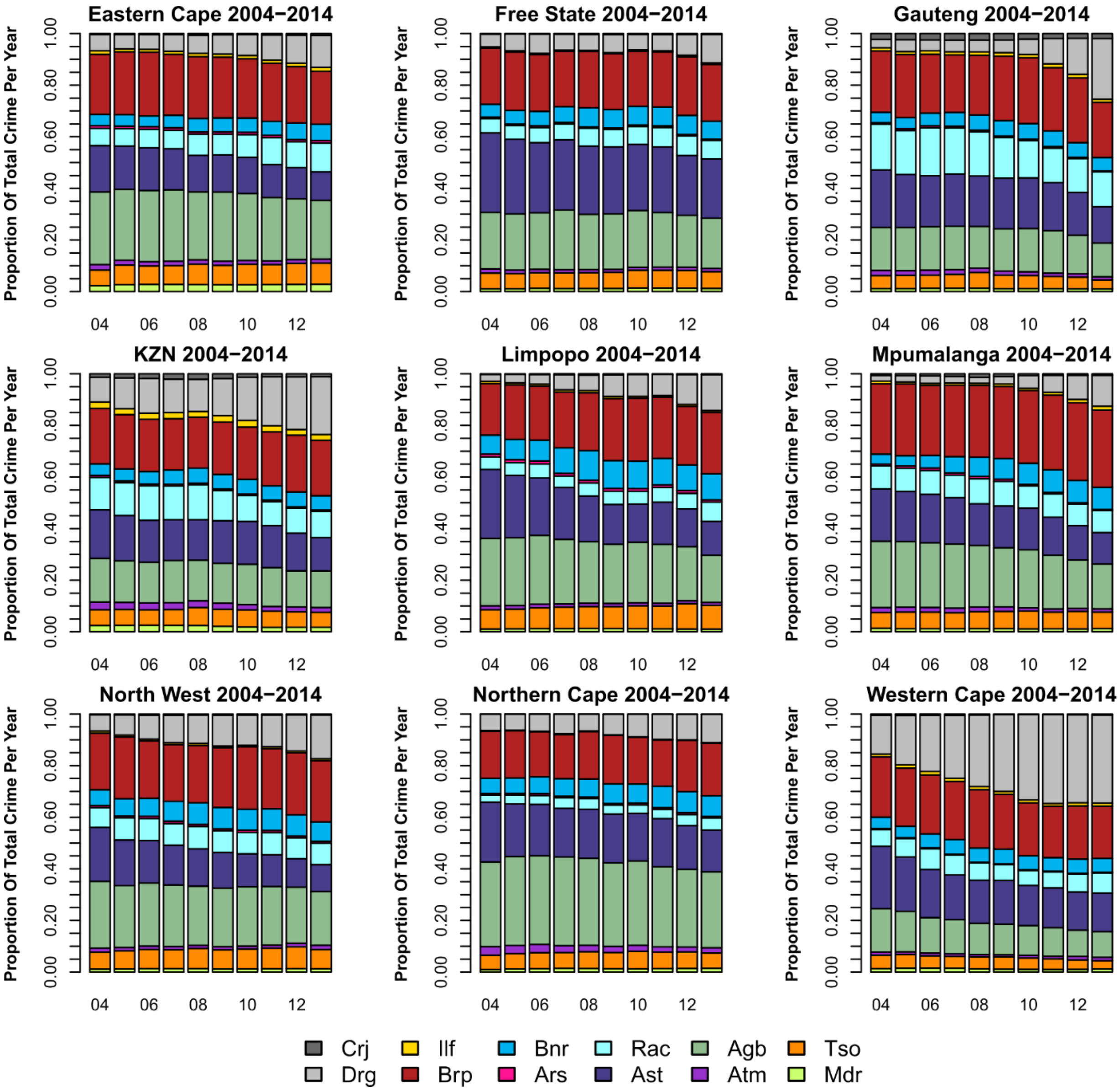

***Figure 4-10**: Trellis chart displaying the proportional contribution stacked bar plot for all provinces for the 2004-2013 period.*

Continuing the trend set in Figure 4-7, the stacked bar plots for all 9 provinces are presented simultaneously in Figure 4-10. At a glance the provinces which experience large changes in their crime composition can easily be detected. The Eastern Cape and the Free State appear to be fairly consistent over the years, as opposed to the Western Cape which experiences the largest variation. Although the magnitude of the variation differs between provinces, the crimes for which proportions increase or decrease are generally the same throughout the provinces. Regrettably, it is still difficult to view the changes of the smaller contributions in the graph. When Figure 4-10 is viewed in conjunction with Figure 4-4, these small changes are more noticeable.

The most visible trends seen in Figure 4-10 coincide with those seen in Figure 4-4. Carjacking is faintly visible in most of the bar plots in Figure 4-10 other than for Gauteng and KwaZulu-Natal, suggesting that in these two provinces carjacking has a larger proportional contribution than in the other provinces. There is a consistent decrease for all provinces in the proportion of burglary at residential premises, common assault, and assault with the intent to cause grievous bodily harm. Burglary at non-residential premises only appears to have increased in Mpumalanga, Limpopo, KwaZulu-Natal and the Eastern Cape, most noticeably in Mpumalanga. It was previously mentioned that in the line graphs, robbery with aggravating circumstances was located much higher for KwaZulu-Natal and Gauteng. This trait is evident in Figure 4-10, as both of these provinces have much larger segments representing this crime type. Similarly, the Northern Cape and the Eastern Cape clearly have the largest proportions for assault with the intent to cause grievous bodily harm. A fact which is clearly seen in Figure 4-4, as assault with the intent to cause grievous bodily harm is located at the top of the line graphs for both these provinces.

As was the case in the line graphs, the development which is most prominent is the increase in drug-related crime throughout all of the provinces, specifically in the Western Cape. Gauteng, KwaZulu-Natal and the North West also experience pronounced increases. Although Gauteng has around 13% less reported drug-related crimes in 2013 than the Western Cape, their difference in Figure 4-10 appears much greater. This is due to the fact that although in frequency they may not differ by a substantial amount, drug-related crimes make up a much larger proportion of the total crime frequency in the Western Cape than in Gauteng.

The majority of the changes visible in Figure 4-10 are of a direct response to increases and decreases visible in Figure 4-4. It follows that increases or decreases in proportional contributions of crimes are mostly due to the corresponding changes in frequencies for that specific crime. The example of Gauteng’s drug-related crime proportion above shows how some increases can be masked by other factors. Although in this particular application the proportional and absolute frequency plots drew similar conclusions, the motivation for utilizing both types of displays is clearly conveyed.

4.2.3 The Mosaic Display

An alternative method of displaying contingency tables is the mosaic plot. Although it does not fall under the category of commonly used methods in the study of crime, it warrants mentioning in the univariate analysis as the mosaic plot has the ability to not only handle two-way tables, but multi-way tables as well. The mosaic plot was introduced by Hartigan & Kleiner in 1981, and can be interpreted as an area-proportional visualization of frequencies (Hartigan & Kleiner, 1981). The mosaic plot consists of tiles (corresponding to the cells of the contingency table) created by recursively splitting the rectangle vertically and horizontally (Chen *et al.*, 2008:590). Each tile’s area is proportional to the size of the cell, and its shape and location are determined during the construction process.

The construction of a p dimensional mosaic plot can be explained in a short recursive algorithm. Let $\boldsymbol{X}$ be a contingency table with p categorical variables $X_1, \dots, X_p$. Let c_i be the number of categories of variable X_i, $i = 1, \dots, p$. The construction of the mosaic plot of $\boldsymbol{X}$ is as follows (Chen *et al.*, 2008:619):

1. The construction begins with one single rectangle r_0 of width w_0 and height h_0, and set $i = 1$.
2. Dissect rectangle r_{i-1} into c_i pieces: find all the observations corresponding to rectangle r_{i-1}, and obtain the compositions for each variable X_i (the frequencies of each category within variable X_i). Split the width (heights) of rectangle r_{i-1} into c_i pieces, where the widths (heights) are proportional to the compositions, and keep the heights (widths) of each the same as r_{i-1}. Let the new rectangles be defined as r_i^j, where $j = 1, \dots, c_i$.
3. Increase i by 1.
4. Repeat steps 2 and 3 for all r_{i-1}^j with $j = 1, \dots, c_{i-1}$, while $i \leq p$.

It is evident from the above algorithm that the hierarchical construction of mosaics places great emphasis on the order of the variables in the plot. In the case of multi-way contingency tables, different variable orders in mosaics emphasize different aspects of the data. However, the order of which the variables must appear is not immediately apparent for multi-way contingency tables (Chen *et al.*, 2008:620). The ordering of the variables is not of great importance in the study of the South African crime data, as the tables are presented in the form of two-way contingency tables. However, it is possible to manipulate two or more of the South African crime tables into one single multi-way table. The function `mosaicplot()` from the VCD package (https://cran.r-project.org/web/packages/vcd/index.html) is capable of constructing mosaic plots of p-way contingency tables (Meyer, Zeileis & Hornik, 2015). The `mosaicplot()` function has been used to produce a mosaic plot of the 2004 South African crime data, presented in Figure 4-11 below.

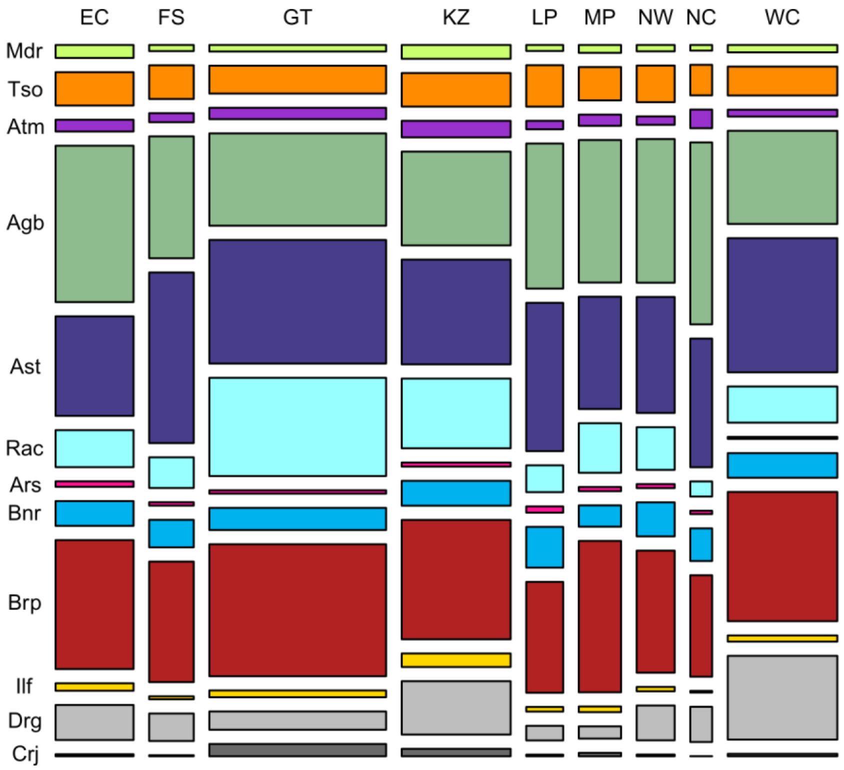

***Figure 4-11**: Mosaic plot of the 2004 South African crime data.*

Figure 4-11 is constructed from a two-way table, and as such, the resulting mosaic plot closely resembles the proportional contribution bar plot of Section 4.2.2. This is due to the fact that the original rectangle is split vertically, proportional to the total frequencies in each province, and then the nine columns are segmented proportionally to their respective counts in each crime category. However, the mosaic plot provides one additional feature over the proportional contribution bar plot, and that is the column widths are proportional to the total crime frequencies per province, providing a reference to the total crime counts in each province.

In order to demonstrate additional features of the mosaic plot, the 2004 and 2013 South African crime tables have been manipulated into a single three-way contingency table. This is achieved by the following R code:

```
CrimeThreeWayTable<-
array(c(as.vector(as.matrix(Crime.2004)),as.vector(as.matrix(Crime.2013))
),dim=c(12,9,2)).
```

The 2004 and 2013 South African crime data is now in the form of a three-way array. Figure 4-12 is a three-way mosaic plot of the aforementioned three-way table. As previously mentioned, there are multiple ways to order the variables for multi-way tables, and the resulting mosaic plot is heavily dependent on said ordering. For the construction of Figure 4-12, the variable ordering was set as: *Year, Province,* and *Crime.* At first glance Figure 4-12 appears to be two separate mosaic plots positioned next to one another. However, this is not the case. Due to the ordering of the variables, the very first dissection is with respect to the *Year* variable. This split is proportional to the total count in each of the crime tables. As the total counts of each table are approximately equal (1 208 769 and 1 193 752), both segments appear to be of approximately equal size. The 2004 and 2013 squares are then subsequently dissected horizontally such that the width of each province rectangle is proportional to the total count for the respective provinces in each respective year. Lastly, the province rectangles are dissected vertically with respect to the individual crime counts for the respective provinces in the respective years. Such ordering of the variables produces what appears to be two individual mosaic plots of the 2004 and 2013 South African crime data side by side.

Two distinct features of the South African crime data are evident in Figure 4-12, that is the increase in the total number of reported crimes in the Western Cape (seen by increase in the width of the Western Cape rectangle from 2004-2013), and the large increase in the proportion of total crimes made up by drug-related crimes (seen by the increase in the size of grey segments between 2004-2013). These results were evident from the results of the figures from earlier sections. The improvement here is that the results are now visible in a single figure.

The mosaic plot has been discussed here due to its ability to graphically represent multi-way data. The mosaic plot provides and improvement upon the proportional contribution bar plot by providing an indication of the proportion of crime per province, and proves to have a relatively simple construction and interpretation. However, by scaling the widths of the columns in the proportional contribution bar plots relative to the total crime per province or year, the same information can be accounted for. Additionally, due to the nature of the South African crime data, the three-way mosaic plot does not prove to be great improvement, and can be viewed as a side-by-side figure of individual two-way mosaic plots. As such, the mosaic plot is not utilised further in this study.

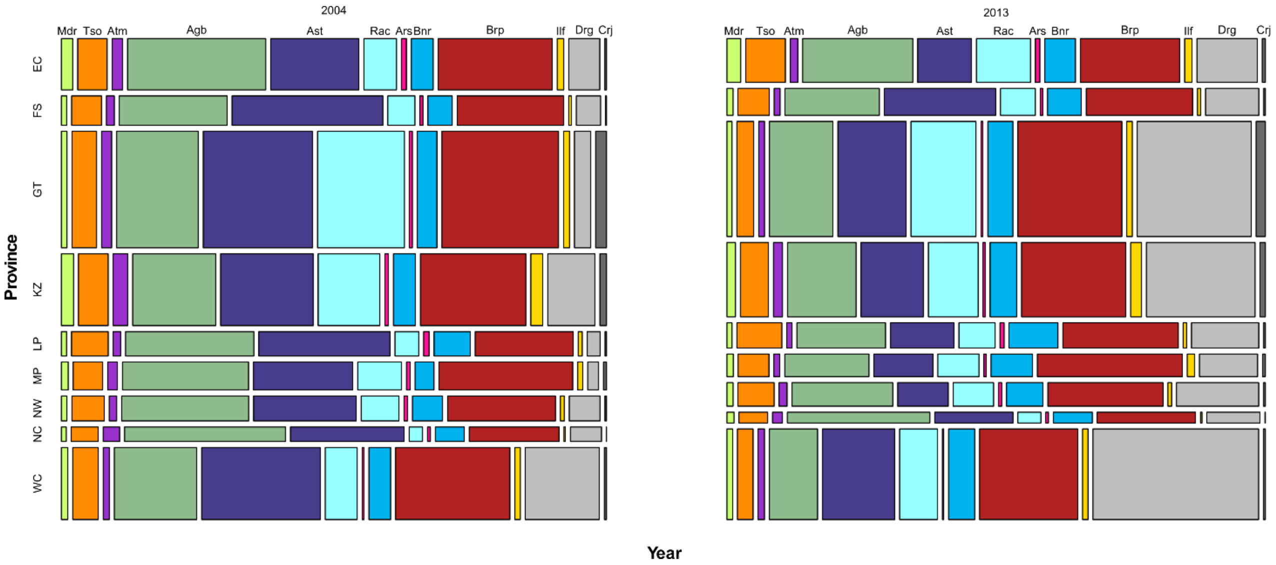

***Figure 4-12**: Three-way Mosaic plot of the 2004 & 2013 South African crime data.*

4.3 Summary

The analysis above progressed from line graphs to the proportional contribution bar plots and mosaic plots. The line graphs proved cumbersome to analyse and it is clearly conveyed that a large number of figures is required with such a technique to produce a worthwhile analysis. Conversely, the proportional contribution bar plot proved to be very useful in describing the data. Additionally, the mosaic plot provided similar information to that of the proportional contribution bar plots, with slight improvements. From the proportional contribution bar plot it was concluded that the provinces, in general, appeared to have very similar compositions of their total crime frequency. Gauteng, KwaZulu-Natal, and the Western Cape are highlighted as they tend to deviate away from the other provinces. It was also noted that drug-related crime has become a major contributor to the total crime frequencies over the ten year period. In order to further investigate the differences and patterns in the provinces' compositions, the analysis shifts to correspondence analysis and biplots in Chapter 6. Chapter 5 provides a theoretical overview to the GDA methods utilized in Chapter 6.

Chapter 5: A Theoretical Overview of GDA Methods

Correspondence analysis and biplots are graphical techniques which are utilised in order to view and interpret multidimensional data in low-dimensional sub-spaces. Such techniques can be broadly grouped under the heading of multidimensional scaling. Multidimensional scaling, commonly referred to as MDS, is defined by Cox & Cox as:

"the search for a low dimensional space, usually Euclidean, in which points in the space represent the objects, one point representing one object, and such that the distances between the points in the space, match, as well as possible, the original dissimilarities." (Cox & Cox, 2001:1)

Similarly, MDS is the search for a low-dimensional sub-space which best approximates the dissimilarities between points as distances in the new sub-space. As it is infeasible to display data in more than three dimensions, there is a need to approximate higher dimensions in a low-dimensional sub-space in order to simultaneously display more than three variables in a single plot.

The methodology of such techniques is explored in this chapter.

5.1 Important Matrix Results

Concepts such as the spectral decomposition and the singular value decomposition of a matrix play an invaluable role in both the correspondence analysis and biplot algorithms. The following results are used repeatedly throughout the study and therefore warrant further explanation.

5.1.1 The Spectral Decomposition

Let $\boldsymbol{X}$ be an $n \times n$ symmetric matrix with eigenvalues $\{\sigma_i\}$ and eigenvectors $\{\boldsymbol{v}_i\}$, then $\boldsymbol{X}$ can be decomposed as:

$$\boldsymbol{X} = \boldsymbol{V\Sigma V}^T = \sum_{i=1}^{n} \sigma_i \boldsymbol{v}_i \boldsymbol{v}_i^T, \qquad (5\text{-}1)$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$, and $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_n]$ is an orthogonal matrix. If $\boldsymbol{X}$ is nonsingular (has an inverse), then the spectral decomposition may be written as:

$$\boldsymbol{X}^m = \boldsymbol{V\Sigma}^m \boldsymbol{V}^T, \qquad (5\text{-}2)$$

with $\boldsymbol{\Sigma}^m = diag(\sigma_1^m, \dots, \sigma_n^m)$ for any integer m. (Cox & Cox, 2000:26)

5.1.2 The Singular Value Decomposition (SVD)

If $\boldsymbol{X}$ is an $n \times p$ matrix of rank k, then $\boldsymbol{X}$ can de decomposed as:

$$\boldsymbol{X} = \boldsymbol{U\Sigma V}^T = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T. \qquad (5\text{-}3)$$

$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_k)$ is a diagonal matrix comprising of the singular values of $\boldsymbol{X}$, with $\sigma_1 \geq \sigma_2 \geq, \dots, \geq \sigma_k > 0$. Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are of size $n \times k$ and $p \times k$ respectively.The k columns of $\boldsymbol{U}$ are the normalized eigenvectors of $\boldsymbol{XX}'$ corresponding to the eigenvalues $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$. The k columns of $\boldsymbol{V}$

are the normalized eigenvectors of $\boldsymbol{X'X}$ corresponding to the eigenvalues $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$. As the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are normalized eigenvectors of the symmetric matrices, they are mutually orthogonal and thus $\boldsymbol{U^T U} = \boldsymbol{V^T V} = \boldsymbol{I}$. (Cox & Cox, 2001:26; Rencher, 2002:36)

5.2 Correspondence Analysis

Correspondence analysis, a subset of MDS which is commonly referred to as CA, is a method for graphically representing tabular data (Greenacre, 2007:1). CA provides a useful method for defining intervals between categories of variables. The distances between the categories and their ordering have substantive meaning in a CA map (Greenacre, 2007:3). However, CA is not a tool for testing statistical significance but rather a graphical method which allows researchers to see patterns of association in the data itself, generating hypotheses to be tested at a later stage (Greenacre, 2007:7).

Profiles and profile spaces are at the heart of correspondence analysis theory. A *profile* is a vector of relative frequencies. Due to the fact that CA is based on two-way tables, it is possible to have two different sets of relative frequencies, one set for rows and another set for columns. Table 5-1 below represents the cross-tabulation of types of violent crimes occurring in three South African provinces. The table itself is a sub-table extracted from the 2013/2014 South African crime data used in this study. The column headings (from left to right) represent the Eastern Cape, Free State, and Gauteng respectively. Similarly, the row headings represent (from top to bottom) murder, total sexual offences, attempted murder, assault with the intent to cause grievous bodily harm, and common assault.

The row profiles of a two-way table are obtained by dividing each element of the row by its respective row total. For example, the first row profile of Table 5-1 is calculated as:

$$\begin{bmatrix} \frac{3\,453}{7\,732} & \frac{946}{7\,732} & \frac{3\,333}{7\,732} \end{bmatrix} = [0.447 \quad 0.122 \quad 0.431].$$

The profiles of CA mimic that of compositional data, as the profile elements sum to unity. Compositional data analysis is explored further in Chapter 7. The “Row Masses”, which play an important role in subsequent calculations, are calculated as the individual row totals divided by the total summation of all frequencies of the table. The average row profile is then calculated by dividing the column totals by the grand total. The respective row profiles of Table 5-1 are provided for below in Table 5-2. Similarly, the column profiles are obtained by interchanging the role of rows and columns in the above description, so that the average row profile contains the column masses associated with the respective columns (see Table 5-3).

***Table 5-1**: A sub-table of the South African crime data for 2013/2014.*

Crime	**EC**	**FS**	**GT**	***Total***
Mdr	3 453	946	3 333	7 732
Tso	9 897	4 814	11 021	25 732
Atm	1 858	911	3 901	6 670
Agb	27 451	14 531	41 581	83 563
Ast	13 392	17 124	44 748	75 264
Total	56 051	38 326	104 584	198 961

Table 5-3 depicts the column profiles of Table 5-1. An important property of correspondence analysis is evident in both Table 5-2 and Table 5-3. Note that in both tables the final row and column values stay constant from Table 5-2 to Table 5-3. This consistency materialises as the symmetric relationship between the rows and columns of the table in Correspondence Analysis. The choice to argue via the row or column profiles has no material impact on the outcomes of the analysis, hence the symmetry. Only the interpretation of the findings will differ depending on the choice to argue via the rows or the columns (Greenacre, 2007:11).

***Table 5-2**: Row profiles of Table 5-1.*

Crime	**EC**	**FS**	**GT**	***Row Masses***
Mdr	0.447	0.122	0.431	0.039
Tso	1.280	0.623	1.425	0.129
Atm	0.240	0.118	0.505	0.034
Agb	3.550	1.879	5.378	0.420
Ast	1.732	2.215	5.787	0.378
Average Row Profile	0.282	0.193	0.526	1.000

***Table 5-3**: Column profiles of Table 5-1.*

Crime	**EC**	**FS**	**GT**	***Average Column Profile***
Mdr	0.062	0.025	0.032	0.039
Tso	0.177	0.126	0.105	0.129
Atm	0.033	0.024	0.037	0.034
Agb	0.490	0.379	0.398	0.420
Ast	0.239	0.447	0.428	0.378
Column Masses	0.282	0.193	0.526	1.000

Table 5-1 is a table of size 5 × 3, which can be interpreted as five row profiles which consist of three elements, or as three column profiles which consist of five elements. The required number of dimensions to accurately display such a table is the min(number of columns -1; number of rows -1) of the table. In this case 5 > 3, therefore the required number of dimensions to perfectly represent the data is 2. In cases such as this, where the profiles are made up of three elements, its common practices to use a triangular or *ternary* coordinate system to display the data (Greenacre, 2007:12).

A *ternary* plot is a special case of a triangular plot in two dimensions. Profiles which contain more than three elements can also be displayed in a triangular plot, with the dimensionality being one less the number of elements in a profile. The general name for such systems is *barycentric coordinate system* (Greenacre, 2007:14). For further details on triangular plots see Chapter 2 and Chapter 3 of *Correspondence Analysis in Practice* (Greenacre, 2007). As ternary plots will not be utilised in the study the construction of said plots is not explained, however, an example of which is provided for below.

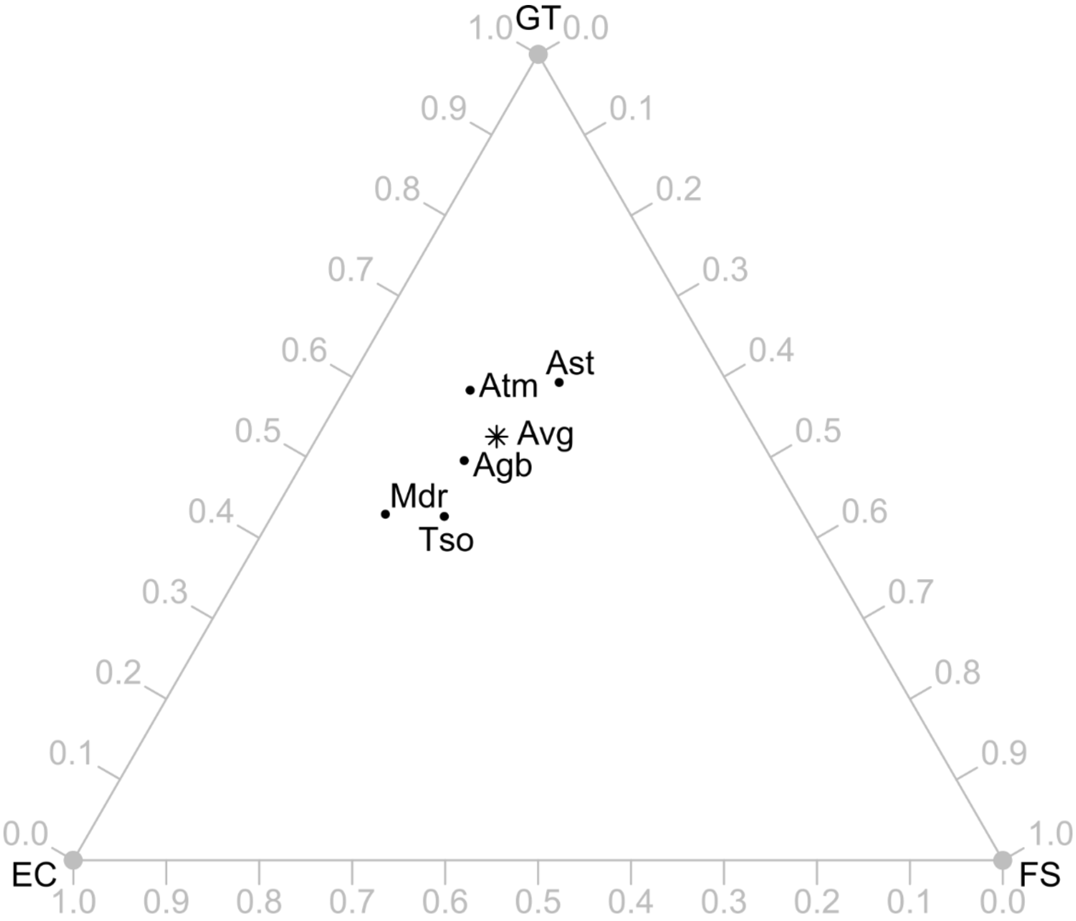

***Figure 5-1**: Ternary plot of the 2013/2014 South African crime data.*

Figure 5-1 consists of three vertices, one for each province, and the five row profiles which are plotted within this space. The average row profile is represented by the star inside the ternary plot. No scaling has taken place for this data, only a suitable system for displaying the data has been found. Potential relationships between types of crimes and provinces can be determined in Figure 5-1 by analysing the distance between them. In Figure 5-1 most of the crimes are positioned in the upper left part of the triangle by the Gauteng vertex, with murder and total sexual offenses migrating slightly towards the Eastern Cape. The positioning of the profile points then suggests that Gauteng has a relatively strong association to common assault, attempted murder and assault with the intent to cause grievous bodily harm. The Eastern Cape appears to be more closely related to murder and total sexual offenses. Alternatively, the Free State appears to be removed from all types of crimes, not suggesting that it does not have a relationship to these crimes, but rather indicates a weaker relationship. Again, the method of CA is to identify possible hypotheses to be tested rather than it being a statistical test itself in which case additional tests are required to accurately determine if relationships do exist.

The above example is that of an ideal scenario. The problem at hand is that very rarely are profiles made up of exactly three elements. The *barycentric coordinate system* has previously been noted as a possible candidate for displaying profiles with elements greater than three. However, for high-dimensional profiles such a coordinate system would be difficult to visualise. Fortunately, CA can accommodate the visualisation of multidimensional profiles.

5.2.1 Symmetric and Asymmetric CA Plots

The words *symmetric* and *asymmetric* appear numerous times in CA literature. The symmetry refers to that of the two-way table. If the table is viewed as symmetric then choosing to argue via the row or column profiles has no material effects on the final conclusions. If the table is viewed as an asymmetric table, then it is viewed as either a set of rows or a set of columns.

An asymmetric CA diagram is a joint display of both the profile and vertex points of a two-way table. In such a display one set of points is scaled in principal coordinates and the other in standard coordinates (Greenacre, 2007:68). In an asymmetric plot the distances between profiles and the average profile can be easily interpreted. If the analysis is focused on the columns, then the column points would be in principal coordinates and the row points in standard coordinates (Greenacre, 2007:68). The opposite holds for a focus on the rows of a two-way table.

In a symmetric CA diagram both sets of profile points, from two different underlying spaces, are overlaid on top of each other. The plot then contains row and column profiles which are from two separate spaces (Greenacre, 2007:70). In this case both sets of points are in principal coordinates. Unfortunately, in a symmetric CA the distances between row and column profiles have no simple interpretation and are best ignored (Greenacre, 2007:72).

5.2.2 Important Calculations

The theory of correspondence analysis is centred about the independence assumption. The independence assumption assumes that the profiles are homogeneous with respect to the categories of a two-way table. If the assumption holds then all of the profiles have similar compositions to that of the average profile, and would only differ due to random sampling fluctuations. If the assumption does not hold, the profiles will be considered heterogeneous with respect to the categories, and therefore oppose the assumption of independence.

The well-known chi-square statistic plays a prominent role in CA. Not only does it provide a measure of association between the rows and columns of a two-way table, but is also used in the calculation of *Inertia*. The chi-squared statistic is computed by calculating the differences between the observed and expected frequencies of the rows or columns of a contingency table under the assumption of independence. Due to the symmetry in CA, arguing via the rows or columns of the table will result in equivalent chi-square statistics and, by association, total inertia values. The general equation for the chi-square statistic is:

$$\chi^2 = \sum \frac{(obeserved - expected)^2}{expected} \qquad (5\text{-}4)$$

(Greenacre, 2007:27). The expected row frequencies of equation 5-4 are calculated by multiplying the row totals by the average row profile. The expected column frequencies are calculated in a similar manner. Depicted below in Table 5-4 are the expected row frequencies of Table 5-1.

***Table 5-4**: Expected row frequencies of Table 5-1.*

Crime	**EC**	**FS**	**GT**	***Total***
Mdr	2 178	1 489	4 064	7 732
Tso	7 249	4 957	13 526	25 732
Atm	1 879	1 285	3 506	6 670
Agb	23 541	16 097	43 925	83 563
Ast	21 203	14 498	39 563	75 264
Total	56 051	38 326	104 584	198 961

The corresponding chi-square calculation to Table 5-4 then follows as:

$$\chi^2 = \frac{(3\,453 - 2\,178)^2}{2\,178} + \frac{(946 - 1\,489)^2}{1\,489} + \cdots + \frac{(44\,748 - 39\,563)^2}{39\,563} = 11\,215.$$

The larger the value of the chi-square statistic, the less evidence there is that the assumption of independence holds. In order to determine if the value of 11 215 is in fact a value large enough to reject the assumption of independence, the critical value is required from a chi-square distribution with $(n-1) \times (p-1)$ degrees of freedom, where n is the number of rows and p is the number of columns of the table.

Following on from the chi-square statistic is the crucially important *Inertia* value. A generalized formula of the inertia value is:

$$Inertia = \sum_i (i^{th}\ mass) \times (\chi^2 - distance\ from\ i^{th} profile\ to\ centroid) \qquad (5\text{-}5)$$

(Greenacre, 2007:29).

A useful relationship exists between the chi-square statistic and the inertia value. Due to this relationship the inertia value can simplified to:

$$Inertia = \frac{\chi^2}{n}, \qquad (5\text{-}6)$$

where n is the total summation of the two-way table frequencies (Greenacre, 2007:28). The total inertia value will be low when profiles are situated close to the average profile, and high when they are located away from the average profile. If all profiles lie at the same point, then the distance between them is zero and thus the inertia will be zero (the minimum inertia). The maximum inertia is obtained when the profiles lie exactly at the vertices. In the case of Table 5-1 the maximum inertia value is equal to 2, the dimensionality of the table. As previously stated, the calculation of the chi-square statistics are equal for both the row and columns and thus the inertia values are equal for both. For the current example the inertia value is 0.0564. As the maximum possible inertia value is 2, the total inertia value appears to be relatively small.

The inertia value, which is a measure of variability, is then used in order to analyse the deviation away from the independence assumption. The total inertia value does not change for a data set, regardless of the dimensionality of the approximation, but the amount of inertia being accounted for in a display does change. This leads to an evaluation method for correspondence analysis based on the inertia value.

5.2.3 Dimension Reduction

Unlike the previous example, many profiles will contain more than three elements. Such cases will require a coordinate system of more than two dimensions. In order to plot a high-dimensional (more than 3 dimensions) contingency table in a low-dimensional sub-space, a process of dimension reduction must be conducted.

Dimension reduction is achieved via the CA process. Similarly to the definition of MDS, CA is the process of finding a low-dimensional sub-space which approximately contains the profiles of a contingency table (Greenacre, 2007:43). Once a low-dimensional sub-space is found to represent the data, a measure of the quality of the display is needed. The total inertia is used as this measure. The quality of fit is determined by the discrepancy between the exact positions of the profiles and their approximations in the new space. This is evaluated by representing the inertia of the low-dimensional sub-space as a percentage of the total inertia of the full space. The lower the loss of inertia the greater the quality of the graph, and the greater the loss the worse the quality of the graph (Greenacre, 2007:43).

The criterion of dimension reduction is to reduce the loss of information when approximating profiles in a sub-space. Assuming that there are several dimensions to a CA, the goal is to eliminate the dimensions where there is the least variation between the profiles. If all the profiles are positioned relatively close to one another on a plane, then a minimal amount information is gained from this dimension and thus can be “removed” without much implication. If an appropriate sub-space of lower dimensionality is found, the profiles are then orthogonally projected onto this sub-space. Their respective positions will thus only be an approximation of the true positions. It then follows that the χ^2 distances between the profiles will too be an approximation of the original distances (Greenacre, 2007:46).

There are a number of possible sub-spaces for which the profiles can be perpendicularly projected upon. The ideal sub-space is a two-dimensional plane of which the distances from profile points to the new sub-space is minimised. The χ^2 distance is used in order to determine which plane is closest to the original space. As not all points have equal weighting, the masses of the individual profiles needs to be accounted for in the search for the closest sub-space. The search would be that of a least-squares problem if not for the profile weights. For the case of the two-way table, it is the respective row or column masses that assume the role of the weights. Taking the weights into account creates a weighted sum of squared distances problem (Greenacre, 2007:46).

The CA problem is further simplified when viewed in terms of the singular value decomposition (SVD). The SVD is a very powerful mathematical result and is especially useful in dimension reduction (Greenacre, 2007:47). By applying the SVD to the two-way contingency table, the optimal solution is neatly obtained in a few simple steps.

5.2.4 The CA argument in terms of the SVD

The most widely used form of correspondence analysis is that which approximates the individual profile deviations from the independence assumption. Denoting the two-way table as $\boldsymbol{X}$, then the vectors containing the row and column sums of $\boldsymbol{X}$ are defined as $\boldsymbol{r} = \boldsymbol{X}\mathbf{1}$ and $\boldsymbol{c} = \boldsymbol{X}^T\mathbf{1}$ respectively. Matrices $\boldsymbol{R}$ and $\boldsymbol{C}$ are then diagonal matrices of $\boldsymbol{r}$ and $\boldsymbol{c}$ respectively.

The deviation from the independence model can be written as:

$$\boldsymbol{X} - \boldsymbol{E} = \boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}. \qquad (5\text{-}7)$$

Introducing the weights of the row and columns into the equation produces the weighted deviation model:

$$\boldsymbol{R}^{-½}(\boldsymbol{X} - \boldsymbol{E})\boldsymbol{C}^{-½} = \boldsymbol{R}^{-½}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\boldsymbol{C}^{-½} = \boldsymbol{S}. \qquad (5\text{-}8)$$

It is then possible to find the optimal solution to the CA problem by simply taking the SVD of the weighted independence model. In this case the optimal solution is:

$$\boldsymbol{S} = \boldsymbol{U\Sigma V'}. \qquad (5\text{-}9)$$

The matrix $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values in descending order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. The principal inertias (λ_k) then follow as the square singular values, thus $\lambda_k = \sigma_k^2$. The principal inertias represent the inertia accounted for by the k^{th} principal axis. The respective coordinates for the rows and columns points of the two-way table then follow as:

$$Principal\ coordinates\ of\ rows: \boldsymbol{F} = \boldsymbol{R}^{-½}\boldsymbol{U\Sigma} \qquad (5\text{-}10)$$

$$Principal\ coordinates\ of\ columns: \boldsymbol{G} = \boldsymbol{C}^{-½}\boldsymbol{V\Sigma} \qquad (5\text{-}11)$$

$$Standard\ coordinates\ of\ rows: \boldsymbol{\Phi} = \boldsymbol{R}^{-½}\boldsymbol{U} \qquad (5\text{-}12)$$

$$\mathit{Standard\ coordinates\ of\ columns}: \boldsymbol{\Gamma} = \boldsymbol{C}^{-½}\boldsymbol{V} \qquad (5\text{-}13)$$

The optimal r-dimensional solution is then obtained by plotting the first r columns of a respective set of principal and standard coordinates. An important relationship exists between principal and standard coordinates. Scaling the standard coordinates, by multiplying them by the singular values from the SVD of the independence model, will provide their respective principal coordinates. Similarly, the standard coordinates can be produced from principal coordinates. It is important to note that this study makes strict use of the asymmetric CA, thus in each figure there is a set of principal coordinates and standard coordinates. A figure containing both sets of principal coordinates is known as a symmetric CA and is disregarded in this study.

5.2.5 Optimal Scaling

The CA problem can be addressed in a variety of manners. An alternative method for obtaining the optimal CA solution is by optimal scaling.

As defined by Greenacre, *"optimal scaling provides a way of obtaining quantitative scale values for a categorical variable, subject to a specific criterion of optimality"* (Greenacre, 2007:50). The specific criterion of optimality in this case is to maximize the variance between groups by assigning values to the various levels of the categorical variable.

The concept of optimal scaling is then demonstrated by applying a numerical scaling value to the types of crime in Table 5-1. Once a starting set of scale values has been selected for the crime types, an optimization procedure is then executed such that the variability between the provinces is maximised. Applying a scale of 1 to 5 for the crime types in Table 5-1 produces Table 5-5.

***Table 5-5**: A reproduction of Table 5-1 where the crime categories have been replaced with numerical scale values.*

Crime	**EC**	**FS**	**GT**	***Total***
1	3 453	946	3 333	7 732
2	9 897	4 814	11 021	25 732
3	1 858	911	3 901	6 670
4	27 451	14 531	41 581	83 563
5	13 392	17 124	44 748	75 264
Total	56 051	38 326	104 584	198 961

From Table 5-5 the average crime value is calculate as:

$$\frac{[(7\,732\times1)+(25\,732\times2)+\cdots+(75\,264\times5)]}{198\,961} = 3.9695.$$

Now consider the three different provinces and their individual average crime value. For the Eastern Cape the average crime value is calculated as:

$$\frac{[(3\,453\times1)+(9\,897\times2)+\cdots+(13\,392\times5)]}{56\,051} = 3.6678.$$

Similarly, the average crime values for the Free State and Gauteng are 5.3641 and 4.0842 respectively. The weighted variance can then be calculated between the provinces as:

$$\frac{56\,051}{198\,961}(3.6678-3.9695)^2+\frac{38\,326}{198\,961}(5.3641-3.9695)^2+\frac{104\,584}{198\,961}(4.0842-3.9695)^2=0.4072,$$

with a standard deviation of $\sqrt{0.4072}=0.6381$.

The weighted variance calculation is dependent on the choice of the initial integer scale for the crime categories. Scores are then introduced as a setting for optimising this weighted variance. If the integer scale values for the crime categories are set as random variables, denoted as $v_1, v_2, \ldots, v_5$, the initial overall average would then assume the more general form:

$$\frac{[(7\,732\times v_1)+(25\,732\times v_2)+\cdots+(75\,264\times v_5)]}{198\,961}=Overal\ Average\ Crime\ Level\ .$$

Following from this, the average for the Eastern Cape, and similarly the other provinces, would be calculated as:

$$\frac{[(3\,453\times v_1)+(9\,897\times v_2)+\cdots+(13\,392\times v_5)]}{56\,051}=Average\ Crime\ Level\ for\ the\ Eastern\ Cape.$$

When the calculations are denoted in such a manner they are then known as *scores* (Greenacre, 2007:51). Let s_1, s_2, s_3 represent the scores for the Eastern Cape, Free State, and Gauteng respectively. The desirable property for the scores is that they are as distinct from one another as possible. This translates to maximising the variance across the scores, and the scale values which lead to the maximum variance is what is known as *Optimal Scaling*.

There are numerous optimising methods which can be used to solve the scaling problem but luckily, as it so happens, the positions of the crime categories along the best-fitting CA dimension solve this problem exactly (Greenacre, 2007:52). Due to the connection between the two approaches, the maximum variance of the optimal scaling is equivalent to the inertia on the optimal CA dimension. The scaling values and the corresponding scores for the optimal scaling solution of Table 5-1 are then given by the optimal CA solution of the standard coordinates of the crimes and the principal coordinates of the provinces, respectively.

5.3 Biplots

Biplots are plots, usually in two or more dimensions, which display both the observations and variables of a data matrix simultaneously (Cox & Cox, 2001:153). Most common forms of biplots are in a two-dimensional space, but one- and three-dimensional biplots can also be constructed. The methodology behind biplots involves the search for a r-dimensional sub-space upon which the sample points can be orthogonally projected, where $r < p$, and p represents the number of variables of a data matrix. A great deal of literature is available on the search for said sub-space (Gower & Hand, 1995; Cox & Cox, 2001; Greenacre, 2010; Gower *et al.*, 2011)

A special case of the biplot, where the sample points are projected onto a sub-space with only two variables (axes), is the common scatterplot. Through biplot methodology it is possible to display more than two variables (or axes) in a single diagram along with the sample points, this process is briefly explained below.

5.3.1 The Theory of Biplots

The methodology and corresponding matrix algebra which forms the body of biplot theory is presented in this subsection. There are many detailed texts on this topic which are readily available, *Understanding Biplots* by Gower, Le Roux, & Lubbe is strongly advised for further reading (Gower *et al.*, 2011). Additionally, *Biplots in Practice* by Greenacre is recommended for its wide variety of applications (Greenacre, 2010). The following sections are derived from Chapter 2 of *Understanding Biplots* (Gower *et al.*, 2011:11).

Let $\boldsymbol{X}$ be a data matrix of size $n \times p$, then the complete singular value decomposition (5-3) of matrix $\boldsymbol{X}$ can be written as:

$$\boldsymbol{X} = \boldsymbol{U}^*\boldsymbol{\Sigma}^*(\boldsymbol{V}^*)', \tag{5-14}$$

where $\boldsymbol{U}^*$ is an $n \times n$ orthogonal matrix with columns known as the left singular vectors of $\boldsymbol{X}$, $\boldsymbol{V}^*$ is a $p \times p$ orthogonal matrix with the columns known as the right singular vectors of $\boldsymbol{X}$, and $\boldsymbol{\Sigma}^*$ is a matrix of the singular values (of size $n \times p$) of $\boldsymbol{X}$ taking the form of $\boldsymbol{\Sigma}^* = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$. Note that $\boldsymbol{\Sigma}$ in this case is a $k \times k$ matrix with the diagonals as the singular values of $\boldsymbol{X}$, where k is the rank of a matrix, given by the number of non-zero singular values of the data matrix $\boldsymbol{X}$.

It then follows that the reduced SVD of $\boldsymbol{X}$ can be written as:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}', \tag{5-15}$$

where $\boldsymbol{U}$ is an $n \times k$ matrix with columns known as the first k non-zero left singular vectors of $\boldsymbol{X}$, $\boldsymbol{V}$ is a $p \times k$ orthogonal matrix with the columns known as the first k non-zero right singular vectors of $\boldsymbol{X}$. The matrix of singular values, $\boldsymbol{\Sigma}$, remains as defined above.

Biplots focus on displaying multiple variables and observations in a low-dimensional sub-space. When a low-dimensional sub-space is found in which to display a high-dimensional data matrix, the display is then no longer a true representation of the data, but rather an approximation thereof. The r-dimensional approximation of matrix $\boldsymbol{X}$ can be written as:

$$\widehat{\boldsymbol{X}}_{[\boldsymbol{r}]} = \boldsymbol{U}\boldsymbol{\Sigma}_{[r]}\boldsymbol{V}', \tag{5-16}$$

where $\boldsymbol{\Sigma}_{[r]}$ is a $k \times k$ matrix with diagonals as the singular values of $\boldsymbol{X}$, but where the remaining $k - r$ smallest diagonal values become zero. This r-dimensional approximation is optimal in the least squares sense such that

$$\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2 = tr\{(\boldsymbol{X} - \widehat{\boldsymbol{X}})(\boldsymbol{X} - \widehat{\boldsymbol{X}})'\} \tag{5-17}$$

is minimized for all matrices of $\hat{\boldsymbol{X}}_{[\boldsymbol{r}]}$, of rank no larger than r (Gower *et al.*, 2011:17).

In *Understanding Biplots*, Gower, Le Roux, & Lubbe introduce the $\boldsymbol{J}$-notation (Gower *et al.*, 2011:17). This notation acts as a convenient method of denoting an r-dimensional approximation of data matrix $\boldsymbol{X}$ in the setting of the complete SVD of $\boldsymbol{X}$. The $\boldsymbol{J}$ matrix of size $p \times p$ is denoted as:

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \qquad \textit{(5-18)}$$

where $\boldsymbol{I}_r$ is a matrix of size $r \times r$ with the diagonal values of unity and all others as zero (Gower *et al.*, 2011:17).

Using the $\boldsymbol{J}$-notation, the r-dimensional approximation, $\hat{\boldsymbol{X}}_{[\boldsymbol{r}]}$, can be denoted as:

$$\hat{\boldsymbol{X}}_{[\boldsymbol{r}]} = \boldsymbol{U\Sigma JV}' = \boldsymbol{UJ\Sigma V}' = \boldsymbol{UJ\Sigma JV}'. \qquad \textit{(5-19)}$$

(Gower *et al.*, 2011:17). After introducing the $\boldsymbol{J}$ matrix in the calculation, $\boldsymbol{UJ}$ and $\boldsymbol{VJ}$ have zeros in the final $p - r$ columns but the matrices remain of size $p \times p$.

The coordinates for the r-dimensional approximation for the rows (samples or cases) and columns (variables) of $\boldsymbol{X}$ are obtained from the first r columns of $\boldsymbol{UJ\Sigma}$ and $\boldsymbol{JV}'$ respectively. By plotting both sets of coordinates in the same space, both representations of the rows and columns of the matrix $\boldsymbol{X}$ are obtained. Often the row coordinates are represented as points and the column coordinates are represented as vectors extending from the origin.

The following subsection provides details of the interpretation of biplot figures.

5.3.2 Interpreting Biplots

Biplots have a unique interpretation method. Unlike CAs, biplots focus on the scalar products between vectors, rather than inter-point distances. The scalar products between two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, can be denoted as:

$$\boldsymbol{x}^T\boldsymbol{y} = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \cdot \cos(\theta). \qquad \textit{(5-20)}$$

In biplot theory, the two sets of coordinates are graphed in a joint display, and are provided for by the left and right matrices of the decomposition of the target matrix (Greenacre, 2010:21). Either set of coordinates can be chosen to portray the biplot axes, and is often represented by points which are connected by vectors to the origin. The remaining set of points assume the role of biplot observation points. The values of the target matrix can then be recovered by perpendicularly projecting the biplot points onto the relative biplot axes. The projection of vector $\boldsymbol{x}$ onto the biplot axis $\boldsymbol{y}$ is simply calculated as the length of $\boldsymbol{x}$ multiplied by the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus, the scalar product of vector $\boldsymbol{x}$ and $\boldsymbol{y}$ can be interpreted as the projection of $\boldsymbol{x}$ on $\boldsymbol{y}$, multiplied by the length of $\boldsymbol{y}$ (Greenacre, 2007:98). It is these scalar products which recover the elements of the target matrix. A graphical representation of such results is provided in Figure 5-2.

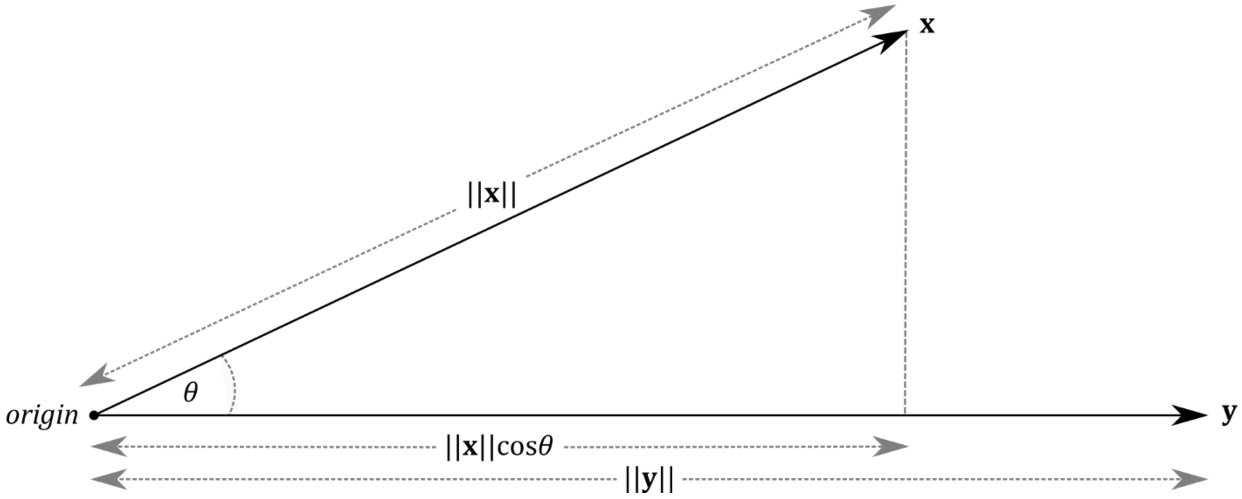

Figure 5-2: A graphical representation of the orthogonal projection of vector x onto vector y.

5.3.3 Lambda-Scaling

Lambda-scaling plays an important role in optimizing the biplot display. Often either the points of the rows, $\boldsymbol{A}$, or the points of the columns, $\boldsymbol{B}$, have much greater dispersion than the other. Lambda-scaling can be utilized in order to scale the coordinates of the row and column points to adjust for this difference in dispersion. From equation 5-15 it follows that:

$$\boldsymbol{X} = \boldsymbol{U\Sigma V}' = \boldsymbol{AB}, \tag{5-21}$$

where $\boldsymbol{A} = \boldsymbol{U\Sigma}$ and $\boldsymbol{B} = \boldsymbol{V}'$ are the coordinates for plotting the row and column points respectively. It then follows that:

$$\boldsymbol{X} = \boldsymbol{AB} = (\lambda\boldsymbol{A})(\boldsymbol{B}/\lambda) \tag{5-22}$$

(Gower *et al.*, 2011:24).

Both the rows and columns are scaled by lambda, but inversely of the other in order to maintain the inner products. The advantage of using Lambda-scaling over other general scaling methods is that it has a minimal effect on distances. General scaling affects distances such as the χ^2, Mahlanobis, and Pythagorean distances severely (Gower *et al.*, 2011:24). This is of particular importance when utilizing correspondence analysis (see Section 5.2) which approximates deviations from independence by a measure of distance, usually the χ^2 distance. Lambda-scaling is utilised numerous times when producing figures in later chapters.

5.3.4 Manipulating the Display

An important property of biplots is that the plot itself can be rotated and/or reflected about the horizontal or vertical axes without violating the inner products of equation 5-17. Following from equation 5-19:

$$\widehat{\boldsymbol{X}}_{[r]} = (\boldsymbol{UJ\Sigma})(\boldsymbol{VJ})' = (\boldsymbol{UJ\Sigma Q})(\boldsymbol{VJQ})' = \boldsymbol{A}_{[r]}\boldsymbol{B}_{[r]}, \tag{5-23}$$

which holds for any $p \times p$ orthogonal matrix $\boldsymbol{Q}$ (Gower *et al.*, 2011:20). The row and column markers in r dimensions are given by $\boldsymbol{A}_{[\boldsymbol{r}]}$ and $\boldsymbol{B}_{[\boldsymbol{r}]}$ respectively in equation 5-23. Equation 5-23 represents one of the most functional properties in biplot theory. Due to the fact that the SVD of data matrix $\boldsymbol{X}$ is only unique up to a multiplication of modulus one, it is not guaranteed that the same calculation will yield the same result when replicated on different computers. Due to this property and the results of equation 5-23 it is possible to subject a biplot to orthogonal rotation and/or reflection without loss of information. This property becomes of great importance when trying to replicate or reproduce results on separate computers. Often a rotation and or reflection is required to perfectly match the original result.

5.3.5 Calibrating Axes

The value of a biplot is truly emphasized when the calibration of the axes is introduced. Although a considerable amount of information is gained by simply introducing the uncalibrated axes, an additional level of interpretation is gained by having calibrated tick marks. Adding the calibrated tick marks provides a method for reading off the values of the target matrix by simply projecting the points onto the relative biplot axis. As the projected values are directly proportional to the target values, it is possible to scale the biplot axis with a scale which depends on the length of the respective axis. After which, the values of the target matrix can simply be recovered by reading off the projections of the respective points onto the calibrated axes, and no multiplication by the length of the axis is required.

In order to determine the length of one unit on a biplot axis, it is necessary to think of how the value of value i in the target matrix is recovered:

$$target\ value\ in\ row\ i = \begin{pmatrix} length\ of\ projection \\ of\ i^{th}\ biplot\ point \end{pmatrix} \times (length\ of\ biplot\ vector) \tag{5-24}$$

(Greenacre, 2010:22). Following from equation 5-24, the length of one unit can be determined by inverting this formula for a target value of unity:

$$length\ of\ projection\ of\ 1\ unit = \frac{1}{length\ of\ biplot\ vector} \tag{5-25}$$

(Greenacre, 2010:22). Therefore, the inverse of the lengths of the biplot vectors provide the unit lengths on the biplot axes. Once these lengths are known, the appropriate tick marks can be added to the biplot axes to improve the interpretability of the figures. Note that the target matrix being approximated may sometimes provide the transformed values of the original variable of interest. In such cases, it may be desirable to be calibrated the axes in terms of the untransformed values. An example of this is that of Correspondence analysis, where the original counts are replaced with row and/or column scaled deviations from the independence model. Several variants of the CA Biplot are discussed in Subsection 5.4.1, and as such, the tick marks will need to be calibrated according to the various methods.

5.4 The CA Biplot

Both the biplot and correspondence analysis algorithms are structured around the singular value decomposition. It is not surprising then that these two methods are closely interlinked, in fact an asymmetric CA diagram with both the rows and column points plotted in the same space is technically a biplot. The two techniques have been neatly integrated by Gower, Le Roux, & Lubbe in Chapter 7 of *Understanding Biplots* (2011:289). Prior to this, Gower & Hand introduced the concept of correspondence analysis as a type of biplot in their book *Biplots* (1996:175). Gower, Le Roux & Lubbe expanded on previous theories to provide several variants of the CA biplot. By manipulating the original weighted deviations model of the CA algorithm, several different metrics can be displayed in a CA biplot.

5.4.1 CA Biplot Variants

The various types of CA biplots are discussed in great detail in Chapter 7 of *Understanding Biplots* (Gower *et al.*, 2011:289). Provided below is an explanation of the several variants of the independence model that can be approximated in the CA biplot.

5.4.1.1 *Pearson's chi-squared:*

By making use of the weighting the independence model, the contributions to the Pearson's chi-squared statistic can be approximated. Thus following from equation 5-17 the model which must be minimised is:

$$\left\| \boldsymbol{R}^{-½}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\boldsymbol{C}^{-½} - \widehat{\boldsymbol{X}} \right\|^2, \tag{5-26}$$

where $\widehat{\boldsymbol{X}}$ is the k-dimensional approximation of $\boldsymbol{R}^{-½}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\boldsymbol{C}^{-½}$ (Gower *et al.*, 2011:291). Note that the SVD of $\widehat{\boldsymbol{X}}$, the r-dimensional approximation of $\boldsymbol{X}$, is written as $\boldsymbol{UJ\Sigma V'}$. The r-dimensional approximation of the χ^2 contributions requires plotting of the first r columns of $\boldsymbol{U\Sigma}^{½}$ and $\boldsymbol{V\Sigma}^{½}$. Such biplots provide a display from which it is possible to identify the elements of $\widehat{\boldsymbol{X}}$ which contribute most to the total inertia as well as which element diverge from the assumption of independence.

5.4.1.2 *Approximating the deviations from independence:*

This method seeks to approximate the deviations from independence, $\boldsymbol{X} - \boldsymbol{E}$, which are weighted by the inverse square roots of the row and column totals. This translates to the minimization of:

$$\left\| \boldsymbol{R}^{-½}\left\{\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right) - \widehat{\boldsymbol{X}}\right\}\boldsymbol{C}^{-½} \right\|^2 \tag{5-27}$$

(Gower *et al.*, 2011:291). Equation 5-27 is minimized by $\boldsymbol{R}^{-½}\widehat{\boldsymbol{X}}\boldsymbol{C}^{-½} = \boldsymbol{U\Sigma JV'}$. Solving for the r-dimensional approximation of $\boldsymbol{X}$ gives $\widehat{\boldsymbol{X}} = \boldsymbol{R}^{½}\boldsymbol{U\Sigma JV'C}^{½}$. By plotting the first r columns of $\boldsymbol{R}^{½}\boldsymbol{U\Sigma}^{½}$ and $\boldsymbol{C}^{½}\mathbf{V\Sigma}^{½}$ a biplot which represents the deviations away from independence is achieved.

5.4.1.3 *Approximation to the contingency ratio:*

The contingency ratio, as defined in *Correspondence Analysis in Practice* (Greenacre, 2007:101) is the ratio of observed proportions of a contingency table to the expected proportions. A CA biplot which approximates the contingency ratios is defined as the minimization of:

$$\left\| \boldsymbol{R}^{-½} \left\{ \boldsymbol{R}^{-1} \left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n} \right) \boldsymbol{C}^{-1} - \widehat{\boldsymbol{X}} \right\} \boldsymbol{C}^{-½} \right\|^2 \tag{5-28}$$

(Gower *et al.*, 2011:292). In equation 5-28 the matrix $\widehat{\boldsymbol{X}}$ approximates $\boldsymbol{R}^{-1}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\boldsymbol{C}^{-1} = \boldsymbol{R}^{-½}\boldsymbol{U\Sigma V'}\boldsymbol{C}^{-½}$. Note that a contingency ratio of unity is desirable, and the closer the value is to unity the closer the data are to independence (Greenacre, 2007:101). By plotting the first r columns of $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma}^{½}$ and $\boldsymbol{C}^{-½}\boldsymbol{V\Sigma}^{½}$ a biplot which depicts the deviation from the unity contingency ratio is obtained. Note that the biplot does not provide a direct deviation from unity but rather the deviation divided by n. This characteristic of the biplot has no material influence on the biplot but must be accounted for when calibrating the axes.

5.4.1.4 *Approximation to the Chi-squared distance:*

The independence assumption is evaluated by the χ^2 statistic. For any two-way table, the assumption can be tested via the χ^2 distance between rows or the χ^2 distance between the columns of the table. Therefore, there are two possible solutions for when approximating the Chi-squared distances in a contingency table.

The chi-squared distances between the ith and i'th rows of $\boldsymbol{X}$ are defined as:

$$d_{ii'}^2 = \left(\frac{\boldsymbol{x_i}}{x_{i.}} - \frac{\boldsymbol{x_{i'}}}{x_{i'.}}\right)' \boldsymbol{C}^{-1} \left(\frac{\boldsymbol{x_i}}{x_{i.}} - \frac{\boldsymbol{x_{i'}}}{x_{i'.}}\right) \tag{5-29}$$

(Gower *et al.*, 2011:293). Equation 5-29 is of the form of weighted Euclidean distances between the rows of $\boldsymbol{X}$. The chi-squared distances are then calculated as the normal Euclidean distances between the rows of $\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{-½}$. This can also be expressed as $\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{-½} - \frac{\boldsymbol{11'C}^{½}}{n}$, where $\frac{\boldsymbol{11'C}^{½}}{n}$ is the translation term. The translation term has no material effect on the chi-squared measures between rows of $\boldsymbol{X}$ but translates the points about the centroid in the biplot. The translated chi-squared distances can then be written as:

$$\boldsymbol{R}^{-1}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\boldsymbol{C}^{-½} = \boldsymbol{R}^{-1}(\boldsymbol{X} - \boldsymbol{E})\boldsymbol{C}^{-½} \tag{5-30}$$

(Gower *et al.*, 2011:294). The biplot of the chi-squared distances between the rows of matrix $\boldsymbol{X}$ is then based upon the minimization of:

$$\left\| \boldsymbol{R}^{½} \left\{ \boldsymbol{R}^{-1} \left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n} \right) \boldsymbol{C}^{-½} - \widehat{\boldsymbol{X}} \right\} \right\|^2, \tag{5-31}$$

which simplifies to:

$$\left\| \boldsymbol{U\Sigma V}' - \boldsymbol{R}^{½}\widehat{\boldsymbol{X}} \right\|^2. \tag{5-32}$$

The approximation, $\widehat{\boldsymbol{X}}$, is then obtained from the inner product of $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma V}'$ and the r-dimensional biplot is obtained by plotting the first r columns of $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma}$ for the row points and $\boldsymbol{V}$ for the column points (Gower *et al.*, 2011:294). Note that $\boldsymbol{V}$ is an orthogonal matrix and has no effect on the distances between the row coordinates.

Similar arguments follow for the chi-squared distances between the columns of $\boldsymbol{X}$. The chi-squared distances between the jth and j'th columns of $\boldsymbol{X}$ are defined as:

$$d_{jj'}^2 = \left(\frac{\boldsymbol{x}_j}{x_{.j}} - \frac{\boldsymbol{x}_{j'}}{x_{.j'}}\right)' \boldsymbol{R}^{-1} \left(\frac{\boldsymbol{x}_j}{x_{.j}} - \frac{\boldsymbol{x}_{j'}}{x_{.j'}}\right). \tag{5-33}$$

This leads to a biplot which attempts to minimise:

$$\left\| \left\{ \boldsymbol{R}^{-½} \left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n} \right) \boldsymbol{C}^{-1} - \widehat{\boldsymbol{X}} \right\} \boldsymbol{C}^{½} \right\|^2, \tag{5-34}$$

which simplifies to:

$$\left\| \boldsymbol{U\Sigma V}' - \widehat{\boldsymbol{X}}\boldsymbol{C}^{½} \right\|^2. \tag{5-35}$$

The approximation, $\widehat{\boldsymbol{X}}$, is then obtained from the inner product of $\boldsymbol{U\Sigma V}'\boldsymbol{C}^{-½}$ and the r-dimensional biplot is created by plotting the first r columns of $\boldsymbol{C}^{-½}\boldsymbol{V\Sigma}$ for the row points and $\boldsymbol{U}$ for the column points. Similarly, $\boldsymbol{U}$ is also an orthogonal matrix which has no effect on the distances between the points representing the columns of $\widehat{\boldsymbol{X}}$.

It is common practice to plot both the row and column chi-squared distances in one single plot, although the relationships between these two measures has no meaningful interpretation. This plot is simply achieved by plotting $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma}$ and $\boldsymbol{C}^{-½}\boldsymbol{V\Sigma}$ simultaneously as two sets of points.

5.4.1.5 *Canonical Correlation Approximation:*

Canonical correlation is associated with the optimal CA solution. The optimal solution can be viewed as a function of the correlation between the rows and columns of the two-way table. When the CA solution is aimed to maximize the correlation, then the name of the corresponding correlation for the optimal solution is known as *canonical correlation.* In this variant of the CA biplot there will be a search for the optimal row and column scores such that the correlation between rows and columns is maximised. Due to its lengthy derivation, the explanation of said method is not provided for. Rather, the reader is referred to Chapter 7 of *Understanding Biplots* for a detailed explanation (Gower *et al.*, 2011:296).

5.4.1.6 *Approximating the row profiles:*

The final variant of the CA biplot is centred upon the row profiles, $\boldsymbol{R}^{-1}\boldsymbol{X}$, of the two-way table. Similarly, the approximation of the column profiles can be achieved by simply replacing the input matrix with the transpose of $\boldsymbol{X}$. In this case the suggested biplot model is:

$$\left\| \boldsymbol{R}^{½} \left\{ \boldsymbol{R}^{-1} \left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n} \right) - \widehat{\boldsymbol{X}} \right\} \boldsymbol{C}^{-½} \right\|^2, \tag{5-36}$$

and the approximation $\widehat{\boldsymbol{X}}$ takes the form of $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma V'C}^{½}$ (Gower *et al.*, 2011:298). The accompanying biplot is obtained by plotting the first r columns of $\boldsymbol{R}^{-½}\boldsymbol{U\Sigma}$ for the rows and $\boldsymbol{C}^{½}\boldsymbol{V}$ for the columns. The distances expressed in this biplot are chi-squared distances, and the points for the columns provide axes for the row profile points. Here the row profiles are not the pure row profiles but rather the deviations from the marginal row profile, $\frac{\boldsymbol{1'C}}{n}$.

5.4.2 Functions for Constructing CA Biplots

In their book, Gower, Le Roux, & Lubbe focus heavily on the graphical representation of data (Gower *et al.*, 2011). Accompanying this textbook is a library of functions entitled *UBbipl* and is available at www.wiley.com/go/biplots. This R package contains all the relevant functions to replicate the biplot graphics within the textbook. Of particular interest here is the *cabipl* function which is contained in the *UBbipl* package (Le Roux & Lubbe, 2013). In their textbook, Gower, Le Roux, & Lubbe provide a breakdown of all the functions for constructing CA biplots, as well as an explanation of the function arguments (Gower *et al.*, 2011:306).

The *cabipl* function has the ability to construct several different forms of CA biplots. The argument *ca.variant* has eleven different entries which pertain to the several different forms of the CA biplot which were discussed in Section 5.4.1.

Chapter 5 has thus provided a broad theoretical overview of the GDA methods to be utilised in this study. A great deal of emphasis was placed on both correspondence analysis and the CA biplot. Chapter 6 aims to apply the aforementioned methods to the South African crime data described in Chapter 3. By utilising such methods, it is hoped that a better understanding of the South African crime data can be achieved.

Chapter 6: Bivariate Data Analysis

Bivariate data analysis encompasses all techniques that simultaneously analyse two variables of interest. In the case of the South African crime data, provinces and types of crime are said variables. Correspondence analysis and the correspondence analysis biplot are the bivariate techniques selected for the primary analysis in this thesis. The theoretical grounding of these two techniques was addressed in Chapter 5. The goal of this chapter is to simultaneously represent provinces and crimes in a single space, allowing for an analysis of the relationships between these two variables.

The compositions of the proportional frequency bar plots in Subsection 4.2.2 closely resembled CA profiles. The result of such analysis provided useful insight into the South African crime data, creating an ideal starting point for further analysis. The study now makes use of correspondence analysis in order to further scrutinize the profiles. The association between profiles relative to the vertex points, can be analysed by creating two-dimensional CA maps. In the CA map, the distances between points express the χ^2 distances between them, thus the further the points are away from each other the more they differ, and vice versa. Further details on the interpretation of the CA will be provided for in Section 6.1.

The CA biplot can display a number of various metrics as discussed in Subsection 5.4.1. For demonstration purposes, the analysis makes exclusive use of a CA biplot depicting the contributions to the χ^2 statistic. Said contributions are presented in the form of the Pearson's residuals. From experimentation, it was concluded that the various metrics provided similar insight, and that a negligible amount of information was lost by their exclusion. As the focus in this section is to determine how the provinces differ from one another (or the average profile), the contributions to the Pearson's χ^2 are expressed in the CA biplots which follow.

The chapter commences with a correspondence analysis on the 2004 South African crime data and progresses to the CA biplot for further analysis.

6.1 Correspondence Analysis

6.1.1 An Application to the 2004 South African Crime Data

The tables and graphs below result as an application of the theory provided in Chapter 5. Similar to Chapter 4, there are numerous potential graphs which can be produced. However, only the most relevant of said graphs are included below. The two-dimensional correspondence analysis display for the 2004 South African crime data (Figure 6-1) is used to demonstrate the interpretation of CA displays. In order to aid in interpretation, the reader is referred back to Table 3-1 and Table 3-2 for the abbreviations of the provinces and crimes respectively.

In Figure 6-1, the province points are represented in principal coordinates, and the crime points are in standard coordinates. Alternatively, the analysis is conducted with the provinces assuming the role as profiles, and the crime types as the vertices. This structure of the CA map will be used consistently throughout this chapter. Additionally, Figure 6-1, as well as all other CA displays which appear in this study, are asymmetric displays. The asymmetric display is utilized to ensure that the distances between profiles, relative to the vertex points, have an easily interpreted meaning. On the contrary, the symmetric display can lead to misinterpretation, hence its exclusion. In a symmetric display, both sets of row and column profiles are overlaid in a joint display, despite emanating from two different spaces (Greenacre, 2007:70). In the symmetric display, both sets of coordinates are displayed in principal coordinates, and as such the closeness of a profile point to a vertex point does not necessarily indicate a high association (Greenacre, 2007:72).

The focal point of the CA is to analyse the variation in a two-way table. This is done, similarly to a normal variance calculation, by comparing the individual profiles to the average profile. For correspondence analysis, the total inertia is the measure of variability. The details of the inertia value were explained in Subsection 5.2.2. A graphical setting for interpreting the inertia is provided for in Figure 6-1. The positions of the various province profiles with respect to the average profile (the centroid) are being displayed in Figure 6-1. As seen in Subsection 5.2.2, the formula for the total inertia is:

$$Inertia = \sum_i \left(i^{th}\ mass\right) \times (\chi^2 - distance\ from\ i^{th} profile\ to\ centroid). \qquad \textit{(6-1)}$$

Equation 6-1 is composed of two parts, the distance of the i^{th} profile to the centroid, and the i^{th} mass. Both components of equation 6-1 are displayed in Figure 6-1. Firstly, the distances of the profiles (the dark blue triangles) to the centroid are displayed. As it happens, all the profiles are clustered together in the centre of the graph, making interpretation difficult. This is a result of the low total inertia of the 2004 South African crime data. Asymmetric maps function well when the total inertia is large, but when the total inertia is small the profile points are clustered near the centre (Greenacre, 2007:80). In order to address this problem, a '*zoomed-in'* display of the profile points is provided in Figure 6-2.

In Figure 6-2, the distribution of the profiles about the mean profile (centroid) is more discernible than in Figure 6-1. Although there is minimal dispersion of the profiles away from the centroid, it is evident that the Western Cape, Gauteng, Limpopo, and the Northern Cape are situated furthest away from the average profile. By the analysis of distance alone, it would then appear that these provinces (profiles) would contribute the greatest to the total inertia.

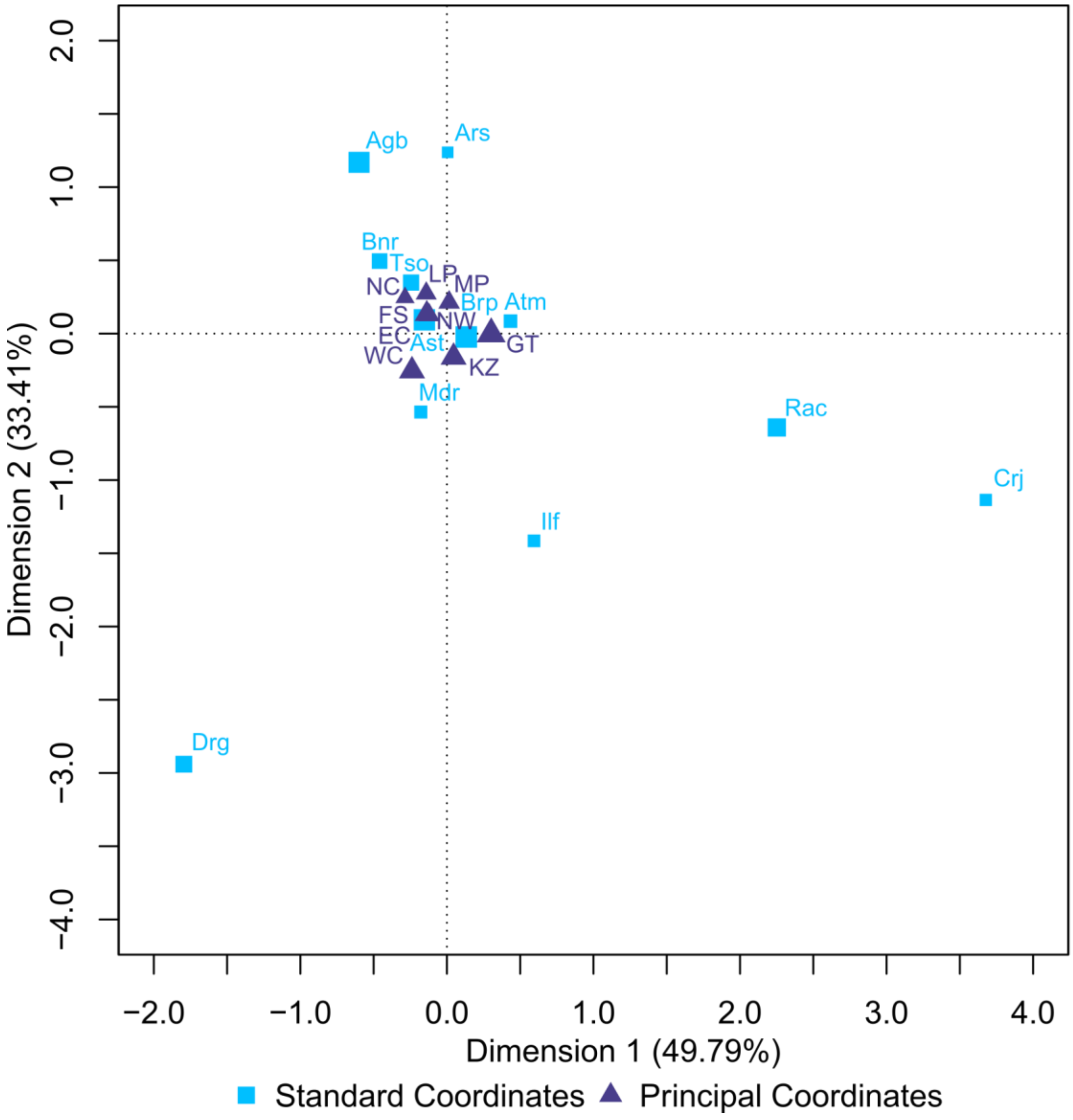

***Figure 6-1**: An asymmetric CA display of the 2004 South African crime data. Note that the plotting symbols are proportional to their respective masses.*

The actual contributions to the total inertia by each province in the 2004 crime data are displayed in Table 6-1. The inertia values, as well as their relative contribution in permills (parts per thousand) appear in Table 6-1. Additionally, conditional highlighting has been utilized in order to highlight any cell contributing more than 10% to the total inertia. This study assumes a contribution of 10% or greater to be meaningful. From Table 6-1, Gauteng and the Western Cape are determined as the two largest contributors to the total inertia value, together contributing more than 50%. Although the Northern Cape and Limpopo are positioned away from the centroid, they provide much smaller contributions. This is due to the fact that the total inertia calculation is dependent on both the distance of the i^{th} profile to the centroid as well as the i^{th} mass. Assuming that Gauteng and the Western Cape have the largest masses, and are situated away from the centroid, it is expected that they would contribute most to the total inertia.

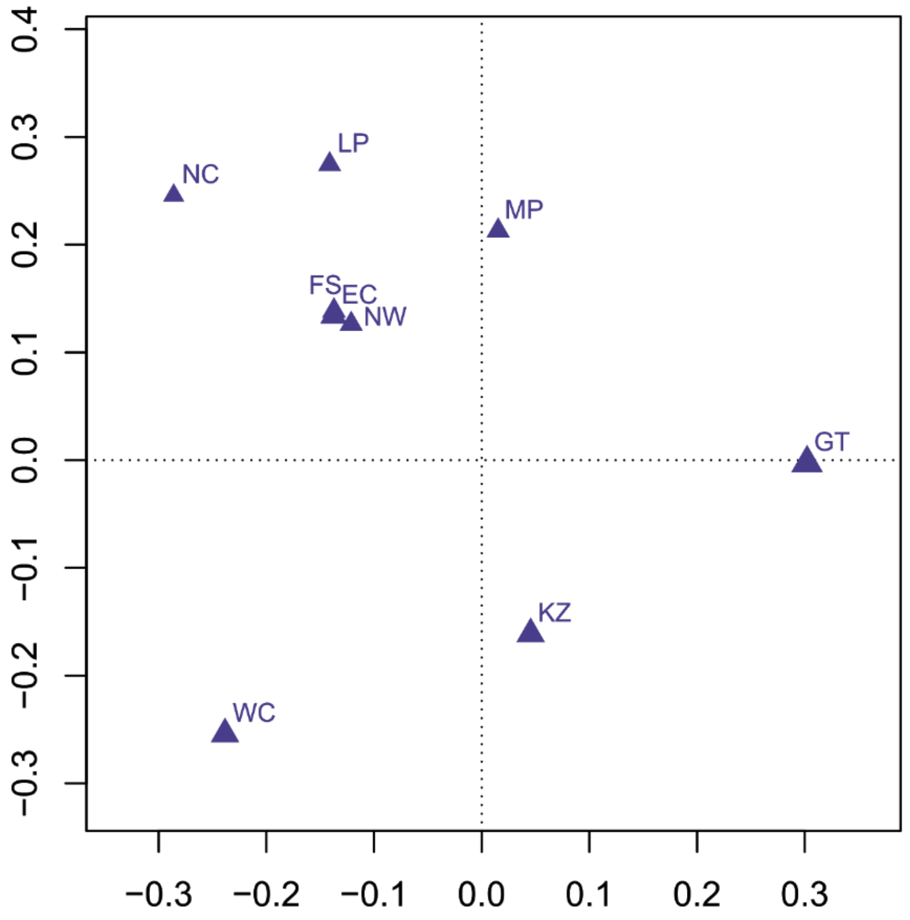

Figure 6-2*: A 'Zoomed-in' extract of the profile points from Figure 6-1.*

Both the profile and vertex masses are displayed in Figure 6-1, as each point is sized relative to its respective mass. The larger the mass, the larger the point, and the greater is its contribution to the total inertia. It can then be confirmed in Figure 6-2, that the Western Cape and Gauteng have the largest masses of all the provinces, reaffirming the assumption made above. The example provided for in Figure 6-1 focuses specifically on the province (profile) masses and their respective distances, and as such the graph is an asymmetric display with provinces assuming the role of the profile points, and crimes as the vertex points. Note that both sets of plotting characters displayed in Figure 6-1 are proportional to their masses, as this provides additional information required for proper interpretation.

***Table 6-1**: The contributions to inertia by the columns of the 2004 South African Crime data.*

Province	Inertia	Permills
EC	0.0071	83
FS	0.0053	62
GT	0.0252	295
KZ	0.0081	94
LP	0.0067	78
MP	0.0039	46
NW	0.0022	26
NC	0.0057	66
WC	0.0214	250

The symmetry between row and column analysis was stressed throughout Chapter 5. Recalling, the total inertia whether arguing via the rows or the columns is identical. The symmetry carries further, such that the relative positions of the vertex points would remain the same if plotted as profile points. This symmetry follows from the direct relationship between standard and principal coordinates. The two sets of coordinates are linked by a scaling factor of the singular values of the independence model, see *Appendix A* of Correspondence Analysis in Practice (Greenacre, 2007). Due to this symmetry, it is beneficial to view the individual crime contributions to the total inertia of the 2004 South African crime data as well. Figure 6-3 displays the vertex points of the 2004 South African crime data and their respective masses. From inspection of the figure, assault with the intent to cause grievous bodily harm, common assault (assault), and burglary at residential premises have the largest masses. As previously mentioned, it is not exclusively the mass, but a combination of mass and distance which determine the individual point's contribution to the inertia. Subsequently, it follows that assault with the intent to cause grievous bodily harm, robbery with aggravating circumstances, and drug-related crime are the largest contributors to the total inertia, as these points are large and positioned away from the centroid. Table 6-2 provides the inertia contributions by each crime type, as well as their relative contributions to the total inertia in permills. From Table 6-2, it can be confirmed that assault with the intent to cause grievous bodily harm, robbery with aggravating circumstances, and drug-related crime are in fact the largest contributors to the total inertia for the 2004 crime data.

The individual contributions by provinces, and by crimes, have thus far been analysed. What is of particular interest, and a strong point of CA, is to determine the contributions to inertia due to the interactions between the row and column of the table. These interactions can be identified in Figure 6-1. The calculations of the individual interactions between the rows and columns of the two-way table (each cell) for the 2004 South African crime data are provided for in Table 6-3. Table 6-3 is thus used to assist in the interpretation of Figure 6-1.

***Table 6-2**: The calculated contributions to inertia by the rows of the 2004 South African Crime data.*

Crime	Inertia	Permills
Agb	0.0125	146
Ars	0.0006	7
Ast	0.0050	59
Atm	0.0018	21
Bnr	0.0017	19
Brp	0.0014	16
Crj	0.0064	75
Drg	0.0269	314
Ilf	0.0026	30
Mdr	0.0020	24
Rac	0.0240	281
Tso	0.0008	9

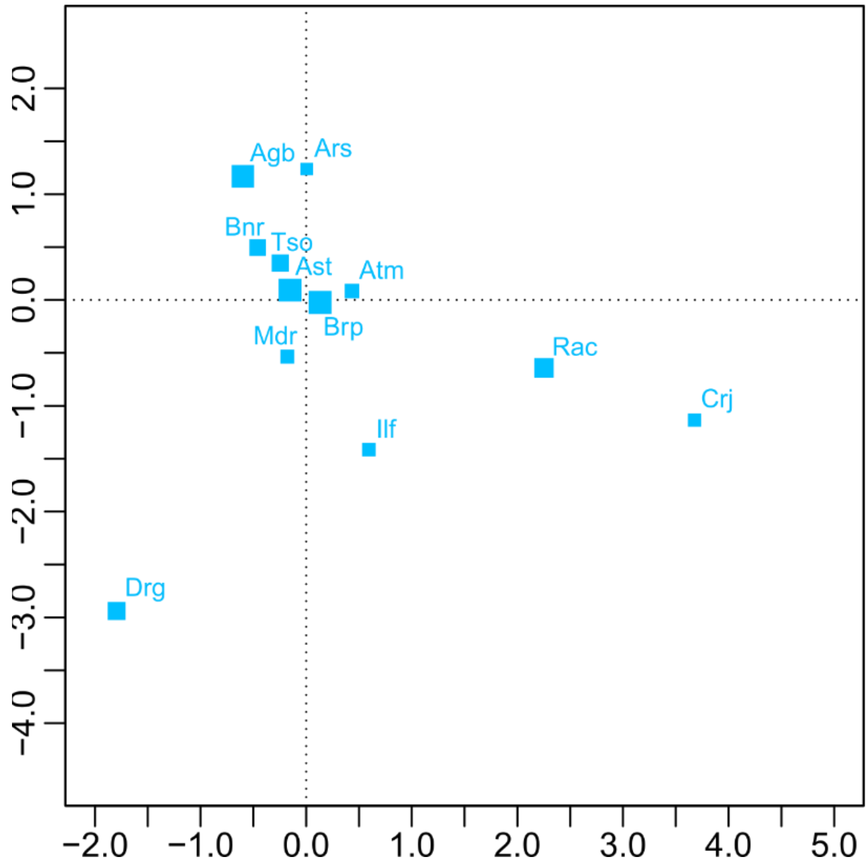

***Figure 6-3**: A 'Zoomed-in' extract of Figure 6-1, where only the standard coordinates are being displayed.*

Gauteng and the Western Cape were noted as the largest contributors to the total inertia value in Table 6-1. Subsequently, their profiles deviate the most with respect to the average profile. Such provinces then form the focal point of the analysis in this study. In the bivariate analysis it is possible to determine which crime types are associated with the respective provinces. The direction in which the profile points deviate from the centroid in Figure 6-1, provide information on their respective associations to vertex points. However, the two-dimensional display is only an approximation of the true space, and hence not all associations are necessarily well approximated. Therefore, care must be taken against over interpreting a CA display.

The axes in Figure 6-1 are marked with a percentage of total inertia accounted for by the individual axes. The first and second dimensions of Figure 6-1 account for 49.79% and 33.41% of the total inertia respectively, a total quality of 83.2%. Thus, almost 50% of the total variation in the 2004 South African crime data are accounted for by the first dimension alone. The purpose of conducting a CA is to find a low dimensional space which can represent high dimensional data well. How well the low dimensional representation approximates the real data are measured by the amount of variation within the data accounted for in the new sub-space. It is desirable that the first and second dimensions account for a large amount of the variation within the data. In this case, a one-dimensional representation of the data would not be adequate as approximately half the variation in the data would not be accounted for. However, as the two-dimensional approximation accounts for 83.2% of the total variation, it can thus be regarded as an adequate representation of the data. This does, however, mean that the relative positioning of the points are not perfectly represented and discretion must be used when interpreting the display.

Assuming that the CA plot is of good quality, it follows that the closer a profile point is to a vertex point, the higher its profile is for that category. Furthermore, the further away the vertex points are from the centre, the less influence they have in the average profile. The results of the univariate analysis suggested that there were certain crime types which were prominent in all of the profiles, and some crimes which were more affiliated with only a few provinces. The crime types which have an equal influence on all province profiles are positioned close to the centroid. Those which do not affect all provinces equally lie further away from the centre.

Deducing from Figure 6-1 and Figure 6-2, it appears that the Western Cape is drawn in the direction of murder and drug-related crime. Due to the fact that the murder vertex is positioned close to the centre, the Western Cape may only have a marginally higher than average profile value for murder. As drug-related crime is located away from the centre, it would be expected that the Western Cape has a large association to drug-related crime as compared to the other provinces. Making use of Table 6-3, it is revealed that the Western Cape and its association to drug-crime account for 18.6% of the total inertia. Conversely, the Western Cape's association to murder only accounts for 0.1% of the total inertia.

***Table 6-3**: Cell contributions to the total inertia for the 2004 South African crime data, expressed in permills.*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	5	1	3	11	1	0	1	1	1
Tso	0	0	3	0	4	0	1	0	1
Atm	0	1	0	9	1	0	1	3	6
Agb	38	1	24	13	10	9	9	28	14
Ast	11	26	0	9	6	1	0	0	4
Rac	19	19	156	8	20	2	5	22	29
Ars	2	0	0	0	2	0	0	0	3
Bnr	0	0	3	0	10	1	3	1	0
Brp	0	0	1	2	2	6	0	4	0
Ilf	0	4	0	20	1	0	1	3	0
Drg	1	5	60	20	18	24	0	0	186
Crj	6	5	44	2	3	1	3	4	6

Gauteng deviates away from the centroid in the direction of robbery with aggravating circumstances, and to a lesser extent in the direction of illegal firearms and carjacking. It is of interest to note that Gauteng does not deviate vertically away from the centroid but only horizontally. Due to this, it would be expected that a majority of the inertia contribution from Gauteng would be with regards to the vertex points which are spread out along the horizontal axis, such as robbery with aggravating circumstances and carjacking. From Table 6-3 it can be deduced that 15.6% of the total inertia is from Gauteng's association with robbery with aggravating circumstances, 6% is from drug-related crime, and 4.4% is from carjacking. Albeit that carjacking is positioned further away than robbery with aggravating circumstances, the latter has a much larger mass, leading to a larger inertia contribution from Gauteng's association to this crime. It is of interest to note that Gauteng's association to drug-related crime is not well represented in Figure 6-1. Although Kwa-Zulu Natal contributes only 2% to the total inertia with respect to drug-related crimes, it is positioned closer to the drug-related crime vertex point than Gauteng. This is due to the fact that in a two-dimensional approximation, not all points and relative distances can be perfectly represented.

A basis to the bivariate analysis has now been established. The general method of interpretation for correspondence analysis has been explained, and the methods have been applied to the 2004 South African crime data. In later CAs, the focus remains on the provinces which are the main contributors to the total inertia for each year. Sections 6.1.3 and 6.1.4 further the correspondence analysis for the following nine years of the reported period. Prior to this, an investigation into an alternative measure of association is investigated in Section 6.1.2.

6.1.2 The Association Matrix and its Corresponding Figures

In their book, Le Roux & Rouanet suggest measures of association between the rows and columns of a contingency table (Le Roux & Rouanet, 2004:32). In particular, they highlight the use of the *association rate* in the study of contingency tables. Denoting the contingency table of size $J \times K$ as $\boldsymbol{X}$, then let the association between j and k ($j = 1, \dots, J$; $k = 1, \dots, K$) be defined as:

$$a_{jk} = \frac{\left(x_{jk} - \frac{x_{j.}x_{.k}}{x_{..}}\right)}{\frac{x_{j.}x_{.k}}{x_{..}}}, \tag{6-2}$$

where $x_{j.}$, $x_{.k}$, and $x_{..}$ are the row, column, and grand totals of matrix $\boldsymbol{X}$ respectively. By denoting the matrix of associations as $\boldsymbol{A}$, equation 6-2 can be written as:

$$\boldsymbol{A} = \frac{\boldsymbol{R}^{-1}}{n}\left(\boldsymbol{X} - \frac{\boldsymbol{R11'C}}{n}\right)\frac{\boldsymbol{C}^{-1}}{n}. \tag{6-3}$$

Let matrices $\boldsymbol{R}$ and $\boldsymbol{C}$ are the diagonal matrices containing the row and column totals of $\boldsymbol{X}$, and let $n = \sum_j \sum_k x_{jk}$. It should be noted that equation 6-3 resembles that of equation 5-7 very closely. However, in the case of equation 6-3, the deviations from the independence model are weighted by $\frac{\boldsymbol{R}^{-1}}{n}$ and $\frac{\boldsymbol{C}^{-1}}{n}$, rather than $\boldsymbol{R}^{-1/2}$ and $\boldsymbol{C}^{-1/2}$.Thus, there is motivation to investigate the feasibility of utilizing such techniques in conjunction with the CA maps.

The association between observation j and variable k is null if $x_{jk} = x_{j.}x_{.k}$, positive (attraction) if $x_{jk} > x_{j.}x_{.k}$, and negative (repulsion) if $x_{jk} < x_{j.}x_{.k}$. The association matrix for the 2004 South African crime data is presented in Table 6-4. Additionally, the two largest attractions have been highlighted in the table. Table 6-4 cannot be compared directly to Table 6-3, as the tables are composed of different metrics. However, it is of interest to note that in both tables, the largest value is that of the Western Cape and drug-related crime. The second largest value is now that of Gauteng and carjacking, and not robbery with aggravating circumstances as was the case previously. Additionally, unlike in Table 6-3, the association matrix is composed of both positive and negative values. Another notable difference between the tables is that Table 6-3 provided a measure of the contributions to inertia, and large values did not necessarily correspond to high associations between provinces and crimes. Whereas, the case of Table 6-4 provides a distinct separation between positive and negative associations.

A graph of attractions can be produced by utilizing the one dimensional approximations for the rows and columns from the correspondence analysis output of matrix $\boldsymbol{X}$. These two sets of points (principal coordinates) are then arranged in two separate columns in descending order, and subsequently positioned next to each other. The distances between the row points and between column points have no meaningful interpretation in the graph of associations. The one dimensional coordinates are used only as a reference for the ordering of the points, and the distances between the subsequent row, and between the subsequent column points are set as constants in the graph of attractions. If the row and column point of Figure 6-1 were orthogonally projected onto the first dimension, the order of which the row and column points were arranged from left to right would provide the ordering of the one dimensional CA approximation required by the graph of associations. Once the two sets of ordered points are plotted next to each other, the row and column points are then connected with lines representing positive associations. A *cut-off* value can be set so that only associations of a certain magnitude are viewed in the figure.

Figure 6-4 and Figure 6-5 are graphs of the attractions between the provinces and crimes of the 2004 South African crime data, as provided for in Table 6-4. Note that “attractions” refer to the positive associations of Table 6-4. The figures draw lines between each of the provinces and crimes for which there are positive associations (attractions). Figure 6-4 only displays the two largest attractions for the 2004 South African crime data, as the *cut-off* value is set at 0.9. Both Gauteng and the Western Cape are noted for having high associations to carjacking and drug-related crimes respectively. This result concurs with the findings of Chapter 4, where the aforementioned provinces were noted for their high rates of carjacking and drug-related crimes. Although only the ordering has been preserved from the CA output, the provinces and crimes with high attractions will tend to be positioned near one another in Figure 6-4 and Figure 6-5. However, only by interconnecting the points with lines representing the attractions, can a measure of the magnitude of the association be portrayed in the graph.

Figure 6-5 displays all the attractions between provinces and crimes for the South African crime data, where the *cut-off* value has been set to zero. In this case, all the lines of attractions are plotted. It is apparent that the figure can quickly become incomprehensible when there are a large number of attractions. However, with closer inspection it is possible to single out all of the attractions. It is of interest to note that both KwaZulu-Natal and the Eastern Cape stand out, as they have 6 and 7 associations respectively. Additionally, it can be determined that the only two provinces with attractions to drug-related crimes are Gauteng and the Western Cape. Similarly, only Gauteng and KwaZulu-Natal have attractions to both robbery with aggravating circumstances and carjacking. The aforementioned attractions were noted in previous chapters.

Table 6-4*: Association rate matrix for the 2004 South African crime data*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	0.50	-0.29	-0.25	0.60	-0.31	-0.09	-0.26	-0.35	-0.15
Tso	0.05	0.06	-0.12	0.06	0.31	0.05	0.15	-0.04	-0.09
Atm	0.05	-0.20	0.01	0.47	-0.25	-0.01	-0.25	0.66	-0.39
Agb	0.37	0.06	-0.19	-0.18	0.27	0.25	0.26	0.59	-0.19
Ast	-0.19	0.39	0.01	-0.15	0.21	-0.09	-0.06	0.05	0.09
Rac	-0.36	-0.47	0.69	0.20	-0.54	-0.15	-0.27	-0.74	-0.38
Ars	0.46	-0.09	-0.10	0.08	0.66	0.10	0.07	-0.11	-0.47
Bnr	-0.04	0.07	-0.14	-0.03	0.59	-0.17	0.32	0.27	-0.04
Brp	0.02	-0.05	0.04	-0.06	-0.13	0.19	-0.04	-0.20	0.02
Ilf	0.05	-0.59	-0.05	0.90	-0.35	-0.21	-0.39	-0.76	-0.13
Drg	-0.09	-0.29	-0.53	0.38	-0.62	-0.68	-0.10	-0.09	1.17
Crj	-0.64	-0.81	1.16	0.31	-0.71	-0.39	-0.69	-0.99	-0.57

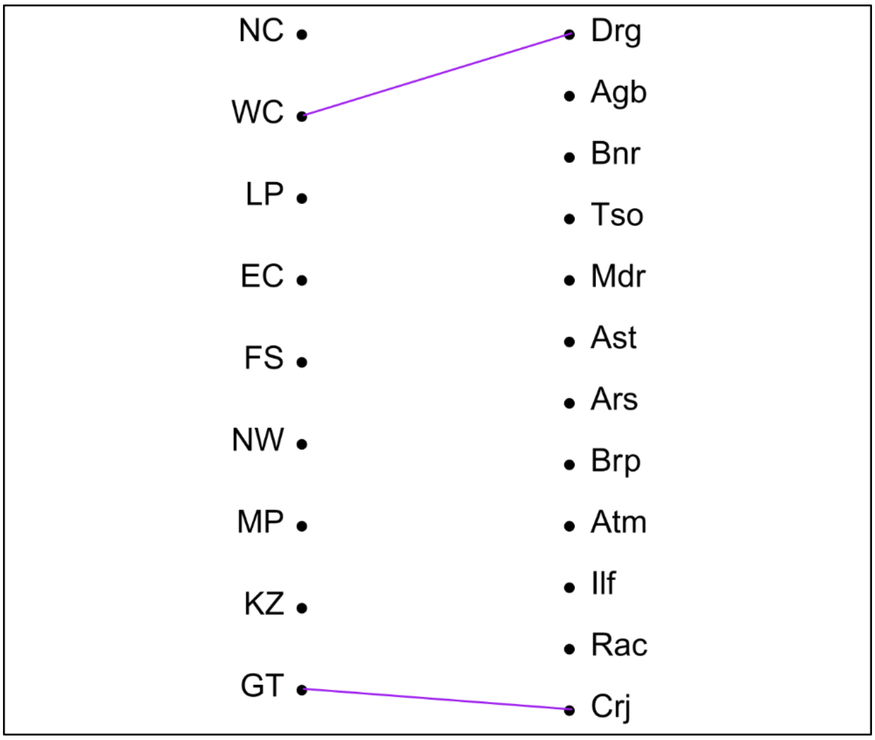

***Figure 6-4**: Graph of attractions for the 2004 South African crime data, displaying only the two largest attractions, cut-off set at 0.9.*

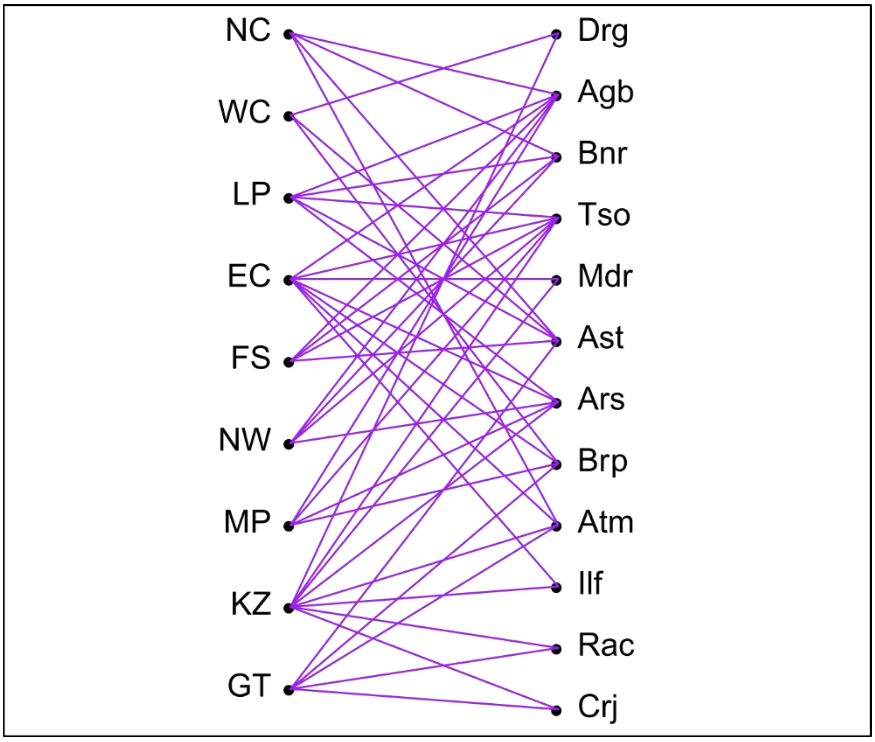

***Figure 6-5**: Graph of attraction for the 2004 South African crime data, with all attractions displayed, cut-off set at 0.*

It is of interest to note that the crimes which have few attractions between themselves and the provinces were situated in the peripheral of Figure 6-1. Conversely, those crimes which have many attractions between themselves and the provinces are located close to the origin or the second quadrant of Figure 6-1. Additionally, many of the associations identified in Figure 6-1 are also accounted for in Figure 6-5. However, interpretability is lost in the graph of attractions due to the fact that there is no reference to the magnitude of the attractions. Imposing an association value cut-off can ensure that only attractions above a set level are plotted, as in Figure 6-4. However, there is still no accurate measurement of the attraction values. Conversely, the distances and directions of the principal coordinates in the CA maps provide a reference to the magnitude of associations between the provinces and crimes.

It can thus be concluded that the association matrix and its respective figures are of interest to note in the study of two-way tables. However, as the CA map provides a more interpretable setting for the associations between the provinces and crimes, this study will not make use of the association rate matrix and its respective figures.

6.1.3 Analysis of Inertia 2004-2013

Subsection 6.1.1 inferred the importance of inertia in correspondence analysis. This section provides an analysis of the total inertia and its contributions for the years 2004-2013. The purpose of such analysis is to provide a broad evaluation of the CA maps which appear in this study. Each of the two-way tables of the South African crime data for the reported period is of size 12 × 9. It follows that the dimensionality of each two-way table for the reported period is 8. As mentioned in Section 5.2, the maximum total inertia of a two-way table is equal to the dimensionality of the table. Therefore, for each two-way table of the South African crime data and its associated CA display, there is a maximum total inertia value of 8. From the example in Subsection 6.1.1, the 2004 South African crime data produced a total inertia value of 0.0856. A total inertia value of 0.0856 is negligible when compared to its potential value of 8. Even though the total inertia value is minimal for the 2004 crime data, there is clearly variation in the profiles in Figure 6-1. Before further CAs are produced, an analysis of the total inertia for all years is warranted.

Figure 6-6 plots the total inertia values for each of the two-way South African crime data tables for the 2004-2013 period. There is a clear increase over the first seven years, where after there is a sharp decrease in the total inertia to a low point in 2013. The increase is due to at least one of the provinces diverging from the average profile over this time period, after which the provinces' profiles quickly started converging. For the first seven years, it is then expected that the profile points in the figures will become more and more dispersed over this period. Where after, the profile points are expected to begin converging to an even tighter cluster than in 2004.

Although the total inertia is of great importance in correspondence analysis, possibly of more importance is the proportion of this value accounted for in the CA display. In Subsection 6.1.1, it was noted that Figure 6-1 accounted for 83.2% of the total inertia of the 2004 South African crime data. The percentage of total inertia accounted for in a two-dimensional display can be used as a measure of the quality of the display. Ideally, it is desirable to account for 100% of the total inertia, but as the two-dimensional display is only an approximation of the eight-dimensional two-way table there is an expected loss of information. Figure 6-7 plots the cumulative inertia value that is accounted for by the number of dimensions comprising the display for each of the South African crime data tables (2004-2013).

As higher dimensional spaces are difficult to represent on paper, the analysis is limited to two-dimensional displays. In a majority of the plots in Figure 6-7, the first dimension accounts for at least 50% of the total inertia. It then appears that together with the second dimension, at least 80% of the total inertia has been accounted for. With exception to the year 2013, it would appear that around 95% of the total inertia can be accounted for in a three-dimensional plot. It can be concluded from the cumulative inertia plots that a two-dimensional display will be adequate for the analysis in this study. However, it is evident that the third dimension provides additional information not accounted for by the first two dimensions. Hence, the two-dimensional approximation must be interpreted with caution, and not regarded as a perfect representation of the data.

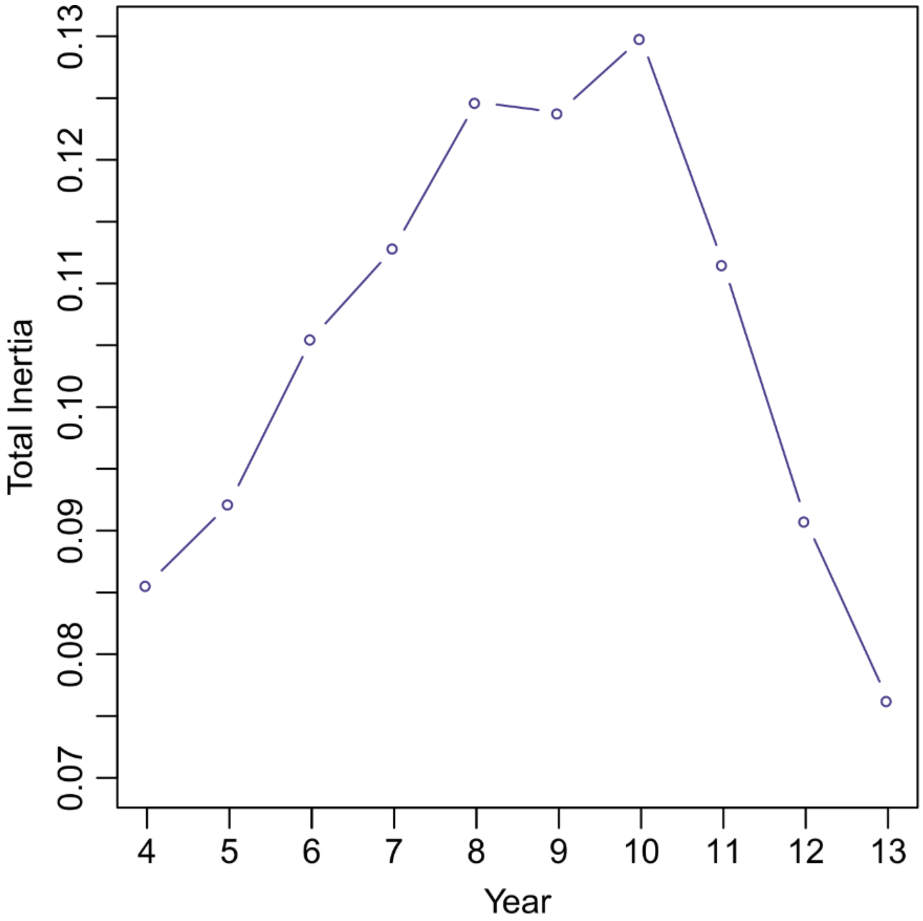

Figure 6-6*: The calculated inertia of the South African crime data for the period 2004-2013.*

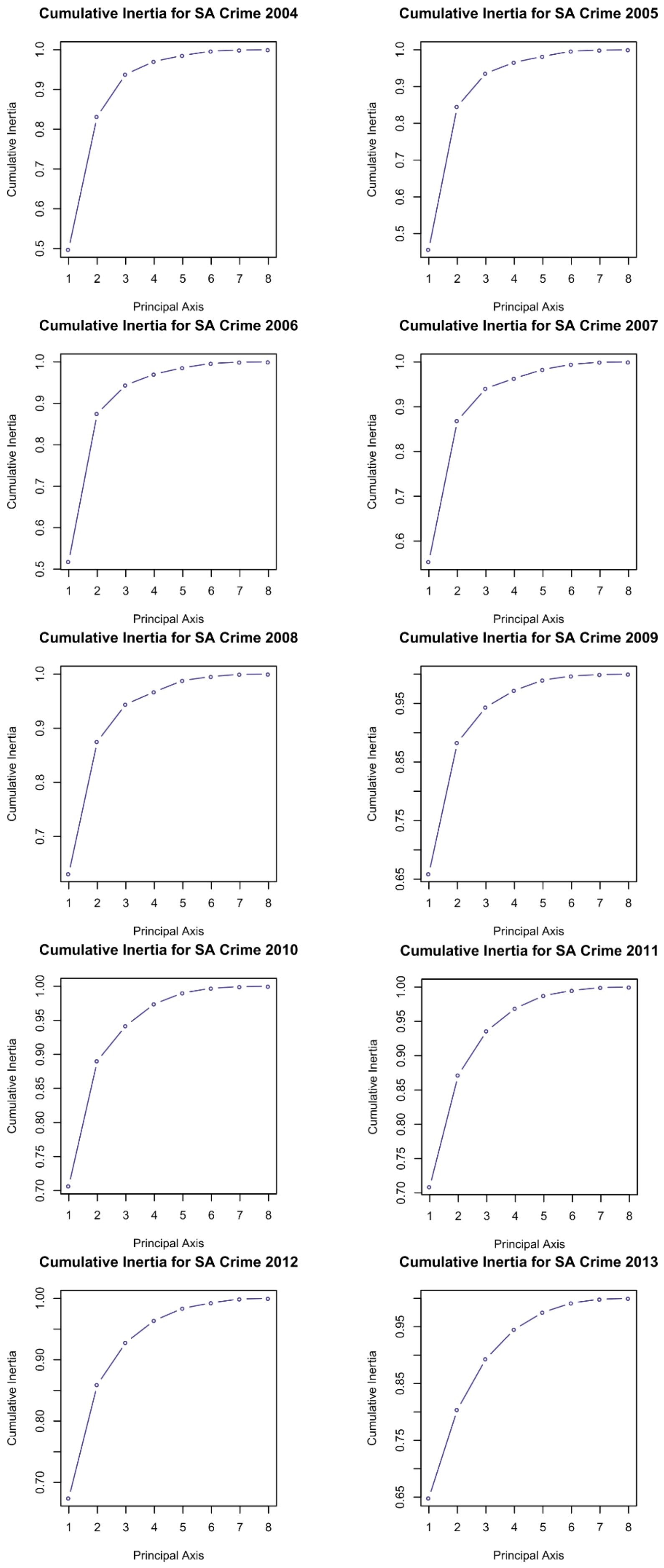

Figure 6-7: *Cumulative inertia plots per year of the South African crime data.*

6.1.3.1 *Correspondence Analysis for the years 2004-2013.*

The CA of the 2004 crime data was analysed in Subsection 6.1.1. This section provides a summary of the CA results for the succeeding nine years of the South African crime data. The CA displays for the years 2006-2009 provide similar results to that of 2005. Due to this, not all displays are included in this subsection, but are available to the reader in Appendix B. Rather, several of the most representative displays will be presented here. Following from the results of Figure 6-6 and Figure 6-7, it would be expected that there would be an initial outward spreading of the profile points as the inertia increases over the first seven years. Where after, the profile points would be expected to converge as the inertia reduces.

Table 6-5 and Table 6-6 provide the individual contributions to the total inertia (in permills) by each province, and by each of the crime types over the reported period respectively. As in Section 6.1.1, conditional highlighting has been utilized in both tables to highlight any cell which contributes 10% or more to the total inertia for a specific year. In Table 6-5 there are clearly three crime types which make up the majority of the total inertia contribution. These crimes are: assault with the intent to cause grievous bodily harm, robbery with aggravating circumstances, and drug-related crime. Assault with the intent to cause grievous bodily harm appears to slowly decrease from 2004 to 2010. After which, it experiences a sharp increase from 2011 to 2013. Robbery with aggravating circumstances experiences a continuous decrease between 2004 and 2012, and only experiences a marginal increase in 2013. Drug-related crime, which is consistently the largest contributor to the total inertia, experiences large increases up until 2010, where subsequently it experiences a sharp decrease. Another noteworthy progression is a consistent increase from 0.9% in 2004 to 10% by 2013 in total sexual offenses.

***Table 6-5**: Individual crime contributions, in permills, to the total inertia per year, for the period 2004-2013.*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Mdr	24	26	20	18	17	16	14	17	22	27
Tso	9	15	16	15	15	16	24	34	65	100
Atm	21	20	17	16	14	12	7	6	4	7
Agb	146	144	146	133	120	120	113	126	161	184
Ast	59	50	38	43	42	38	42	51	52	68
Rac	281	219	207	183	145	124	96	78	75	80
Ars	7	7	7	6	6	6	6	6	9	11
Bnr	19	21	15	19	26	23	21	29	29	40
Brp	16	18	16	15	14	18	22	24	27	28
Ilf	30	26	25	22	17	25	25	24	20	26
Drg	314	382	429	463	529	547	581	563	489	372
Crj	75	73	65	66	55	56	48	45	46	57

Table 6-6: *Individual province contributions, in permills, to the total inertia per year, for the period 2004-2013.*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
EC	83	98	94	86	68	76	71	83	108	159
FS	62	64	54	62	55	48	57	75	89	116
GT	295	236	237	226	189	204	192	138	118	124
KZ	94	98	91	69	55	53	43	38	33	29
LP	78	75	68	55	51	43	44	63	69	84
MP	46	53	49	53	53	50	41	49	51	65
NW	26	21	22	21	19	19	21	29	40	45
NC	66	64	60	48	47	43	41	45	58	78
WC	250	290	324	380	463	465	490	481	434	300

Table 6-6 provides the contributions to the total inertia by each province for the reported period. From the table it is evident that the Western Cape and Gauteng make up a large majority of the total inertia per year. The combined contribution peaks in 2010 at 68%, then quickly declines to a low of 42% by 2013. It is of interest to note that the contributions of the Western Cape and Gauteng do not always increase and decrease simultaneously. Gauteng appears to have a steady decreasing contribution whereas the Western Cape only has notable decreases in the final two years. There is, however, a combined increase until 2010 due to the larger increases by the Western Cape. Additionally, the Eastern Cape is highlighted for the years 2012 and 2013 as it has exceeded the 10% contribution barrier for these years. Finally, the Free State is highlighted only for the year 2013. All other provinces make up the remaining contribution to the total inertia.

Table 6-5 and Table 6-6 have highlighted increases in contributions to the total inertia by crime type and province separately. By making use of CA displays, as well as the calculated contributions to the total inertia for each year of the South African crime data, how the provinces deviate away from the mean profile can be analysed in detail. The first notable change in the CA display occurs between the years 2004 and 2005. The CA of the 2005 South African crime data is presented in Figure 6-8. At first glance the positioning of the points appear to have changed a considerable amount compared to that of Figure 6-1, although relative positioning is still very similar. It then follows that Figure 6-8 and Figure 6-1 can be directly compared if Figure 6-1 is reflected about the Y-axis and rotated 65 degrees. The function `CArotated` contains the arguments reflect and rotate which can be utilised to manipulate the CA plot. Figure 6-9 is produced by inputting the CA output from the 2004 crime data into the `CArotated` function, and subsequently setting the arguments to `reflect=”y”` and `rotate=60`. As mentioned in Subsection 5.3.4, it is possible to reflect and rotate these displays without any loss of information. This being the case, it then appears that there is not a notable amount of change between the displays of 2004 and 2005 after manipulation.

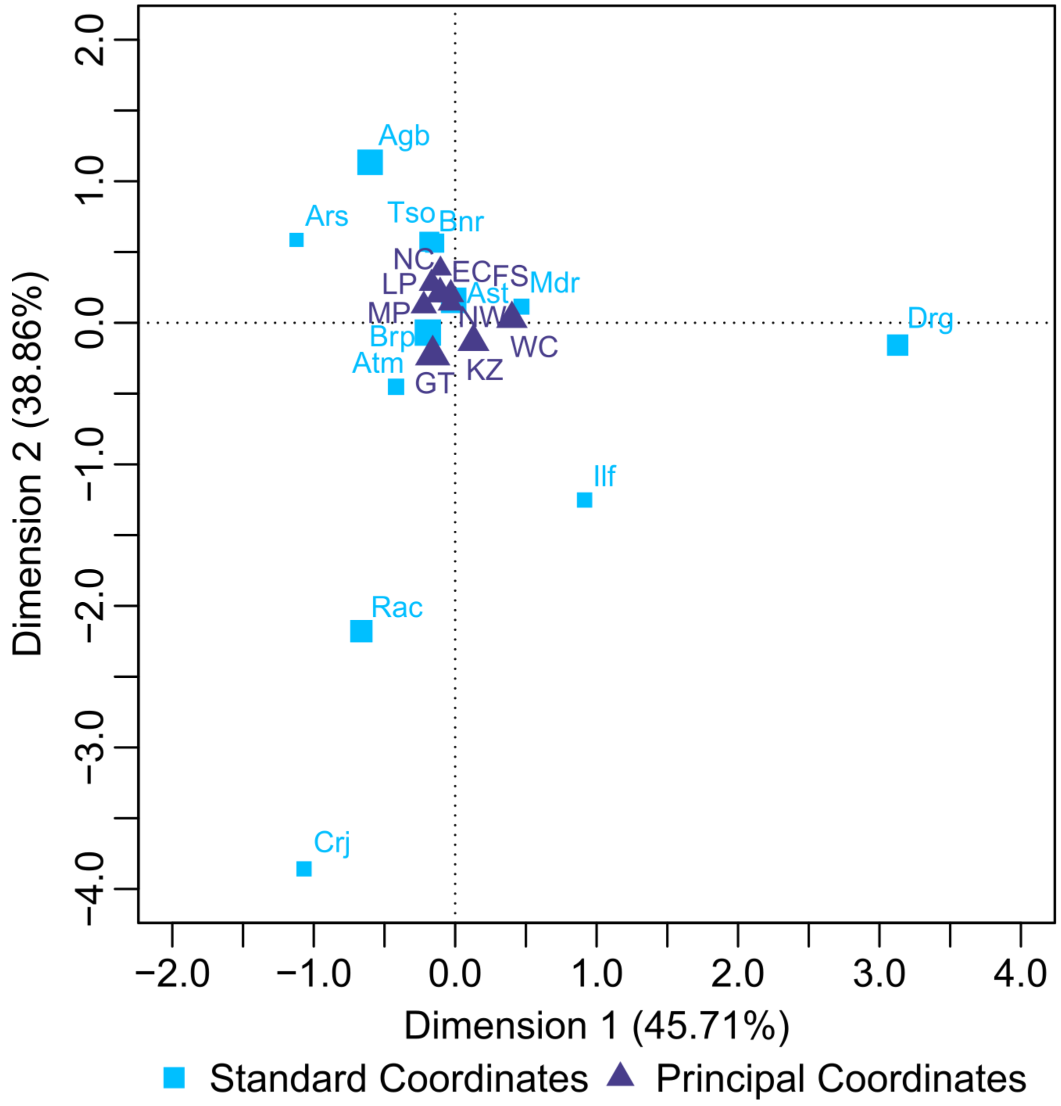

Figure 6-8*: An asymmetric CA display of the 2005 South African crime data.*

Table 6-7*: Cell contributions to the total inertia for the 2005 South African crime data (in permills).*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC	Sum
Mdr	9	1	4	9	1	1	1	0	0	**26**
Tso	5	0	4	0	3	0	1	0	2	**15**
Atm	0	1	0	9	1	0	1	3	6	**21**
Agb	36	1	17	15	11	10	5	33	17	**145**
Ast	8	28	0	6	5	0	2	0	1	**50**
Rac	18	19	116	8	18	2	2	16	20	**219**
Ars	2	0	0	0	2	0	0	0	3	**7**
Bnr	0	0	2	0	11	1	4	1	1	**20**
Brp	1	0	2	4	1	6	0	3	0	**17**
Ilf	0	3	0	17	1	1	1	2	0	**25**
Drg	13	3	52	27	19	30	0	2	235	**381**
Crj	7	7	39	3	4	1	3	4	6	**74**
Sum	**99**	**63**	**236**	**98**	**77**	**52**	**20**	**64**	**291**	**1000**

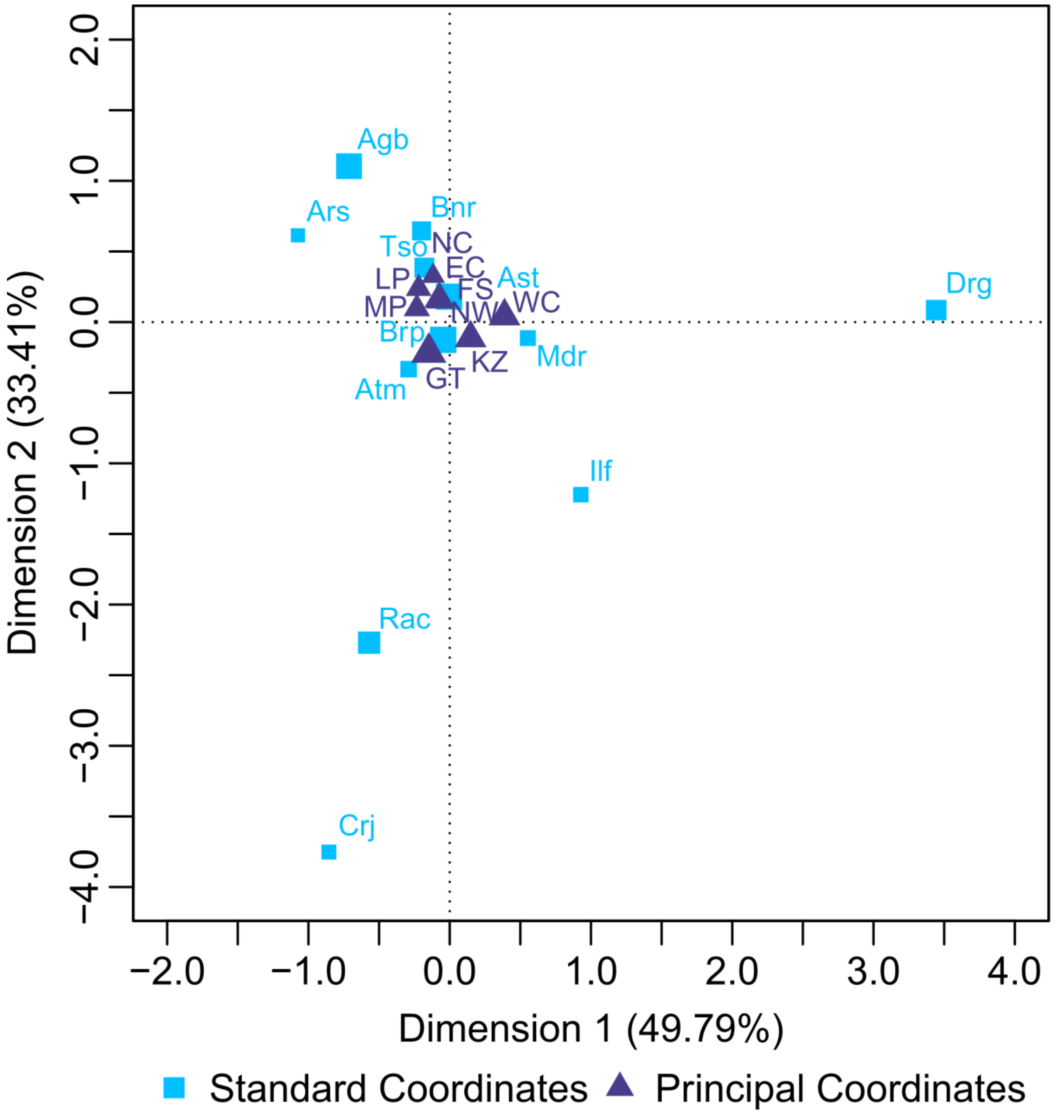

***Figure 6-9**: The CA of the 2004 South African crime data after being subjected to a reflection about the Y-axis, and being rotated by 60°.*

By making use of both Figure 6-8 and Table 6-7 a detailed analysis can be conducted on the 2005 data. Firstly, the graph appears to show a marginal increase in spread with regards to the profile points, a result of the increased inertia. Where previously overlapping, there is now an increase in the dispersion of the Eastern Cape, Free State, and North West Province. Additionally, the Western Cape's profile point has been drawn further out towards drug-related crime. This change is evident in Table 6-7, where the contribution to total inertia by drug-related crime in the Western Cape increased from 186 (18.6%) in 2004 to 235 (23.5%) in 2005. Furthermore, there is a considerable drop in the contribution from Gauteng and robbery with aggravating circumstances, which is not well represented in Figure 6-8. As the relative positioning of both points do not appear to have changed between 2004 and 2005, the decrease in the contribution from Gauteng could in fact be due to the larger increase in the contribution from the Western Cape, and not a direct result of Gauteng itself.

This is a likely scenario as both points are drawn further from the centroid, however the change for the Western Cape is greater, due to its large association to drug-related crime.

Over the next several years similar patterns continue in the CA displays. For this reason, some displays are excluded from the discussion, but are included in Appendix B. Over the reported period, what is most evident is the Western Cape's progression in the direction of drug-related crime. Simultaneously, there is a movement of the drug-related crime vertex point towards the centroid. Gauteng's profile point does not move significantly over the years, but the decrease in its contribution to inertia appears to be largely attributed to the inward movement of the robbery with aggravating circumstances vertex point.

Figure 6-10 provides a CA display for the 2010 South African crime data. The associated contributions to the total inertia for the 2010 data are provided in Table 6-8. The vertex points which were previously positioned nearer to the centre of the graph, have been redistributed about the centroid. The most notable changes are those of murder, and total sexual offenses. As previously stated, it is evident that the profile point of the Western Cape is clearly drawn outwards towards the drug-related crime vertex. The vertex point of drug-related crime has shifted inwards and its size (mass) has also notably increased. The Western Cape's association with drug-related crime now comprises almost 40% of the total inertia. Drug-related crime in total accounts for 58.2% of the total inertia of South African crime data for 2010. Many of the profile points are, again, positioned closely together in Figure 6-10. In general, the profile points have kept a similar structure to that of previous years. However, the vertex points now lie closer to the second dimension than in previous displays. In Figure 6-10 the first dimension accounts for 70.69% of the total inertia. There is a continuous increase in this percentage from the years 2004, where the first dimension accounted for 49.79%, to a high in the year 2011, where the first dimension accounts for 70.90% of the total inertia. It then follows that for the 2011 crime data, that a one dimensional display would provide an adequate representation of the data. However, there is still a considerable amount of inertia accounted for by the second dimension.

From Table 6-8 it is clear that a single cell contributes significantly to the total inertia of the 2010 crime data, that being the interaction of the Western Cape and drug-related crime. As previously noted in Table 6-7, the cell of Gauteng and robbery with aggravating circumstances had a notable contribution. The contribution of Gauteng and robbery with aggravating circumstances in Table 6-8 is less than half the value for the year 2005. By observing the row and column totals of Table 6-7 and Table 6-8, it is evident that between the years of 2005 and 2010 the contribution of robbery with aggravating circumstances has dropped from 21.9% to 9.6%. Similarly, assault with the intent to cause grievous bodily harm experienced a marginal decrease to 11.4% from 14.5%. Conversely, drug-related crimes experienced a large increase from 38.2% to 58.1%. With regards to the provinces' contributions, Gauteng decreases from 23.6% to 19.35%, and the Western Cape increased notably from 29% to 48.9%. Gauteng and the Western Cape remain the only two provinces with notable contributions to inertia.

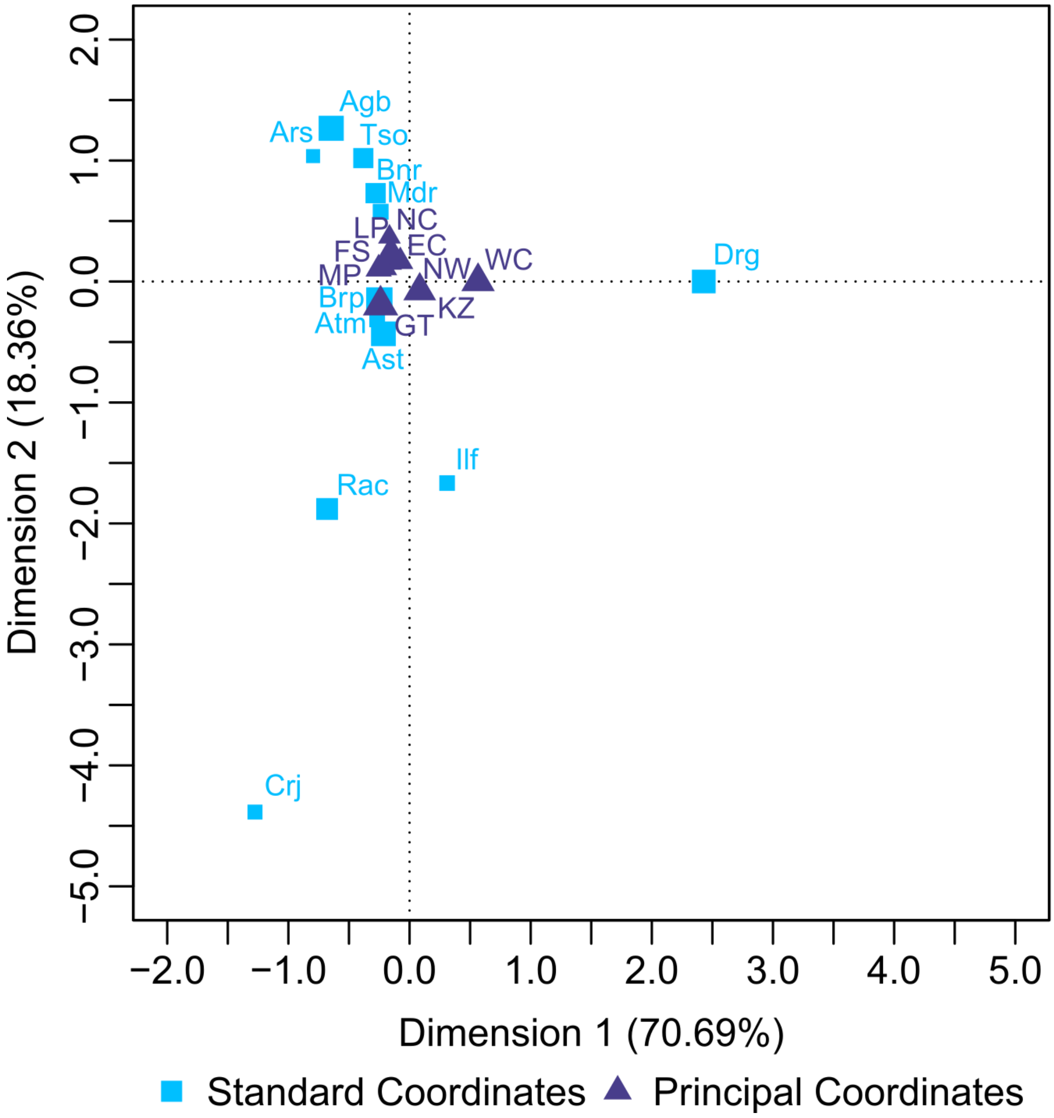

Figure 6-10*: An asymmetric CA display of the 2010 South African crime data.*

***Table 6-8**: Cell contributions to the total inertia for the 2010 South African crime data (in permills).*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC	Sum
Mdr	8	0	1	2	0	0	0	0	2	**13**
Tso	4	1	4	0	5	0	2	0	8	**24**
Atm	0	0	0	3	0	0	0	1	2	**6**
Agb	27	4	3	5	6	4	5	22	38	**114**
Ast	5	21	7	0	1	0	5	0	2	**41**
Rac	0	3	54	1	8	0	1	8	21	**96**
Ars	2	0	0	0	1	0	0	0	2	**5**
Bnr	1	1	0	1	11	2	2	0	2	**20**
Brp	0	1	6	4	0	6	0	2	4	**23**
Ilf	0	3	0	17	1	1	2	2	0	**26**
Drg	21	22	91	8	9	26	2	4	399	**582**
Crj	2	2	27	2	2	0	2	2	9	**48**
Sum	**70**	**58**	**193**	**43**	**44**	**39**	**21**	**41**	**489**	**998**

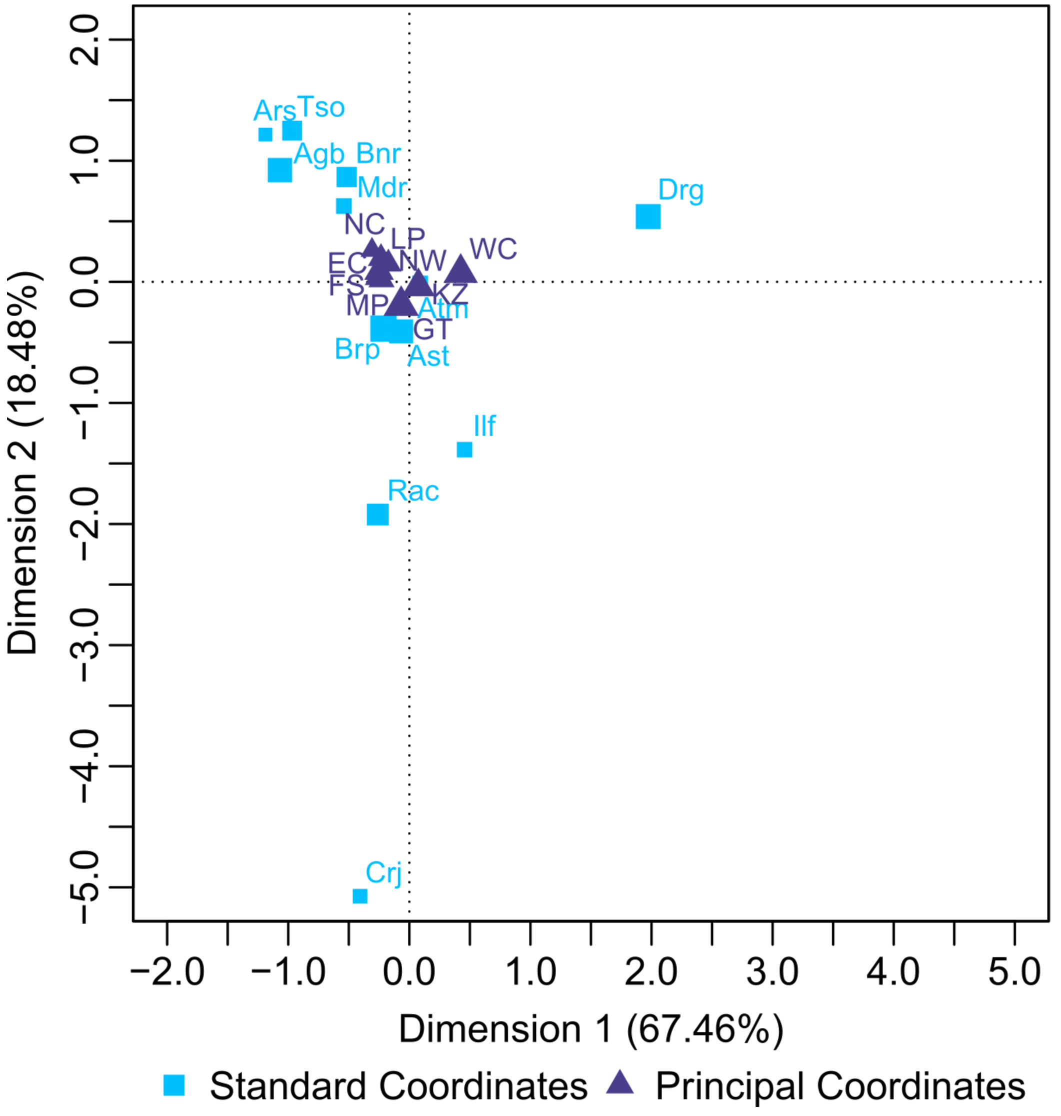

***Figure 6-11**: An asymmetric CA display of the 2012 South African crime data.*

Between the years 2010 and 2012 (Figure 6-11), there are several notable changes. It is evident that the profile points have contracted towards the centroid, as a result of a lower total inertia. Additionally, the vertex points of drug-related crime and carjacking appear to have migrated inward, and outward, respectively. Moreover, the points are positioned similar to that of Figure 6-10, but have experienced a minor clockwise rotation about the centroid. Accompanying this rotation is a marginal increase in the variation of the points relative to the second dimension. The first dimension now accounts for slightly less inertia than before. There is a loss of 2.61% on the first dimension but only a 0.12% gain on the second dimension, resulting in a display that performs marginally worse than Figure 6-10. The contributions to the total inertia value for the South African crime data are contained in Table 6-9. In Table 6-9, it is made evident that there are now three provinces which contribute more than 10% to the total inertia, these are: Eastern Cape, Gauteng, and the Western Cape. Similarly to Figure 6-9, there is only a single cell that has a contribution of more than 10%. The interaction of drug-related crime and the Western Cape has dropped from 399 (39.9%) to 324 (32.4%). Additionally, the overall contribution by drug-related crime has decreased from 582 (58.2%) to 490 (49%). Alternatively, the overall contribution of assault with the intent to cause grievous bodily harm has increased. Albeit not a considerable contributor, the overall contribution of total sexual offenses more than doubles between 2010 (2.4%) and 2012 (6.4%).

***Table 6-9**: Cell contributions to the total inertia for the 2012 South African crime data (in permills).*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC	Sum
Mdr	16	0	2	2	1	0	0	0	2	**23**
Tso	10	1	8	0	18	1	7	0	19	**64**
Atm	0	0	0	3	1	0	0	0	0	**4**
Agb	37	6	3	7	8	2	11	32	54	**160**
Ast	7	31	3	0	0	2	7	1	0	**51**
Rac	1	2	42	0	7	1	1	8	14	**76**
Ars	2	0	0	0	3	0	0	0	2	**7**
Bnr	0	1	2	1	12	4	3	1	4	**28**
Brp	1	0	5	1	0	13	0	1	6	**27**
Ilf	0	3	0	11	1	0	2	3	0	**20**
Drg	33	42	26	6	16	26	6	11	324	**490**
Crj	1	2	26	2	2	1	2	2	8	**46**
Sum	**108**	**88**	**117**	**33**	**69**	**50**	**39**	**59**	**433**	**996**

The final CA to be included in this section is that of the 2013 South African crime data (Figure 6-12). Although it appears that the points are reflected about the second dimension, the positioning of the points has a completely new structure as compared to previous figures. The points of Figure 6-12 have been reflected about the X-axis, producing Figure 6-13. Figure 6-13 is now more directly comparable with Figure 6-11. After reflection, there still appears to be a notable difference between all the previous displays and that of 2013. Similarly to the 2004 data, all provinces are now positioned close to the centroid (average profile) and the majority of the vertex points are creating a cloud-like shape around the centroid. Carjacking is positioned a substantial distance away from all other points, followed by robbery with aggravating circumstances and drug-related crime. With the exception of attempted murder and residential burglary, which are positioned at the centroid, all other points form a fairly consistent cloud-like shape around the centroid. Following from Table 6-10, assault with the intent to cause grievous bodily harm and drug-related crime remain the two largest contributors to the total inertia. However, drug-related crime now accounts for 37.2% of the total inertia, a large decrease from 49% in 2012.

By inspecting Table 6-10, it is evident that the contributions to the total inertia by the different provinces is becoming more evenly distributed. There are now four of the nine provinces with contributions greater than 10%. In previous years the majority of the total inertia was attributable to Gauteng and the Western Cape. As the contributions to total inertia by the Western Cape and Gauteng decrease, the contributions by the remaining provinces increase. This is most likely due to the overall increase in drug-related crime in all provinces, and not only in the Western Cape as it was in earlier years. Additionally, the decrease in robbery with aggravating circumstances in Gauteng has significantly lowered its contribution. As drug-related crime grew in all provinces, the vertex point of drug-related crime was drawn closer towards the centroid. The Western Cape remains the largest contributor to the drug-related crime problem. However, its pronounced association is not as extreme as it was at the beginning of the reported period.

From the initial correspondence analysis, it can be concluded that the profiles of the individual provinces are not static. Rather, the profiles behave in a dynamic manner. Whilst, by 2013 the vertices had new positions, the profile points had returned to positions similar to that of the 2004 data. This could act as an indication of cyclical patterns in the South African crime data. Over the reported period it is evident that there is aggressive growth in drug-related crime, particularly in the Western Cape. However, the profiles quickly find a new equilibrium as they converge towards a uniform composition.

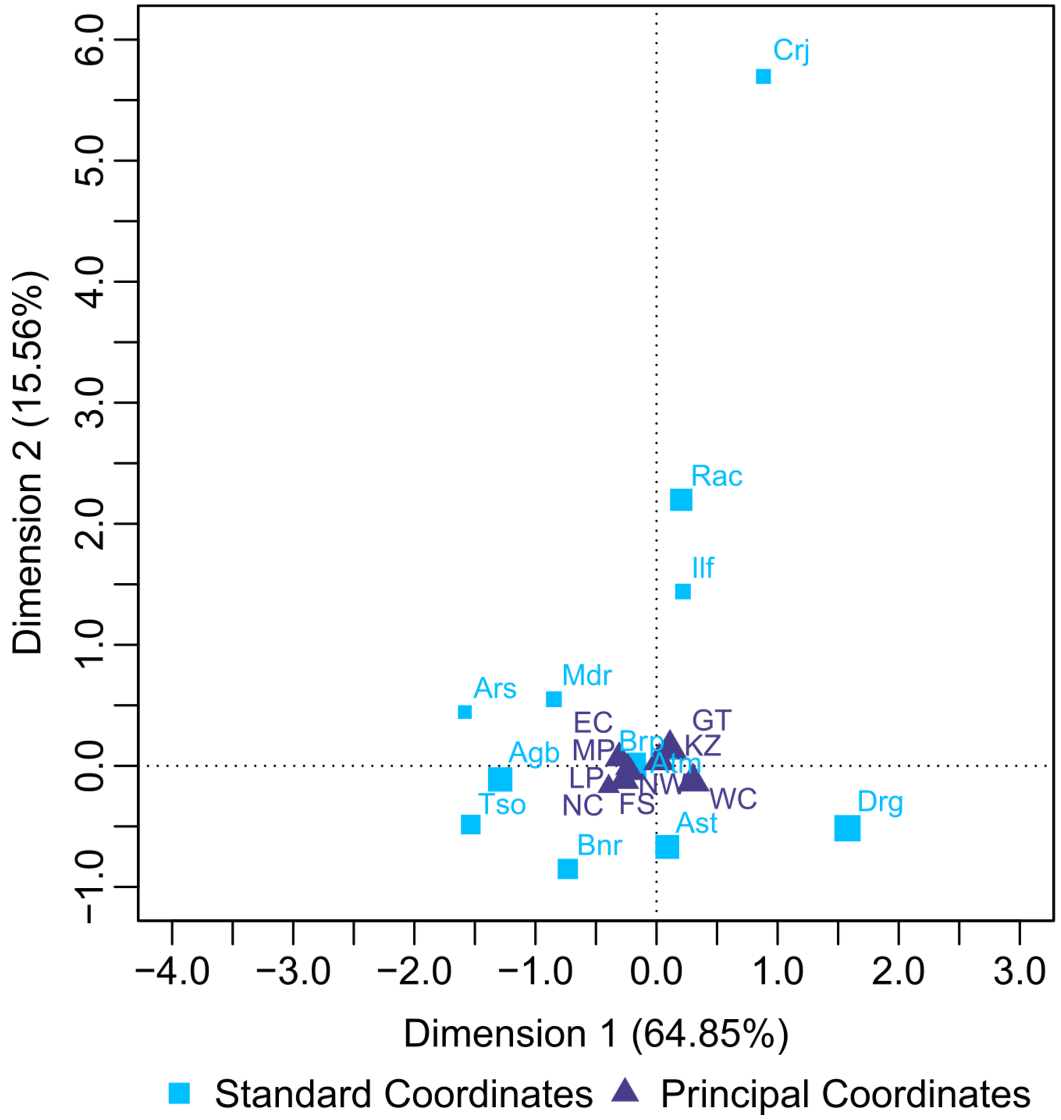

***Figure 6-12**: An asymmetric CA display of the 2013 South African crime data.*

***Table 6-10**: Cell contributions to the total inertia for the 2013 South African crime data (in permills).*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC	Sum
Mdr	19	0	4	2	1	0	0	0	1	**27**
Tso	22	2	21	1	23	2	7	0	21	**99**
Atm	0	0	1	3	1	0	0	1	0	**6**
Agb	47	9	12	2	4	2	14	42	52	**184**
Ast	8	47	0	2	0	2	6	1	2	**68**
Rac	2	6	41	0	5	2	2	9	13	**80**
Ars	5	0	1	0	2	0	0	0	3	**11**
Bnr	0	1	5	2	19	6	2	2	3	**40**
Brp	1	0	0	0	1	21	1	0	3	**27**
Ilf	1	3	0	16	2	0	2	2	1	**27**
Drg	53	44	5	0	22	30	8	18	192	**372**
Crj	1	3	35	1	3	1	2	2	9	**57**
Sum	**159**	**115**	**125**	**29**	**83**	**66**	**44**	**77**	**300**	**998**

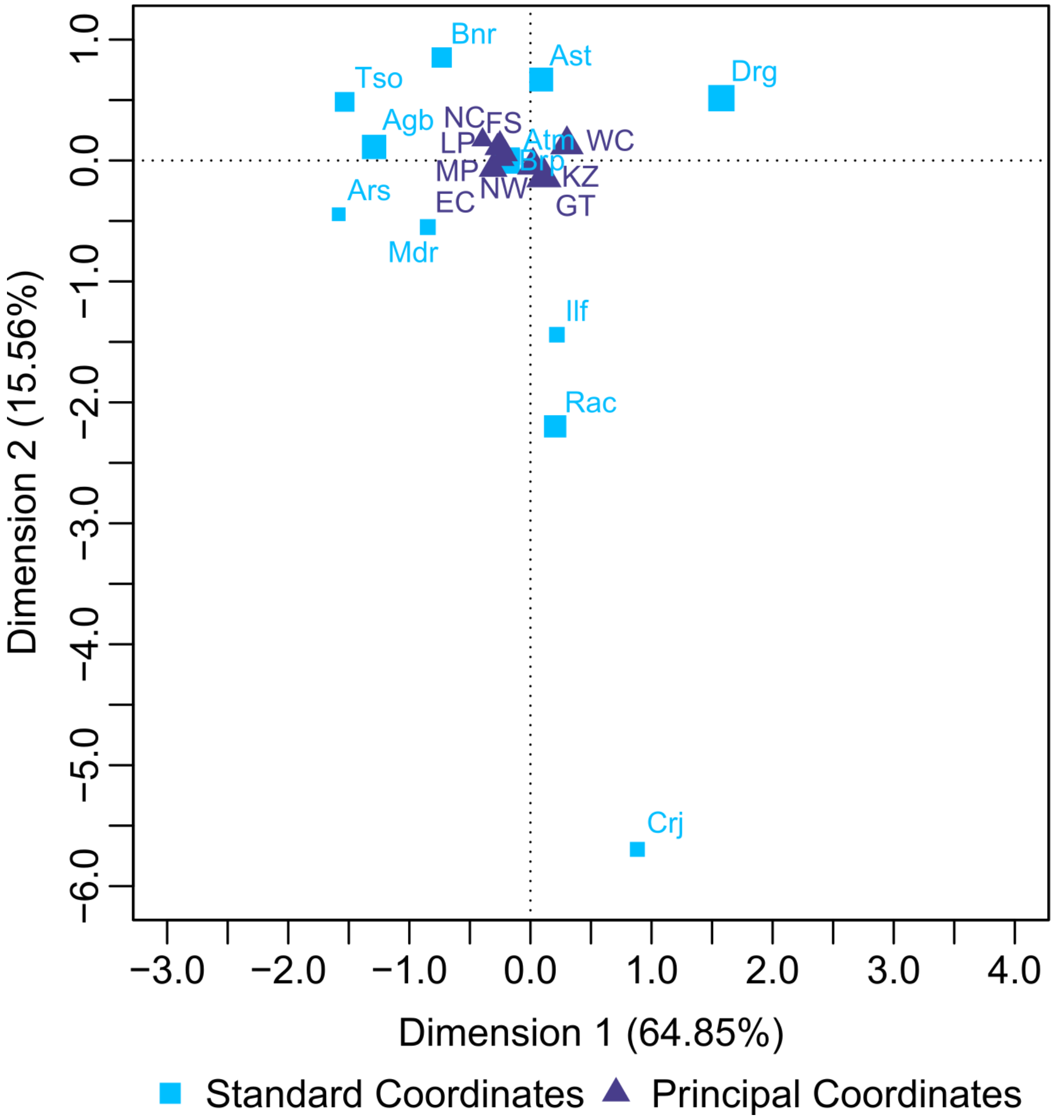

***Figure 6-13**: The 2013 South African crime data after being subjected to a reflection about the X-axis.*

6.2 Correspondence Analysis Biplot

Section 6.1 revealed important aspects of the South African crime data. However, shortfalls of the correspondence analysis were made apparent. The greatest of which is the inability to easily manipulate the CA map. It is often found in asymmetric CA maps, that the principal coordinates are bundled around the origin, whereas the standard coordinates are located in the peripheral. This creates a display which is, at times, difficult to interpret. The ability to scale the coordinates is therefore a desirable property, as it creates a more interpretable graph. The correspondence analysis biplot (CA biplot) addresses this issue by replacing the standard coordinates with calibrated axes, and allowing for Lambda-scaling where needed and applicable. The reader is referred back to Subsection 5.3.3 for details on Lambda-scaling.

The `cabipl()` function from the *Understanding Biplots* text can implement the theory presented in Section 5.4 (Gower *et al.*, 2011). Additionally, the `cabipl()` function can easily account for Lambda-scaling, reflecting, and rotating of a CA biplot. The function call, and its associated arguments, can be found in Chapter 7 of *Understanding Biplots* (Gower *et al.*, 2011:306). By making use of these arguments, graphics can be constructed that are more readily interpretable.

As mentioned in Subsection 5.4.1, there are several variations of the independence model which can be approximated in a CA biplot. For uniformity throughout, the CA biplots in this section are restricted to approximating the contributions to the Pearson's χ^2. This, coupled with the ability to simultaneously display calibrated axes, formulates an effective method for interpreting the figures. The argument `PearsonRes.scaled.markers=TRUE` of the `cabipl()` function provides the ability to calibrate the axes in terms of the square root of the Pearson residuals. Following from equation 5-8, the elements of the weighted deviations $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$ can be written as:

$$\frac{\left(x_{ij}-\frac{x_{i.}x_{.j}}{n}\right)}{\sqrt{x_{i.}x_{.j}}} = \frac{\frac{1}{\sqrt{n}}\left(x_{ij}-\frac{x_{i.}x_{ij}}{n}\right)}{\sqrt{\frac{x_{i.}x_{.j}}{n}}}, \qquad (6\text{-}4)$$

where $x_{i.}$, $x_{.j}$, and n are the row, column, and grand totals of matrix $\boldsymbol{X}$ respectively. Recalling that the Pearson's χ^2 statistic is calculated as:

$$\chi^2 = \sum\frac{(observed-expected)^2}{expected} = \sum_i\sum_j\frac{\left(x_{ij}-\frac{x_{i.}x_{ij}}{n}\right)^2}{\frac{x_{i.}x_{.j}}{n}}, \qquad (6\text{-}5)$$

it then follows that $\sqrt{n}$ times the elements of the weighted deviations gives the square root of the contributions the χ^2 statistic (Gower *et al.*, 2011:291). The argument is used in the construction of the CA biplots in order to allow the actual amount, and direction, of deviation to be easily detected in the figures. Figure 6-14 provides an example of the CA biplot for the 2004 South African Crime data. The transpose of the 2004 data matrix was used as input into the `cabipl()` function. The figure is constructed from the first two columns of $\boldsymbol{U\Sigma}^{-½}$ and $\boldsymbol{V\Sigma}^{-½}$, derived from the SVD of $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$. The axes have been calibrated directly in terms of the Pearson residuals by incorporating the factor $\sqrt{n}$ in the calibrations, as implemented by the argument `PearsonRes.scaled.markers=TRUE`. Setting the argument `predictions.sample=9` and `ort.lty=c(2,1)` plots the predicted sample values for the Western Cape.

Figure 6-14 can be compared to Figure 6-1, as both are constructed from the 2004 South African crime data. The crime type points are no longer displayed in the figure, and have rather been replaced by calibrated axes representing the various crimes. Additionally, the profile points are no longer clustered by the centroid (average profile), as in Figure 6-1. This is a result of the implementation of Lambda-scaling. The profile points have been drawn away from the centroid to make the interpretation of the biplot easier. Additionally, the positioning of the profile points have remained relatively unchanged. However, the Northern Cape appears closer to the centre than in Figure 6-1. Although Figure 6-14 can be compared to Figure 6-1, it must be stressed that their interpretations differ. In the CA figures, the χ^2 distance between profiles and vertices was displayed. Alternatively, in Figure 6-14 the amount by which each profile deviates from the average profile is expressed.

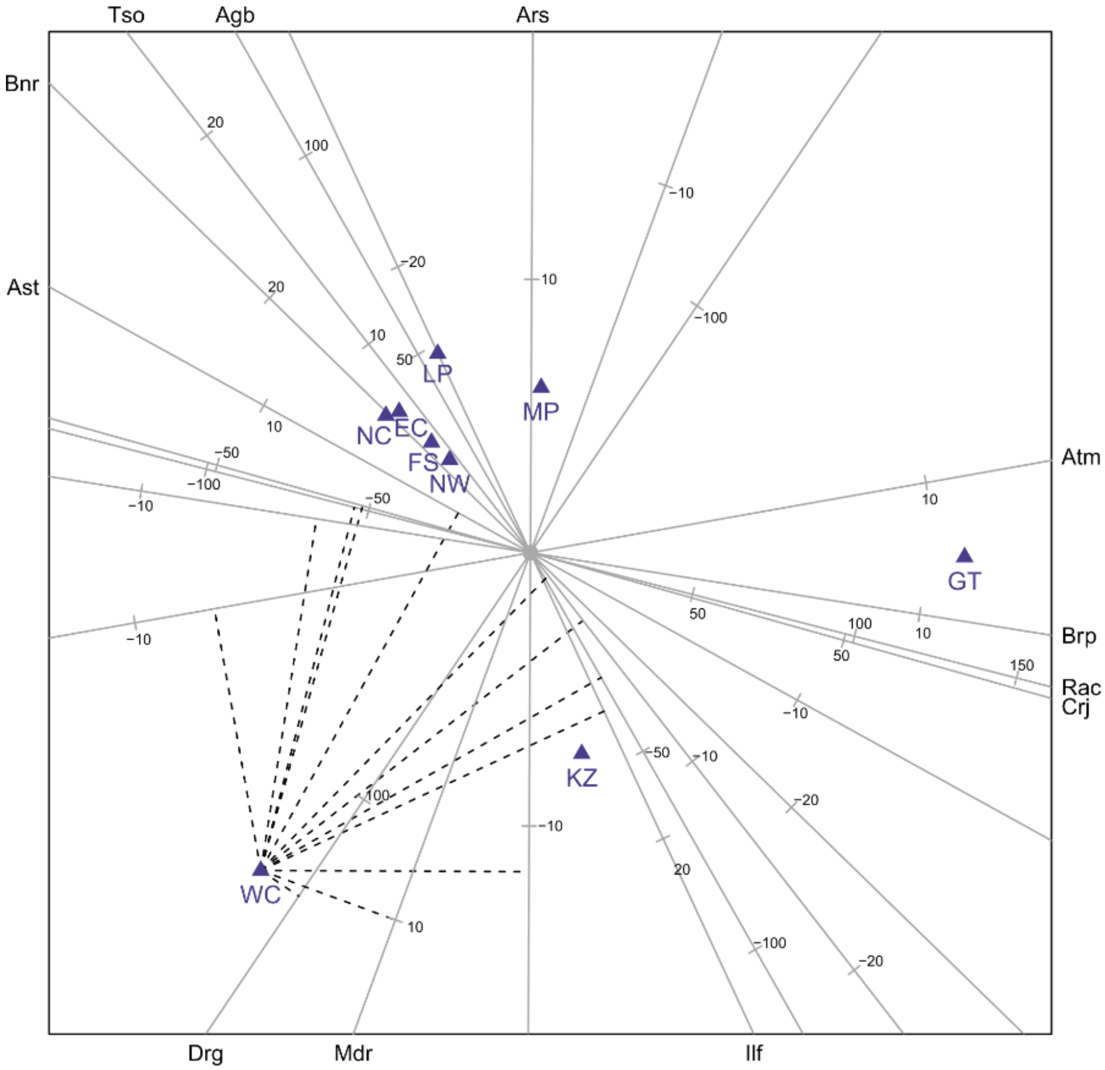

***Figure 6-14**: A two-dimensional CA biplot of the 2004 South African crime data set in terms of the contributions to the Pearson residuals*

For means of demonstration, the predicted values for the Western Cape are plotted in Figure 6-14. The Western Cape's predictions are represented by the dashed lines projected orthogonally onto the axes in Figure 6-14. Just as with the Western Cape, the predicted values can be displayed for all of the nine provinces in a CA biplot. From inspection of the graph, it is approximated that the Western Cape has a drug-related crime value in 2004 which is greater than the average profile, as the projection of the profile point appears on the positive end of the drug-related crime axis. Recalling that the values depicted here are the contributions to the χ^2 statistic, it appears that the Western Cape's value is approximately 130 on the drug-related crime axis.

Table 6-11 provides the two-dimensional approximations for the contributions to the χ^2 statistic. From the inspection of Table 6-11, the two-dimensional approximated value for the Western Cape's drug-related crime contribution is in fact 139.11. Similarly, all other two-dimensional predictions for the Western Cape are provided for in Figure 6-14, and by the last row Table 6-11. Additionally, Table 6-11 provides all the two-dimensional approximations for each of the provinces. In Table 6-11, just as was the case in Table 6-3, Gauteng's association to robbery with aggravating circumstances and the Western Cape's association to drug-related crime are the largest contributors to the χ^2 statistic and thus the total inertia.

The ability to freely reflect and rotate the graph holds true for the CA biplot. Figure 6-15 is produced by reflecting Figure 6-14 about the Y-axis, and subsequently rotating the figure by -60°. Similarly to Figure 6-9, the benefit of such manipulation is to render Figure 6-14 more comparable to the CA biplots of later years.

Table 6-11: *$\sqrt{n}$ times the predictions for the two-dimensional biplot for the 2004 South African crime data, in weighted deviation form $\boldsymbol{R}^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-\frac{1}{2}}$.*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	-2.23	-1.78	-3.73	4.37	-3.96	-4.06	-1.52	-2.02	10.00
Tso	7.27	5.60	-10.21	-7.21	8.11	4.69	4.66	7.46	-3.25
Atm	-2.61	-1.95	10.63	0.38	-1.41	0.98	-1.57	-2.95	-7.99
Agb	41.12	31.76	-47.76	-43.89	48.01	30.35	26.47	41.80	-31.46
Ast	6.01	4.59	-12.54	-4.66	5.82	2.31	3.79	6.33	2.71
Rac	-48.52	-36.86	125.97	29.84	-41.68	-9.21	-30.29	-52.05	-54.34
Ars	5.14	4.02	-0.09	-7.35	7.27	6.05	3.39	5.00	-11.67
Bnr	10.56	8.11	-17.21	-9.71	11.27	5.90	6.73	10.93	-1.58
Brp	-3.85	-2.92	10.90	2.08	-3.11	-0.39	-2.39	-4.17	-5.50
Ilf	-11.63	-9.00	11.84	12.94	-13.94	-9.22	-7.51	-11.76	11.09
Drg	-14.64	-12.15	-80.35	46.32	-38.03	-48.01	-10.73	-11.06	139.11
Crj	-25.28	-19.21	64.46	15.91	-21.96	-5.24	-15.80	-27.07	-26.77

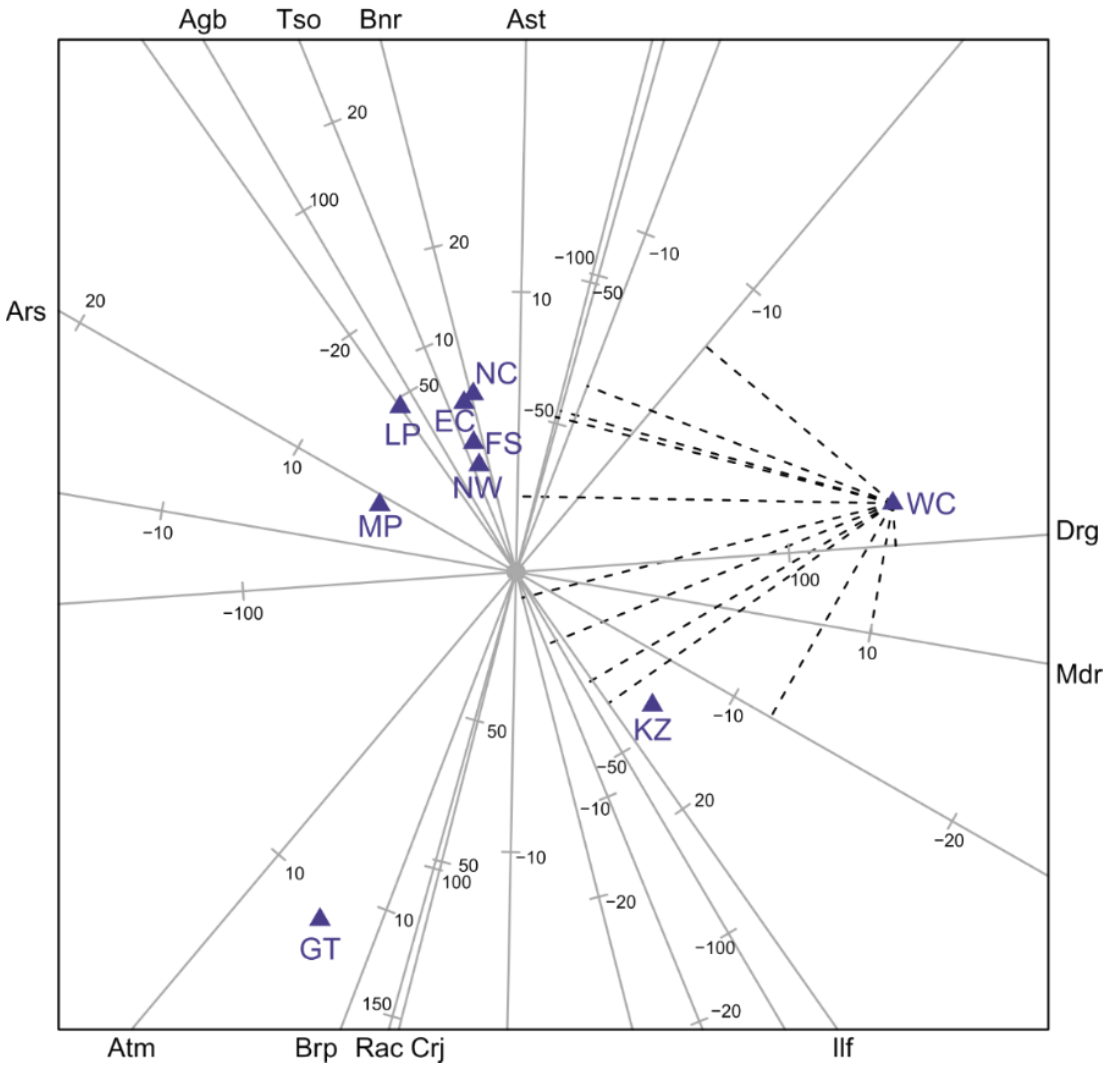

Figure 6-15: A reproduction of Figure 6-12 with `reflect="y"` *and* `rotate.degrees=-60`*.*

***Table 6-12**: Cumulative quality expressed as percentages for the 2004 South African crime data.*

Dimension	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
Quality	49.79	83.20	93.83	97.06	98.57	99.65	99.92	100.00

Table 6-12 provides the cumulative measure of quality by the number of dimensions composing the display for the South African 2004 crime data. The quality of Figure 6-14 and Figure 6-15 attributable to the first two dimensions is 83.20%, exactly that of the CA display of the South African crime data 2004. This property holds for all CAs and CA biplots developed from the same data. Thus, all corresponding CAs and CA biplots (produced from the same data) presented in this chapter are of equal quality. This property is due to the equality of the eigen-values from both the CA and CA biplot algorithms.

An additional measure of performance, known as *predictivity,* is now introduced. Predictivities are a measure of how well the approximations of the CA biplot agree with the corresponding true elements in $\boldsymbol{X}$ (Gower *et al.*, 2011:91). With reference to the CA biplot model from Subsection 5.4.1.1, the respective row and column predictivities of data matrix $\boldsymbol{X}$ are:

$$row\ predictivities = diag\left(\boldsymbol{U}\boldsymbol{\Sigma}^{2}\boldsymbol{J}\boldsymbol{U}'\right)[diag(\boldsymbol{U}\boldsymbol{\Sigma}^{2}\boldsymbol{U}')]^{-1}, \quad (6\text{-}6)$$

$$column\ predictivities = diag(\boldsymbol{V}\boldsymbol{\Sigma}^{2}\boldsymbol{J}\boldsymbol{V}')[diag(\boldsymbol{V}\boldsymbol{\Sigma}^{2}\boldsymbol{V}')]^{-1}, \quad (6\text{-}7)$$

where $\boldsymbol{U\Sigma V'}$ is the SVD of the biplot model (Gower *et al.*, 2011:299). Table 6-13 and Table 6-16 provide the predictivities for both the axes and the sample points of Figure 6-15 respectively. Although Table 6-12 provided an overall quality of the display, the predictivities provides a specific measure of quality for each component of the graph. As expected, by the eighth dimension all components in Table 6-13 and Table 6-16 have a predictivity of 1 (100%). This result is due to the fact that the dimensionality of the 12 × 9 contingency table is 8, and therefore all components would be perfectly represented in said space. Similarly to Section 6.1, conditional highlighting has been utilized in the tables to highlight any predictivity with a value of 0.5 or less. What dictates a poor predictivity is relative, however this thesis assumes that a predictivity less than 0.5 is poor.

As it appears in Table 6-13, a CA biplot constructed from the first two dimensions performs poorly at representing the axes for the 2004 crime data (Figure 6-15). Robbery with aggravating circumstances and carjacking are both represented well by the first dimension. Whereas, arson and murder cannot be accurately represented at all by the first dimension. There is an improvement by the inclusion of a second dimension, as assault with the intent to cause grievous bodily harm and drug-related crimes are also well represented in the two-dimensional approximation. However, several axes are still poorly represented. In particular, arson improves greatly with the addition of the second dimension. There does, however, appear to be a considerable improvement in the axis predictivities with the inclusion of the third dimension. In order to try and incorporate this improvement in the axis predictivities between the second and third dimension in the CA biplot, the following two techniques are proposed.

***Table 6-13**: Cumulative axis predictivities for the 2004 South African crime data in weighted deviation form $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Dim_1	0.01	0.19	0.09	0.25	0.04	0.94	0.00	0.25	0.12	0.08	0.35	0.92
Dim_2	0.07	0.45	0.09	0.90	0.06	0.99	0.48	0.45	0.13	0.36	0.99	0.98
Dim_3	0.83	0.51	0.57	0.97	0.92	1.00	0.73	0.47	0.13	0.82	1.00	0.98
Dim_4	0.87	0.76	0.73	0.98	0.97	1.00	0.82	0.78	0.79	0.93	1.00	0.98
Dim_5	0.96	0.85	0.77	1.00	0.98	1.00	0.88	0.82	0.98	1.00	1.00	0.99
Dim_6	0.97	0.97	0.97	1.00	1.00	1.00	0.94	0.99	0.98	1.00	1.00	0.99
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The first solution stems from the fact that any combination of dimensions can be used to construct the CA biplot. For the likes of Figure 6-14 and Figure 6-15, the first two dimensions are used to construct the two-dimensional figure. As there was a large improvement in the axis predictivities in the three-dimensional approximation, it is of interest to construct a two-dimensional CA biplot using the first and third dimensions. The method for constructing such a figure is much the same as before, however the coordinates for the construction of the figure are obtained by the first and third columns of $\boldsymbol{U\Sigma}^{½}$ and $\boldsymbol{V\Sigma}^{½}$ respectively, rather than the first two. Figure 6-16 below is a CA biplot constructed from the first and third dimensions of the 2004 South African crime data. It is immediately apparent that the positions of the provinces and axes are considerably different to that of Figure 6-14 and Figure 6-15. Notably, Gauteng is in the peripheral of the figure, KwaZulu-Natal has migrated towards the top of the figure, and the remaining provinces now form a much larger cloud of points.

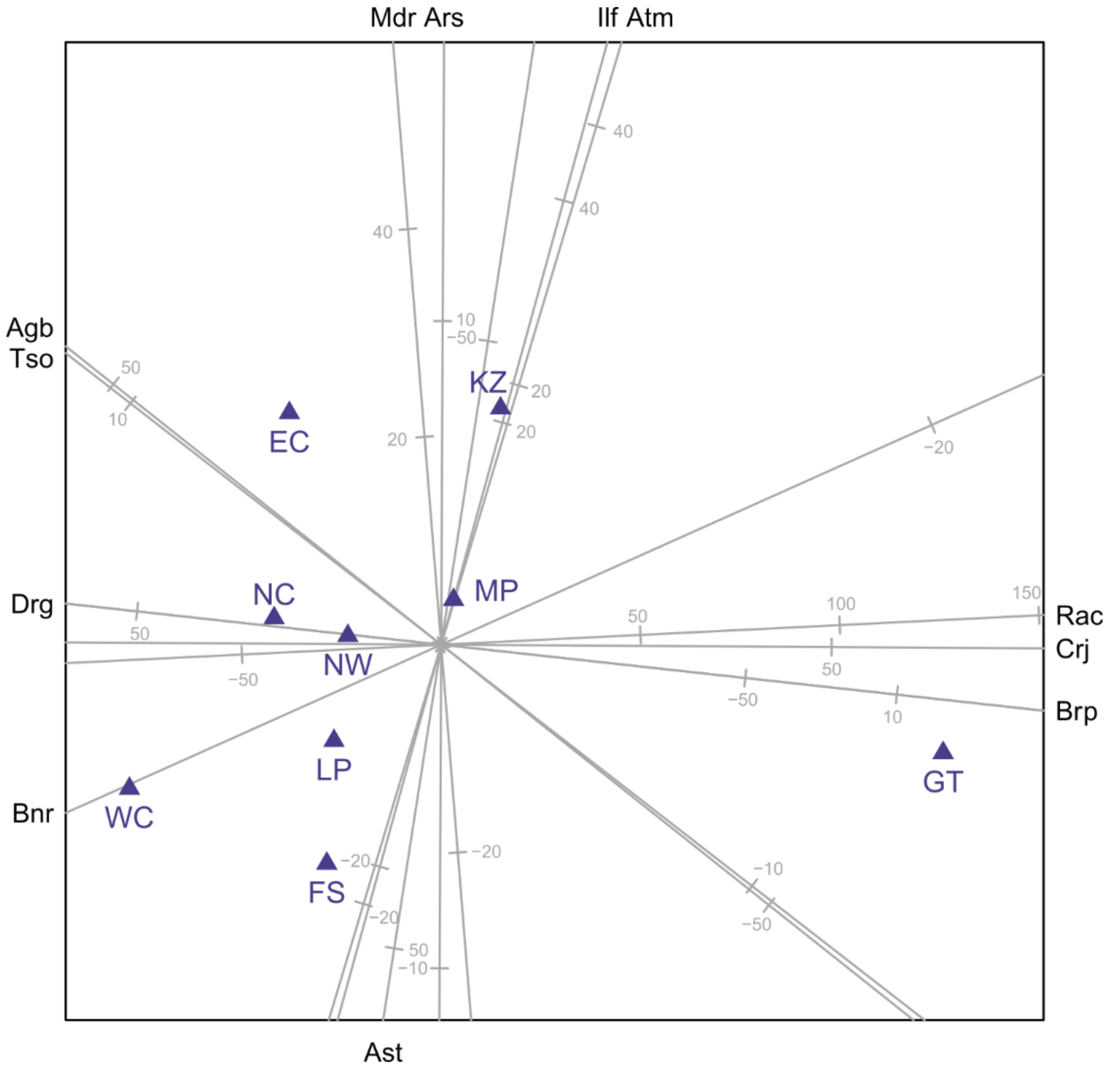

***Figure 6-16**: CA biplot of the 2004 South African crime data, constructed using the 1st and 3rd dimensions.*

The cumulative quality values for Figure 6-16 are provided for in Table 6-14. Similarly, the cumulative axis predictivities for Figure 6-16 are provided for in Table 6-15. Note how the positions of Dim 2 and Dim 3 have changed between the aforementioned tables and Table 6-12 and Table 6-13. Table 6-14 looks almost identical to Table 6-12, and only the second value has changed from 83.20 to 60.41. That is to say that the combination of the first and third dimensions account for approximately 23% less variation in the data than did the first and second. This is an expected result, as the first two dimensions will always account for the greatest variability in the data. However, the intention was not to increase the quality of the graph, but to rather improve the axis predictivities in the two-dimensional approximation.

Table 6-14*: Cumulative quality expressed as percentages for Figure 6-16.*

Dimension	Dim 1	Dim 3	Dim 2	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
Quality	49.79	60.41	93.83	97.06	98.57	99.65	99.92	100.00

Similarly to the above table, Table 6-15 is identical to Table 6-13, baring the second row. Just as was the case before, by the third row, both Table 6-15 and Table 6-13 have identical values. The most apparent difference between Table 6-15 and Table 6-13 is that there are now fewer highlighted cells in the second row. Hence, there are more axes with predictivities greater than 0.5, and as such, there is an improvement as to how well the axes are represented. However, the drug-related crime axis did experience a considerable predictivity decrease from 0.99 to 0.36. This overall improvement confirms that there is merit to utilising different combinations of dimensions other than the first two, however there was a substantial loss in terms of the quality of the figure, and is thus not a definite solution.

Table 6-15*: Cumulative axis predictivities for Figure 6-16 in weighted deviation form* $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$.

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Dim_1	0.01	0.19	0.09	0.25	0.04	0.94	0.00	0.25	0.12	0.08	0.35	0.92
Dim_3	0.77	0.25	0.57	0.33	0.91	0.94	0.26	0.27	0.12	0.53	0.36	0.92
Dim_2	0.83	0.51	0.57	0.97	0.92	1.00	0.73	0.47	0.13	0.82	1.00	0.98
Dim_4	0.87	0.76	0.73	0.98	0.97	1.00	0.82	0.78	0.79	0.93	1.00	0.98
Dim_5	0.96	0.85	0.77	1.00	0.98	1.00	0.88	0.82	0.98	1.00	1.00	0.99
Dim_6	0.97	0.97	0.97	1.00	1.00	1.00	0.94	0.99	0.98	1.00	1.00	0.99
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The second possible solution is to produce a three-dimensional CA biplot. There is no need to recalculate the quality and axis predictivities for the three-dimensional biplot, as these values have been provided for in the above tables. The three-dimensional CA biplot will provide the best quality biplot with the highest predictivities, however, there are several draw backs to using the three-dimensional CA biplot. An example of the three-dimensional CA biplot is presented below. Due to its multidimensional nature, it is difficult to represent in a two-dimensional setting. Figure 6-17 is a three-dimensional CA biplot of the 2004 South African crime data. The three-dimensional CA biplot allows the user to freely rotate and zoom in and out of the figure when viewed in the R graphical user interface, which is necessary for proper interpretation. In order to incorporate this function here, multiple snapshots of the three-dimensional biplot are given in Figure 6-18, Figure 6-19, and Figure 6-20. The three-dimensional representation of the 2004 crime data has a quality of 93.8%, an approximate 10% improvement upon the two-dimensional biplot. By inspection of the Figure 6-18, it is evident that there is a substantial amount of variation between the provinces with respect to the third dimension. The increase in the quality between the two- and three-dimensional biplot is due to the aforementioned variation.

As previously stated, the drug-related crime axis was represented well in the two-dimensional CA biplot of the 2004 South African crime data. Conversely, the murder axis was represented poorly in the two-dimensional biplot. However, the axis predictivity of murder increase dramatically in the three-dimensional biplot. By comparing the position of both axes between the two- and three-dimensional biplots, it is possible to determine why the murder axis was represented poorly in the two-dimensional biplot.

In Figure 6-15 the Western Cape is positioned in the positive direction of both drug-related crime and murder. By the inspection of the three-dimensional biplot, it is apparent that the Western Cape is still positioned in the direction of drug-related crime. However, its position relative to the murder axis has changed greatly. This is due to the fact that the murder axis deviates away from the drug-related crime axis with respect to the third dimension. Conversely, the drug-related crime axis runs almost parallel to the two-dimensional plane, thus its position remains fairly constant when orthogonally projected upon this plane. As such, it has a high predictivity in the two-dimensional biplot. When the murder axis is projected onto the two-dimensional plane (the grey 2D plane represented in Figure 6-17), it is positioned close to the drug-related crime axis. Once projected onto the two-dimensional plane, a considerable amount of information is lost with regards to the murder axis as it is not an adequate representation of its true position.

Figure 6-21 is a three-dimensional CA biplot of the 2004 crime data, which has been positioned to view the figure with respect to the two-dimensional plane. The two-dimensional approximation can be obtained by orthogonally projecting all the points in the three-dimensional biplot onto the two-dimensional plane. Although not exact, it is possible to visualize where the two-dimensional approximation of Figure 6-15 originates from, as the positions of the axes and profile points are similar to that of Figure 6-21. Although the three-dimensional CA biplot can provide an improved representation of the relative positions of the axes and profile points, is can be cumbersome to interpret and accurately represent on paper. Thus, the three-dimensional CA biplot will not be utilised further in this study.

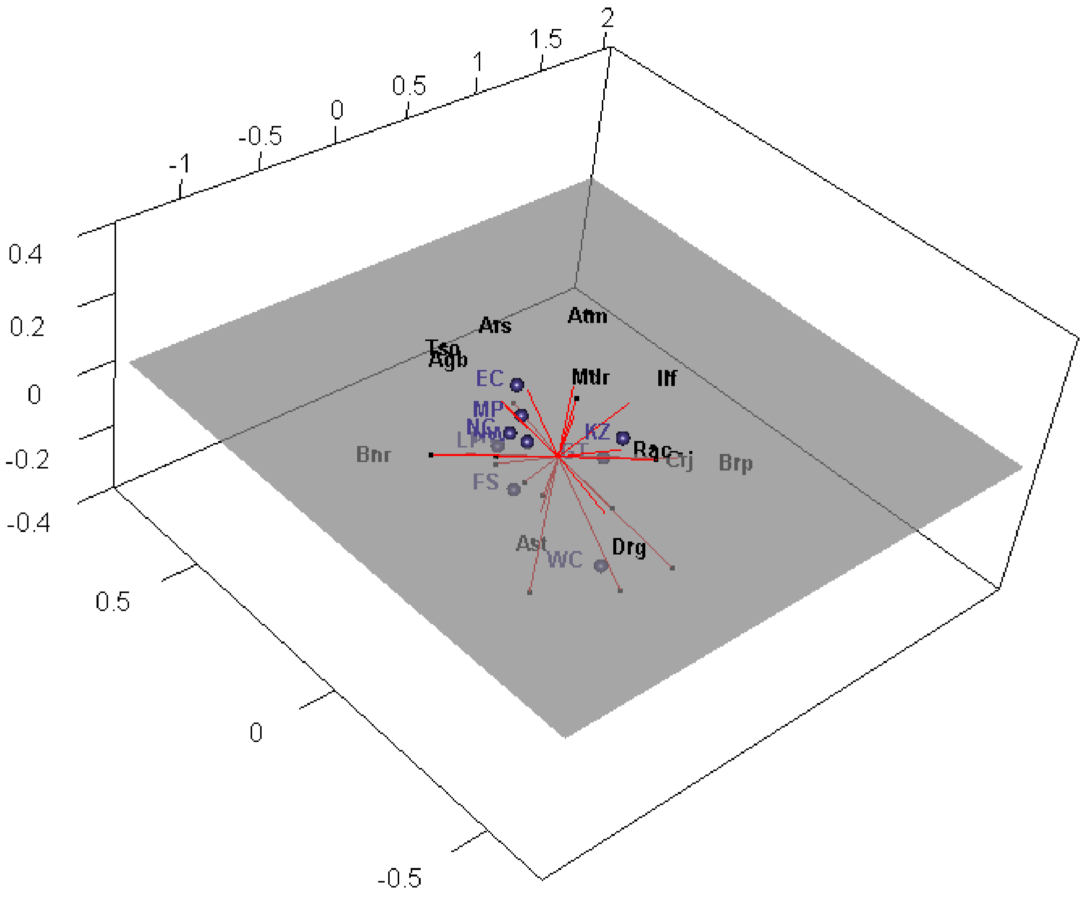

__Figure 6-17__: 3-dimensional CA biplot of the 2004 South African crime data, view A.

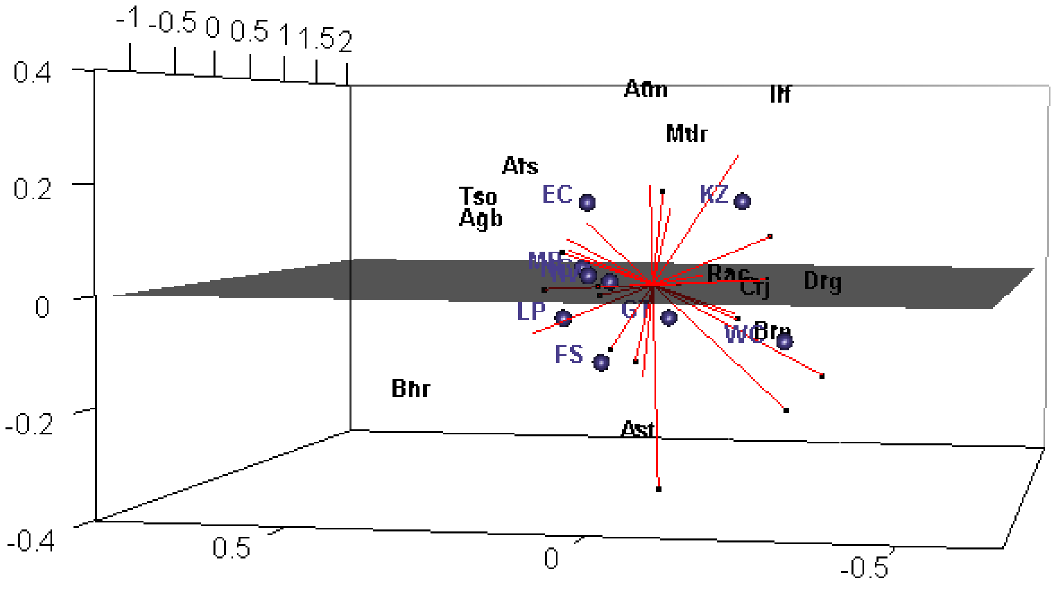

__Figure 6-18__: 3-dimensional CA biplot of the 2004 South African crime data, view B.

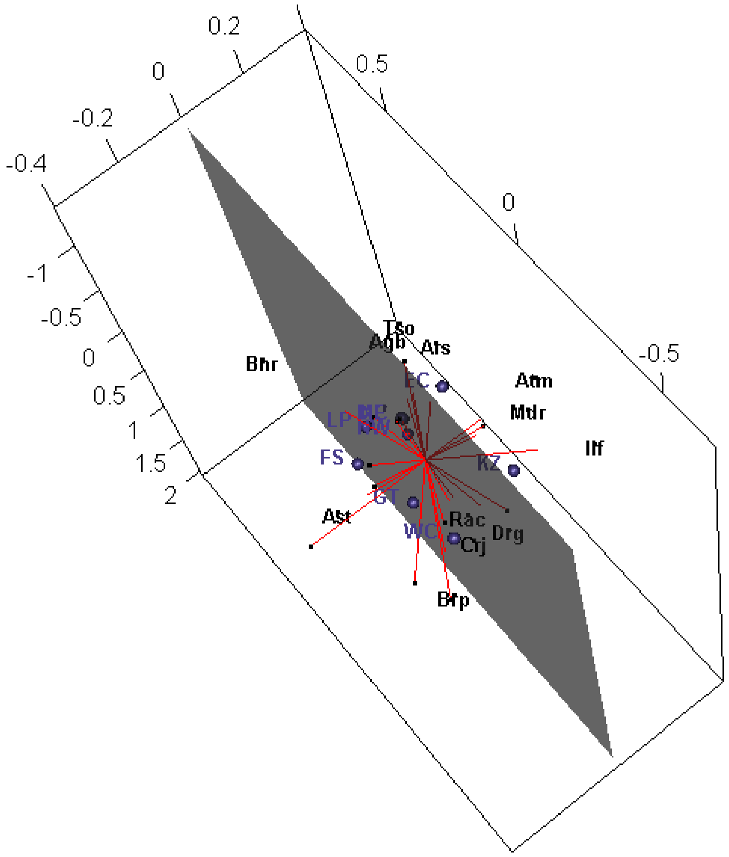

***Figure 6-19**: 3-dimensional CA biplot of the 2004 South African crime data, view C.*

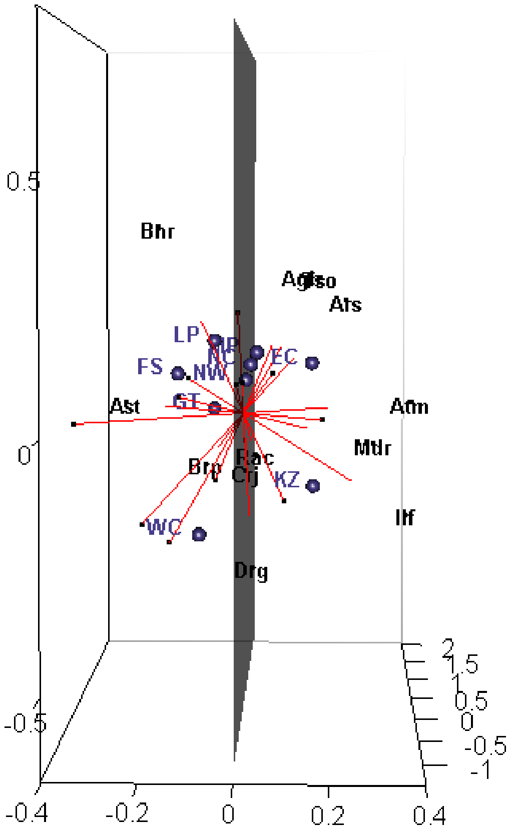

***Figure 6-20**: 3-dimensional CA biplot of the 2004 South African crime data, view D.*

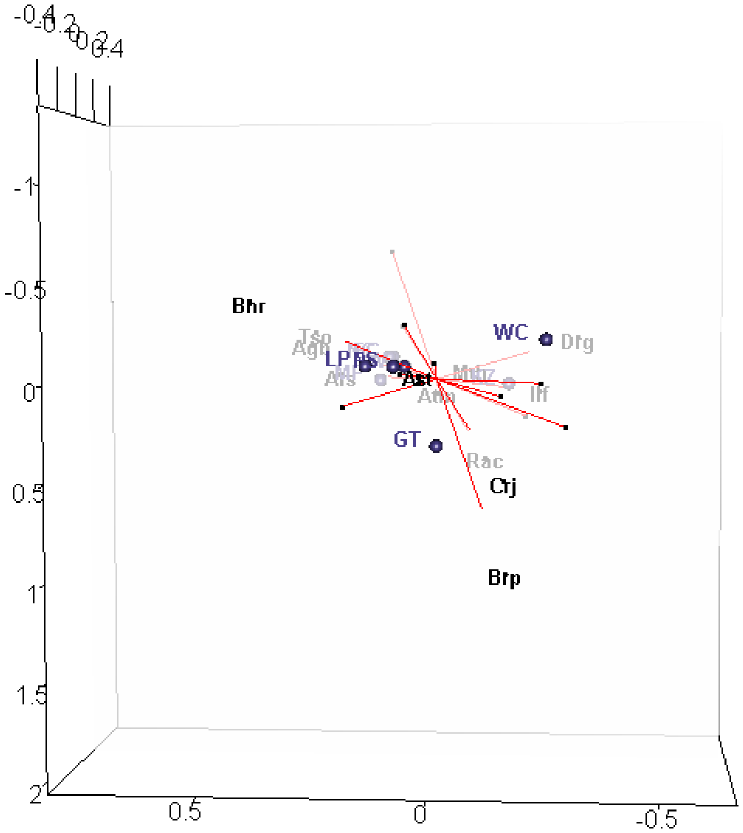

***Figure 6-21**: 3-dimensional CA biplot of the 2004 South African crime data, positioned to mimic a two-dimensional map*

Table 6-16 provides the predictivities for the sample points of the 2004 South African crime data, and has similarly been subject to the same conditional highlighting as in Table 6-13. Likewise, by the eighth dimension all samples are perfectly represented, as was the case in Table 6-13. In general, the first dimension performs far better for the sample predictivities. However, only Gauteng is well represented in the first dimension. With the addition of the second dimension, all provinces except the Free State have a predictivity greater than 0.5. It then follows that the conclusion for Table 6-16 is similar to that of Table 6-13. A three-dimensional approximation is ideal as it provides improvement in the quality of the display, however the analysis is limited to two-dimensional approximations.

***Table 6-16**: Cumulative sample predictivities for the 2004 South African crime data in weighted deviation form $\mathbf{R}^{-½}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-½}$.*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Dim_1	0.32	0.24	0.97	0.04	0.17	0.00	0.39	0.48	0.44
Dim_2	0.61	0.49	0.97	0.58	0.80	0.74	0.81	0.83	0.95
Dim_3	0.95	0.90	1.00	0.89	0.86	0.77	0.81	0.84	0.99
Dim_4	0.99	0.93	1.00	0.99	0.96	0.91	0.81	0.84	1.00
Dim_5	0.99	0.96	1.00	1.00	0.97	0.97	0.85	0.96	1.00
Dim_6	0.99	1.00	1.00	1.00	1.00	0.98	0.98	1.00	1.00
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The CA biplot and its associated tables provide a much more detailed method of analysis than that of the CA. The graph itself is far easier to interpret, and the quality of the representation can be scrutinized in much greater detail. When comparing Figure 6-14 to Figure 6-1, the improvements in interpretability are evident. However, much of the information displayed in Figure 6-14 can be seen in Figure 6-1. The relative positioning of the sample points is similar in both displays, such that the direction of deviation by each province is accounted for in both figures. Due to the similarity of relative positioning, the relationships between sample points and between vertex points is also preserved in both displays. However, the degree to which a province deviates away from the mean is not well represented in Figure 6-1, but due to Lambda-scaling the magnitude of the deviation can always be well represented in a CA biplot.

With the introduction of calibrated axes, the association between provinces and crimes can easily be approximated. All axes which appear in a CA biplot pass through the centroid and have both negative and positive ends. The end at which the axis name appears indicates the positive direction for all CA biplots axes. Following from the goals of this study, the positioning of a province in the CA biplot is of great importance. Appearing on the negative end of an axis indicates a lower than expected value for the individual province with respect to said crime. This is a desirable result, as it represents disassociation between a particular province and a crime type. When a province is positioned at the centre of any axis, this indicates a value for the province equal to that of the average profile, with respect to the individual axis. Of particular interest to this study is to view specific associations between provinces and crimes. Thus, the focus is then on provinces that are located high on the positive end of an axis. Such positioning would indicate a higher than expected value for the province with respect to the crime type represented by the axis.

Applying the above interpretation to the 2004 South African Crime data in Figure 6-15, it is evident that the Western Cape, Kwa-Zulu Natal, and Gauteng each appear to have strong positive associations to specific crimes. Further, the remaining provinces appear to have common associations to the twelve crimes of interest as they are clustered together. With the assistance of the Western Cape’s approximation projections, it is determined that the Western Cape has a strong association to drug-related crime, followed by weaker associations to murder, assault, and illegal firearms. Gauteng is positioned high on the positive end of five axes, indicating higher than average values for carjacking, attempted murder, burglary at residential premises, illegal firearms, and a notably high value for robbery with aggravating circumstances. Kwa-Zulu Natal, which is positioned closer to the centroid than aforementioned provinces, does not appear to be positioned high on any particular axis. However, when compared to other province’s approximations, Kwa-Zulu Natal has notably high values for illegal firearms, drug-related crime, murder and carjacking, followed by weaker associations to burglary at residential premises and robbery with aggravating circumstances.

The remaining six provinces are grouped close together in Figure 6-15, and all are positioned in the same direction on the respective axes. Notably, Mpumalanga is singular among these six provinces for its positive position on the attempted murder axis. These remaining six provinces all have positive deviations in the direction of arson, assault with the intent to cause grievous bodily harm, total sexual offenses, burglary at non-residential premises, and assault. The remaining provinces can subsequently be grouped into larger and smaller deviations. Limpopo, the Eastern Cape, and the Northern Cape having average deviations more than 50% greater than that of Mpumalanga, Free State, and the North-West Province.

In Subsection 6.1.3.1, the display of the 2013 data (Figure 6-12) appeared considerably different to the displays of earlier years. The interpretation was restricted greatly by the fact that all the province points were grouped tightly around the centroid. Figure 6-22 is provided to demonstrate the improvement by the CA biplot on the 2013 data. Figure 6-22 is constructed by following a similar procedure to that of Figure 6-14. By making use of Lambda-scaling, the province points are now well distributed and the interpretability has greatly improved. Just as in Figure 6-12, there are large shifts in the profile points and crime axes in Figure 6-22. Gauteng, the Western Cape, and Kwa-Zulu Natal are still well separated from the remaining provinces. The Eastern Cape, however, has migrated away from the cluster of remaining provinces. The quality of the display is acceptable at 80.41%, the same as that of Figure 6-12. By inspection of Table 6-18, it is evident that the murder, attempted murder, assault, residential burglary and illegal firearm axes are not well represented and must be interpreted with caution. It is of interest to note that all the above mentioned axes were also poorly displayed in Figure 6-15, suggesting that these variables are inherently multidimensional. Similarly, by inspection of Table 6-19, it appears that the majority of provinces are well represented by the two-dimensional approximation. However, Kwa-Zulu Natal remains poorly represented.

The CA biplot has provided much improvement upon the initial correspondence analysis. Although a large amount of information was portrayed in Section 6.1, the CA biplot provided a much more detailed representation of the South African crime data. The development, as well as the interpretation, of the graphs was also demonstrated in this section. The `cabipl()` function proved highly effective in the development and manipulation of CA biplots. The analysis in Section 6.2 was limited to only the 2004 and 2013 South African crime data in order to demonstrate the improvement of the CA biplot over the regular CA. The results of Section 6.2 concurred with those of Section 6.1. However, the interpretability of the results for Section 6.2 are greatly improved.

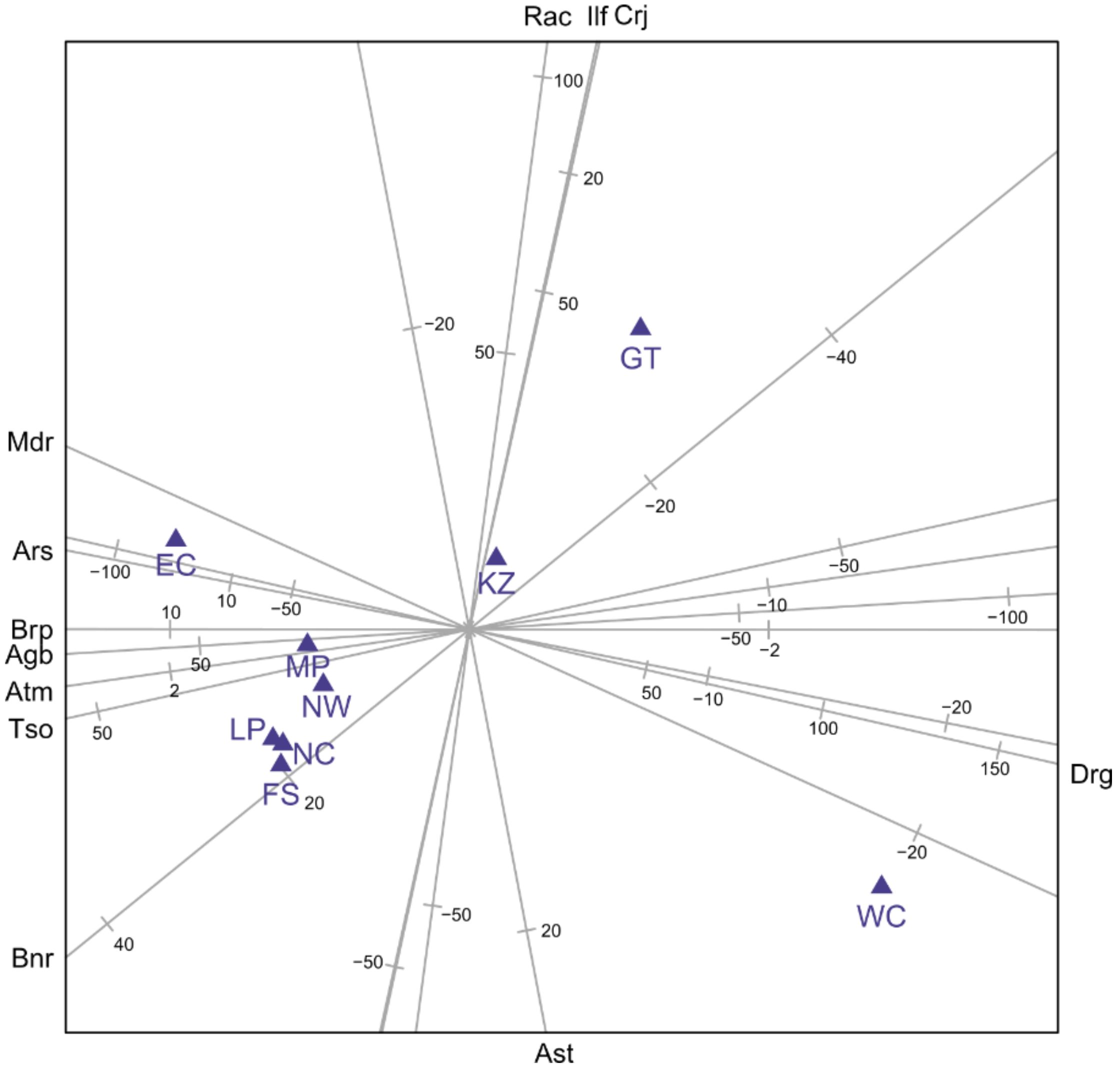

***Figure 6-22**: A two-dimensional CA biplot of the 2013 South African crime data set in terms of the contributions to the Pearson residuals.*

***Table 6-17**: Cumulative quality expressed as percentages for the 2013 South African crime data.*

Dimension	**Dim 1**	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**	**Dim 6**	**Dim 7**	**Dim 8**
Quality	64.85	80.41	89.35	94.54	97.54	99.18	99.86	100.00

***Table 6-18**: Cumulative axis predictivities for the 2013 South African crime data in weighted deviation form* $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$.

	Mdr	**Tso**	**Atm**	**Agb**	**Ast**	**Rac**	**Ars**	**Bnr**	**Brp**	**Ilf**	**Drg**	**Crj**
Dim_1	0.25	0.80	0.03	0.90	0.01	0.03	0.73	0.54	0.19	0.02	0.95	0.08
Dim_2	0.27	0.82	0.03	0.90	0.15	0.97	0.74	0.71	0.19	0.18	0.97	0.91
Dim_3	0.59	0.91	0.31	0.90	0.87	0.98	0.85	0.71	0.20	0.47	0.99	0.97
Dim_4	0.85	0.92	0.48	0.96	0.93	0.98	0.86	0.94	0.86	0.47	1.00	0.97
Dim_5	0.92	0.94	0.56	0.99	0.99	0.98	0.86	0.96	0.87	0.93	1.00	0.98
Dim_6	0.92	0.99	0.76	1.00	1.00	1.00	1.00	0.98	0.97	0.98	1.00	0.98
Dim_7	0.99	1.00	0.99	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

***Table 6-19**: Cumulative sample predictivities for the 2013 South African crime data in weighted deviation form* $\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}$.

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Dim_1	0.79	0.45	0.35	0.04	0.67	0.59	0.70	0.65	0.83
Dim_2	0.83	0.56	0.87	0.16	0.78	0.59	0.75	0.77	0.99
Dim_3	0.92	0.94	0.98	0.50	0.78	0.59	0.84	0.80	1.00
Dim_4	0.98	0.96	0.98	0.51	0.91	0.90	0.87	0.87	1.00
Dim_5	0.98	1.00	1.00	0.92	0.92	0.91	0.97	0.94	1.00
Dim_6	0.99	1.00	1.00	0.94	0.99	0.98	0.98	0.99	1.00
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

6.3 Yearly Progressions of Provinces

Section 6.2 presented additional evidence that four of the nine provinces had profiles notably different to the remaining provinces. Gauteng, the Western Cape, and Kwa-Zulu Natal have consistently been positioned away from all other provinces. By 2013, the Eastern Cape had also migrated away from the cluster of provinces in the direction of the arson and murder axes. Following from this, it is then of interest to investigate the progressions of Gauteng, the Western Cape, Kwa-Zulu Natal, and the Eastern Cape over the ten year period (2004-2013).

In order to visualize the progressions of the profile points over the ten year period, it is necessary to interpolate the profile points of previous years onto an existing CA biplot. The interpolated points play no role in determining the positions of the axes in the CA biplot, and are an addition post development of the figure. For the purpose of clarity, the biplot for which the supplementary points are to be displayed upon, is referred to as the '*base*' biplot. The base CA biplot for the following section will categorically be that of the 2013 crime data (Figure 6-14). Details on the interpolation of supplementary points can be found in Chapter 7 of *Understanding Biplots* (Gower *et al.*, 2011:302–303). A summary of the derivations from this text is provided below.

The general model of a CA biplot is given by:

$$\boldsymbol{Y} = \boldsymbol{W}_1^{-1}\boldsymbol{R}^{-½}(\boldsymbol{X}-\boldsymbol{E})\boldsymbol{C}^{-½}\boldsymbol{W}_2^{-1}. \qquad (6\text{-}8)$$

Following from equation 6-8, the approximation of $\boldsymbol{Y}$ is given by:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{W}_1^{-1}\boldsymbol{U\Sigma V}'\boldsymbol{W}_2^{-1} = \boldsymbol{AB}', \qquad (6\text{-}9)$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ provide the approximated plotting coordinates in r dimensions. In order to display the profile points over a period of ten years, the interpolated row coordinates are required for this period. Due to this, the derivations are restricted to the interpolation of supplementary row points (provinces) only. Following from equation 6-9, it is then possible to denote the row coordinates in a general form as:

$$\boldsymbol{A} = \hat{\boldsymbol{Y}}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}. \tag{6-10}$$

In order to supplement a new row $\boldsymbol{x}': 1 \times q$ to $\boldsymbol{X}$, a new row $\hat{\boldsymbol{y}}': 1 \times q$ is also required. The new values $\hat{\boldsymbol{y}}'$ are thus given by:

$$\hat{\boldsymbol{y}}' = w_1^{-1} r^{-½} \left(\boldsymbol{x}' - \frac{r\mathbf{1}'\boldsymbol{C}}{n}\right) \boldsymbol{C}^{-½} \boldsymbol{W}_2^{-1}, \tag{6-11}$$

where $r = \mathbf{1}'\boldsymbol{x}$ is the total of the new row and w_1 is its associated weight. The new row coordinates $\boldsymbol{a}'$ are then given as:

$$\boldsymbol{a}': 1 \times q = \hat{\boldsymbol{y}}'\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}. \tag{6-12}$$

From equation 6-9 it holds that $\boldsymbol{A} = \boldsymbol{W}_1^{-1}\boldsymbol{U}\boldsymbol{\Sigma}^{\alpha}\boldsymbol{J}$. Recalling that the $\boldsymbol{J}$ matrix is a convenient way of denoting an r-dimensional approximation of a matrix, and having the restriction that $\alpha + \beta = 1$, then $\boldsymbol{A}$ can be expanded to:

$$\boldsymbol{A} = \boldsymbol{W}_1^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}')\boldsymbol{V}\boldsymbol{\Sigma}^{-\beta}\boldsymbol{J} = \boldsymbol{W}_1^{-1}\boldsymbol{R}^{-½}(\boldsymbol{X} - \boldsymbol{E})\boldsymbol{C}^{-½}\boldsymbol{V}\boldsymbol{\Sigma}^{-\beta}\boldsymbol{J}. \tag{6-13}$$

In order to remove the generality from equation 6-13, the specific weights for the Pearson residual can be applied. Given that the weighting matrices are identity matrices for the Pearson residual model, the above formula simplifies to:

$$\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}')\boldsymbol{V}\boldsymbol{\Sigma}^{-\beta}\boldsymbol{J} = \boldsymbol{R}^{-½}(\boldsymbol{X} - \boldsymbol{E})\boldsymbol{C}^{-½}\boldsymbol{V}\boldsymbol{\Sigma}^{-\beta}\boldsymbol{J}. \tag{6-14}$$

The interpolated points for the new row $\boldsymbol{x}'$ are thus given by:

$$\boldsymbol{a}' = r^{-½}\boldsymbol{x}'\boldsymbol{C}^{-½}\boldsymbol{V}\boldsymbol{\Sigma}^{-\beta}\boldsymbol{J}, \tag{6-15}$$

where β is the scaling factor for the row points. Similarly, α is the scaling factor of the column coordinates (see Gower et al., 2011), and their combined value must sum to unity. The formulas for interpolating supplementary points can be implemented by the `cabipl()` function. By inputting the ten year values for a province of interest into the `X.new` argument, the series of profile points for the selected province will be interpolated into the figure.

Figure 6-23 displays the interpolated points of the Gauteng province for the 2004-2013 period, supplemented onto the CA biplot of the 2013 crime data. By inputting the yearly crime frequencies of Gauteng into the `X.new` argument, the respective interpolated points for the period 2004-2013 are displayed on the graph. The axes are calibrated directly in terms of the Pearson residuals by incorporating the factor $\sqrt{n}$ in the calibrations, as implemented by the argument `PearsonRes.scaled.markers=TRUE`.

Accompanying this figure is Table 6-20, which provides the measure of predictivity of the interpolated points. The predictivity for all of the years is greater than 0.5, an acceptable quality of display. However, the points are not perfectly represented and are merely approximations of their true positions. The 2013 supplementary point has a predictivity equal to that of Figure 6-22, 0.87. As Figure 6-22 is developed only from the 2013 data, it is expected that it would perform poorly at representing profiles from previous years. The predictivity for Gauteng in Table 6-16 was at a high of 0.9744. However, in Figure 6-23 the 2004 Gauteng point has a predictivity of only 0.53. Due to this, the interpolated points cannot be followed blindly and should be scrutinized against the yearly CA biplots of Section 6.2 before drawing any conclusions.

The weighted deviation matrix, and the two-dimensional predictions of the deviations for Gauteng are provided for in Table 6-21 and Table 6-22 respectively. By the comparison of these two tables, insight into the poor representation can be gained. The absolute differences between Table 6-21 and Table 6-22 are provided for in Table 6-23. For the case of Figure 6-23, it appears only Gauteng's positions relative to the drug-related crime, carjacking, and robbery with aggravating circumstances are well represented. The movement of the profile relative to all other axes should be scrutinized. Following from this, the movements of Gauteng's profile point can only be an accurate representation of the movement along the well represented axes. Hence, the vertical drop of Gauteng's profile point appears to be a result of its movement along the robbery with aggravating circumstances, and carjacking axes. Where after, Gauteng's profile moves horizontally in the direction of drug-related crime.

The horizontal movement is confined to the 2011, 2012, and 2013 profile points, a result of the large growth experienced in drug-related crimes for this period. This growth in Gauteng's profile was evident in Figure 4-10 where there was a large increase in the drug-related crime portion of the stacked bar plot from 2011-2013. However, the increase in drug-related crime is concealed when analysing the contributions to the total inertia by Gauteng over the reported period (Figure 6-23). Unfortunately, the proportional contributions to the total inertia are not always a good indicator of increases or decreases in the number of reported cases for crimes in a specific province. However, as both the CA biplots of Section 6.2 and the analysis in Section 4.2 draw similar conclusions, it follows that there is a considerable amount of evidence to suggest that Gauteng's association to drug-related crime has increased considerably over the reported period. Additionally, there is a large amount of evidence to support Gauteng's decrease in carjacking and robbery with aggravating circumstances.

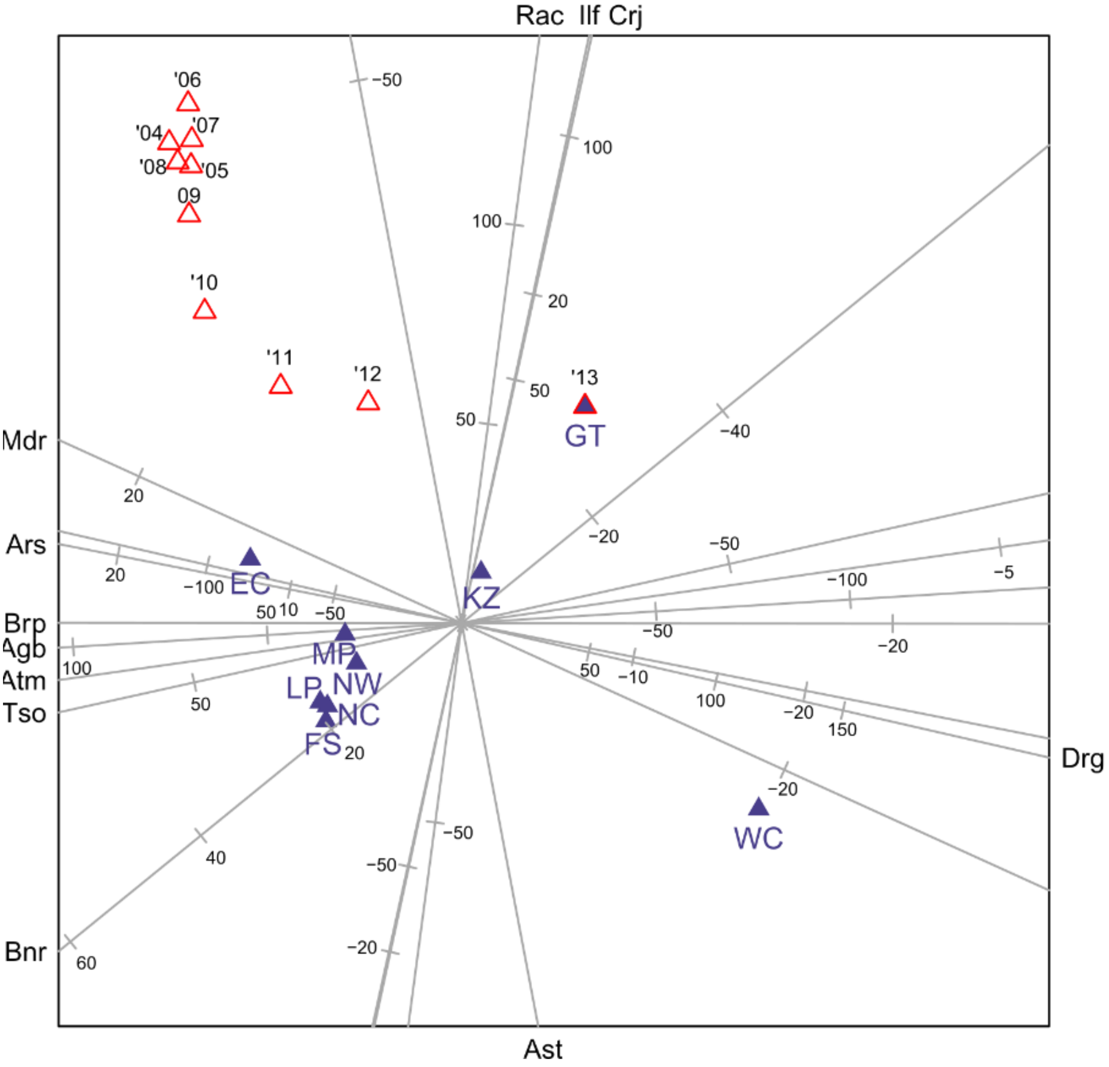

Figure 6-23: A two-dimensional CA biplot of the 2013 South African crime data with Gauteng's yearly progression.

Table 6-20: Predictivities calculated for the interpolated Gauteng points in Figure 6-23.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Predictivity	0.53	0.59	0.66	0.64	0.66	0.61	0.58	0.54	0.53	0.87

***Table 6-21**: The Weighted Deviation Matrix $\boldsymbol{R}_{new}^{-\frac{1}{2}}(\boldsymbol{X}_{new}-\boldsymbol{E}_{new})\boldsymbol{C}^{-\frac{1}{2}}$ of the Gauteng crime data for the period 2004-2013.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
2004	-19.21	-8.91	45.98	29.66	198.26	218.71	15.85	-78.48	37.74	-5.54	-355.70	118.55
2005	-18.30	-5.55	32.14	30.60	151.46	194.75	20.62	-64.72	49.82	-14.49	-320.93	127.56
2006	-8.95	-6.23	35.49	32.47	132.71	232.97	26.85	-42.09	18.19	1.91	-323.12	134.87
2007	-8.12	4.62	29.58	30.80	140.26	210.14	19.74	-27.73	8.05	-5.56	-313.09	145.22
2008	-7.60	30.95	22.00	17.09	127.85	193.74	11.94	-10.30	24.57	4.39	-318.37	143.65
2009	-20.25	-1.18	12.02	22.44	130.07	157.05	5.64	-6.78	58.16	6.04	-311.76	136.89
2010	-17.80	-7.65	3.08	30.36	124.23	115.88	11.63	-4.38	65.13	2.04	-282.02	102.29
2011	-19.49	-19.14	-7.90	22.29	100.26	86.41	11.14	-3.14	46.64	13.82	-207.49	79.47
2012	-22.77	-26.71	-7.60	-8.32	55.47	81.66	-1.79	-15.07	57.96	4.97	-138.05	74.05
2013	-28.16	-68.72	-15.16	-50.89	2.08	96.08	-10.91	-34.82	-6.88	-10.38	31.92	88.56

***Table 6-22**: Two-dimensional predicted values of $\boldsymbol{R}_{new}^{-\frac{1}{2}}(\boldsymbol{X}_{new}-\boldsymbol{E}_{new})\boldsymbol{C}^{-\frac{1}{2}}$ for Gauteng, for the period 2004-2013.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
2004	41.25	52.15	3.22	106.10	-74.90	171.20	34.05	-14.27	21.36	38.07	-236.40	128.01
2005	38.64	47.46	2.95	97.85	-71.10	163.38	31.73	-14.72	19.76	36.40	-220.45	122.40
2006	41.15	44.45	2.87	97.52	-79.82	187.22	33.06	-21.59	19.98	41.99	-230.50	141.26
2007	39.53	45.61	2.88	96.87	-74.71	173.51	32.11	-17.91	19.70	38.79	-223.45	130.47
2008	39.86	50.96	3.13	103.14	-71.97	164.14	32.97	-13.22	20.74	36.48	-228.79	122.63
2009	36.99	51.06	3.08	99.86	-64.27	144.21	31.05	-8.63	19.90	31.87	-215.00	107.11
2010	32.19	52.66	3.05	95.93	-50.37	107.59	28.01	0.49	18.76	23.37	-192.85	78.45
2011	23.30	36.02	2.11	67.13	-37.87	82.43	20.02	-1.68	13.21	18.02	-138.11	60.54
2012	15.63	12.53	0.89	32.27	-33.26	80.57	12.04	-12.43	6.83	18.26	-84.51	61.45
2013	-2.00	-47.97	-2.20	-55.04	-27.01	90.36	-7.11	-43.44	-8.98	22.23	43.33	75.17

Figure 6-24 displays the progression of the Western Cape, and has accompanying predictivities for the interpolated points in Table 6-24. The predictivities of the profile points for the years 2004, 2005, and 2006 are low, resulting in poor representations of the Western Cape's profile points for this period. Despite the initial low values, all other years have high predictivities, particularly from 2009 onwards. Similarly to Table 6-23, Table 6-25 provides the absolute percentage differences between the actual and predicted values for the Western Cape's interpolated points. By analysis of the average difference, it appears that the Western Cape's movement relative to the assault with the intent to cause grievous bodily harm, robbery with aggravating circumstances, and carjacking are well represented. When the 2004-2006 results are disregarded, the remaining years are well represented with respect to the drug-related crime, total sexual offenses, and arson axes. The movement of the Western Cape's profiles point is then a result of its progression along the well represented axes.

In Figure 6-24 the Western Cape's progression forms an '*S*'-like shape in the direction of drug-related crime. Noted both in Chapter 4 and Subsection 6.1.1, the Western Cape has a very strong connection to drug-related crime. The progression in Figure 6-24 shows the magnitude, as well as the increase of the Western Cape's association to drug-related crimes over the reported period. Additionally, the 'S'-like shape is a result of the profile point's movement down the assault with the intent to cause grievous bodily harm, total sexual offenses, and arson axes. Although the start and end positions of the profile point with respect to the robbery with aggravating circumstances axis are similar, the oscillation of the point along this axis provides an accurate representation of its movement over the reported period. With the exception of arson, the perceived changes in the Western Cape's profile point relative to the aforementioned axes was evident in the univariate analysis. This provides strong evidence of the Western Cape's association to these crimes. The remaining axes perform poorly at representing the Western Cape's progression and should be interpreted with caution.

Table 6-23*: Absolute percentage difference between the actual and two-dimensional approximations of* $\boldsymbol{R}_{new}^{-\frac{1}{2}}(\boldsymbol{X}_{new}-\boldsymbol{E}_{new})\boldsymbol{C}^{-\frac{1}{2}}$ *for Gauteng, for the period 2004-2013.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
2004	315	685	93	258	138	22	115	82	43	787	34	8
2005	311	956	91	220	147	16	54	77	60	351	31	4
2006	560	813	92	200	160	20	23	49	10	2094	29	5
2007	587	887	90	215	153	17	63	35	145	798	29	10
2008	624	65	86	504	156	15	176	28	16	730	28	15
2009	283	4425	74	345	149	8	451	27	66	428	31	22
2010	281	789	1	216	141	7	141	111	71	1045	32	23
2011	220	288	127	201	138	5	80	46	72	30	33	24
2012	169	147	112	488	160	1	773	18	88	267	39	17
2013	93	30	85	8	1401	6	35	25	31	314	36	15
Average	344	909	85	265	274	12	191	50	60	685	32	14

Table 6-24*: Predictivities calculated for the interpolated Western Cape points in Figure 6-24.*

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Predictivity	0.25	0.22	0.32	0.66	0.88	0.93	0.96	0.97	0.99	0.99

Table 6-25: Absolute percentage difference between the actual and two-dimensional approximations of $\boldsymbol{R}_{new}^{-\frac{1}{2}}(\boldsymbol{X}_{new} - \boldsymbol{E}_{new})\boldsymbol{C}^{-\frac{1}{2}}$ for the Western Cape, for the period 2004-2013.

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
2004	137	3075	115	71	90	12	147	169	67	53	63	58
2005	145	840	103	159	87	16	103	140	80	311	97	47
2006	206	97	99	73	81	3	73	141	112	418	3726	3
2007	329	46	97	36	69	8	47	136	149	264	48	7
2008	47	25	95	19	43	13	33	116	181	206	11	14
2009	14	99	95	9	36	8	9	127	46	28	3	22
2010	27	61	90	17	34	12	19	90	22	198	7	15
2011	20	51	87	19	46	8	39	71	38	21	8	10
2012	31	19	23	5	130	8	35	30	1	146	5	4
2013	76	3	33	6	77	4	18	13	18	4	4	18
Average	103	432	84	41	69	9	52	103	71	165	397	20

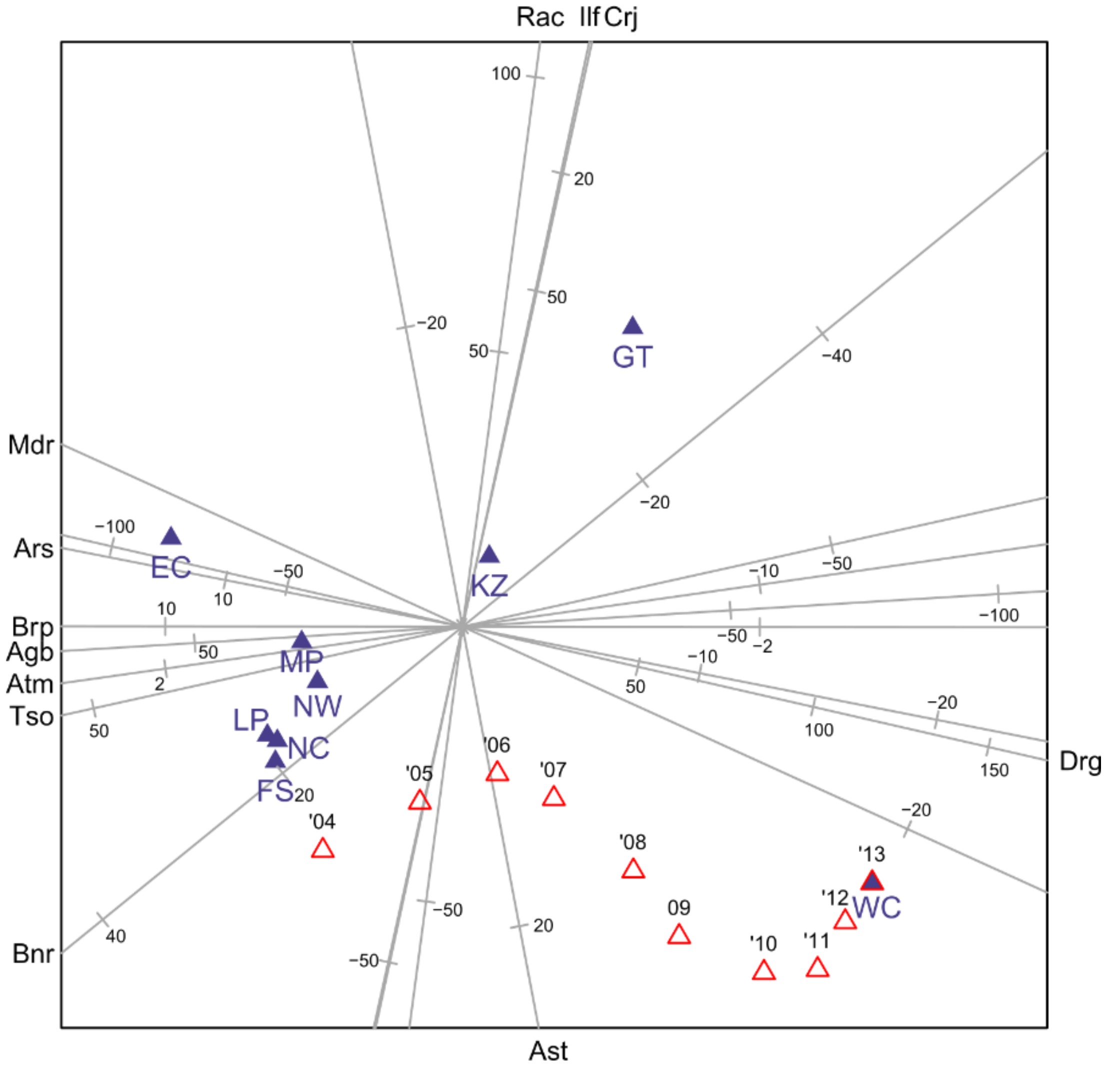

Figure 6-24: A two-dimensional CA biplot of the 2013 South African crime data with the Western Cape's yearly progression.

Kwa-Zulu Natal and its progression points appear in Figure 6-25. As opposed to Figure 6-23 and Figure 6-24, the representation of Kwa-Zulu Natal in Figure 6-25 is poor. By the inspection of Table 6-26, only the years 2005-2008 are adequately displayed in Figure 6-25. This results in numerous unreliable profile points representing Kwa-Zulu Natal in Figure 6-25. Inference is then difficult, as the positioning of the Kwa-Zulu Natal profile points are not representative of their true location. The predictivities of Kwa-Zulu Natal for the CA biplots for the years 2004-2013 are presented in Table 6-27. For the two-dimensional approximation, the predictivities remain low with only two values greater than 0.6, and many of the predictivities are below 0.5. This may suggest that Kwa-Zulu Natal is inherently multidimensional and that it cannot be well represented in a two-dimensional space. The generally poor representation of Kwa-Zulu Natal in the CA biplots may then account for the inadequate representation in Figure 6-25. In order to circumvent this problem, a CA biplot has been constructed based solely on the yearly reported crime frequencies for Kwa-Zulu Natal over the reported period (Figure 6-26).

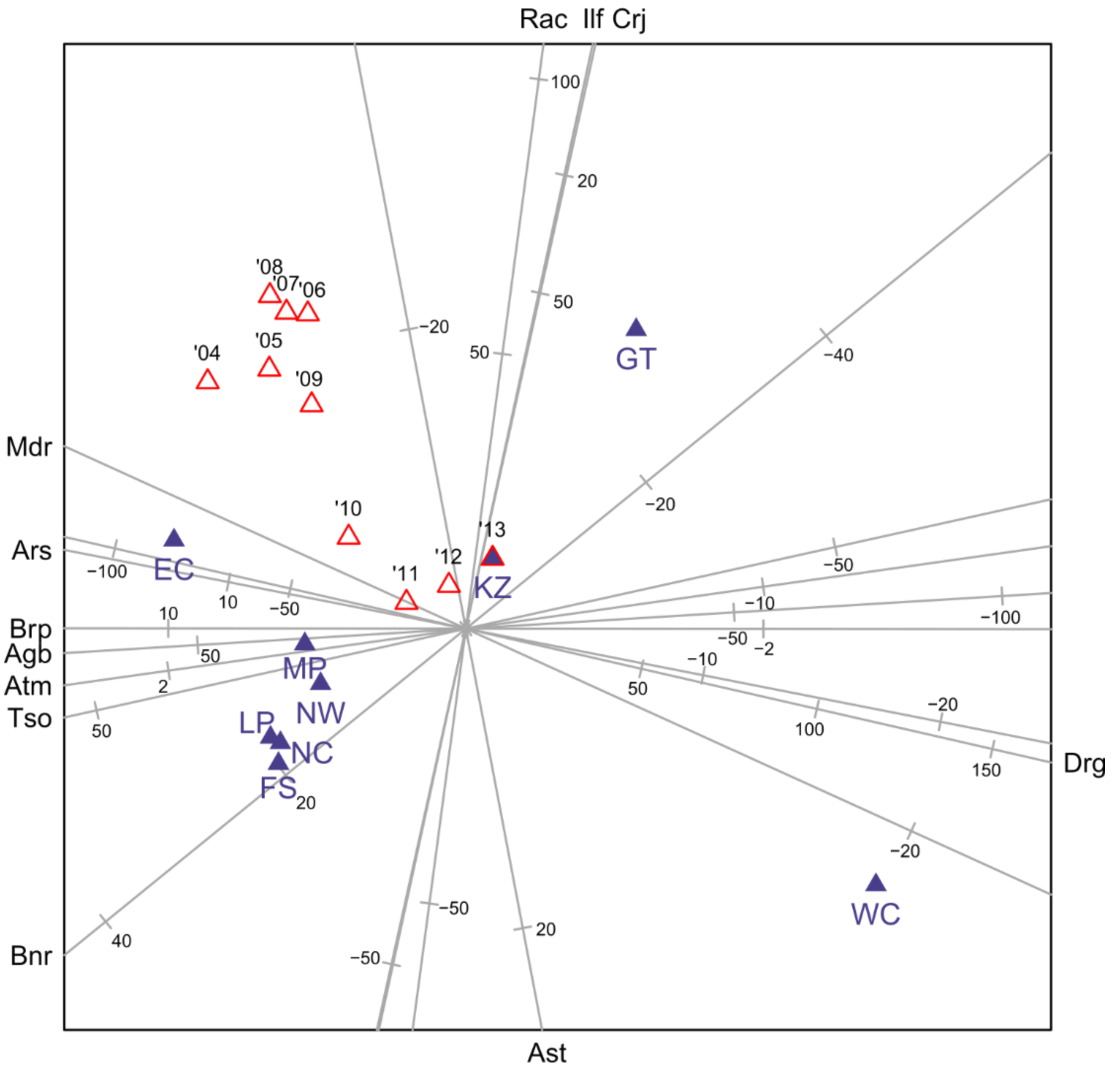

***Figure 6-25**: A two-dimensional CA biplot of the 2013 South African crime data set with KwaZulu-Natal's yearly progression.*

***Table 6-26**: Predictivities calculated for the interpolated Kwa-Zulu Natal points in Figure 6-25.*

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Predictivity	0.47	0.50	0.56	0.62	0.65	0.49	0.30	0.12	0.06	0.16

Accompanying Figure 6-26 is Table 6-28, which depicts the quality of display by number of dimensions. From inspection of Table 6-28, it appears that Figure 6-26 has a quality measure of 94.6%, a large improvement on the inadequacies of Figure 6-25. The associated sample predictivities for Figure 6-26 appear in Table 6-29. It then follows that all yearly profile points are well represented in the two-dimensional biplot, with exception to 2009, and to a lesser extent 2010. The associated axis predictivities, which appear in Table 6-30, suggest that the total sexual offenses, burglary at residential premises, and the illegal firearm axes do not perform well. Following from the above evidence, with the exception to the aforementioned years and axes, the progression of the yearly Kwa-Zulu Natal profile points can be interpreted with confidence.

***Table 6-27**: The predictivities of the Kwa-Zulu Natal profile points for the CA biplots developed from the yearly South African crime data, for the period 2004-2013.*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Dim_1	0.04	0.32	0.37	0.16	0.04	0.10	0.23	0.34	0.34	0.04
Dim_2	0.58	0.66	0.69	0.56	0.51	0.44	0.42	0.48	0.44	0.16
Dim_3	0.89	0.91	0.93	0.92	0.86	0.87	0.79	0.61	0.66	0.50
Dim_4	0.99	0.99	0.98	0.98	0.96	0.92	0.91	0.75	0.70	0.51
Dim_5	1.00	1.00	0.99	0.98	0.99	1.00	1.00	1.00	0.96	0.92
Dim_6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.94
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Kwa-Zulu Natal's progression can be separated into three parts, its starting point in 2004, the movement from 2004 to 2008, and the subsequent progression from 2008 to 2013. The profile point initially drops vertically until 2008, where after its movement is reversed upward and horizontally along the drug-related crime axis. In 2004, Kwa-Zulu Natal had a strong association to assault, assault with the intent to cause grievous bodily harm, and arson. At this point in time, Kwa-Zulu Natal has a strong disassociation to drug-related crime. Over the 2004-2008 period, the profile point makes a quick progression towards the robbery with aggravating circumstances, and carjacking. Simultaneously, there is smaller growth in its association to total sexual offenses and drug-related crime. Accompanying this growth is a large decrease in the association to assault, and to a lesser extent assault with the intent to cause grievous bodily harm, and arson. The slight decrease in murder is well represented, however, the position relative to attempted murder is poor.

The most noticeable aspect of the 2008-2013 progression is the substantial increase in the association to drug-related crime. Additionally, there is a large decrease in carjacking, assault with the intent to cause grievous bodily harm, assault, robbery with aggravating circumstances, arson, attempted murder, murder, and total sexual offenses. When the entire 2004-2013 progression is viewed as a single movement, the most apparent change is that the profile point has progressed from the negative extreme of the drug-related crime axis to the positive end. The disadvantage of Figure 6-26 is that there is no longer a complete analysis being conducted. As the figure is only developed from the Kwa-Zulu Natal yearly data, there is no accounting for the behaviour of other provinces. Unlike the biplots appearing in Figure 6-23 and Figure 6-24, Figure 6-26 does not display the deviation from the average South African province profile. Rather, it displays the deviation from the average Kwa-Zulu Natal profile itself. In this sense, the analysis of Figure 6-26 is more of an internal, rather than a comparative analysis.

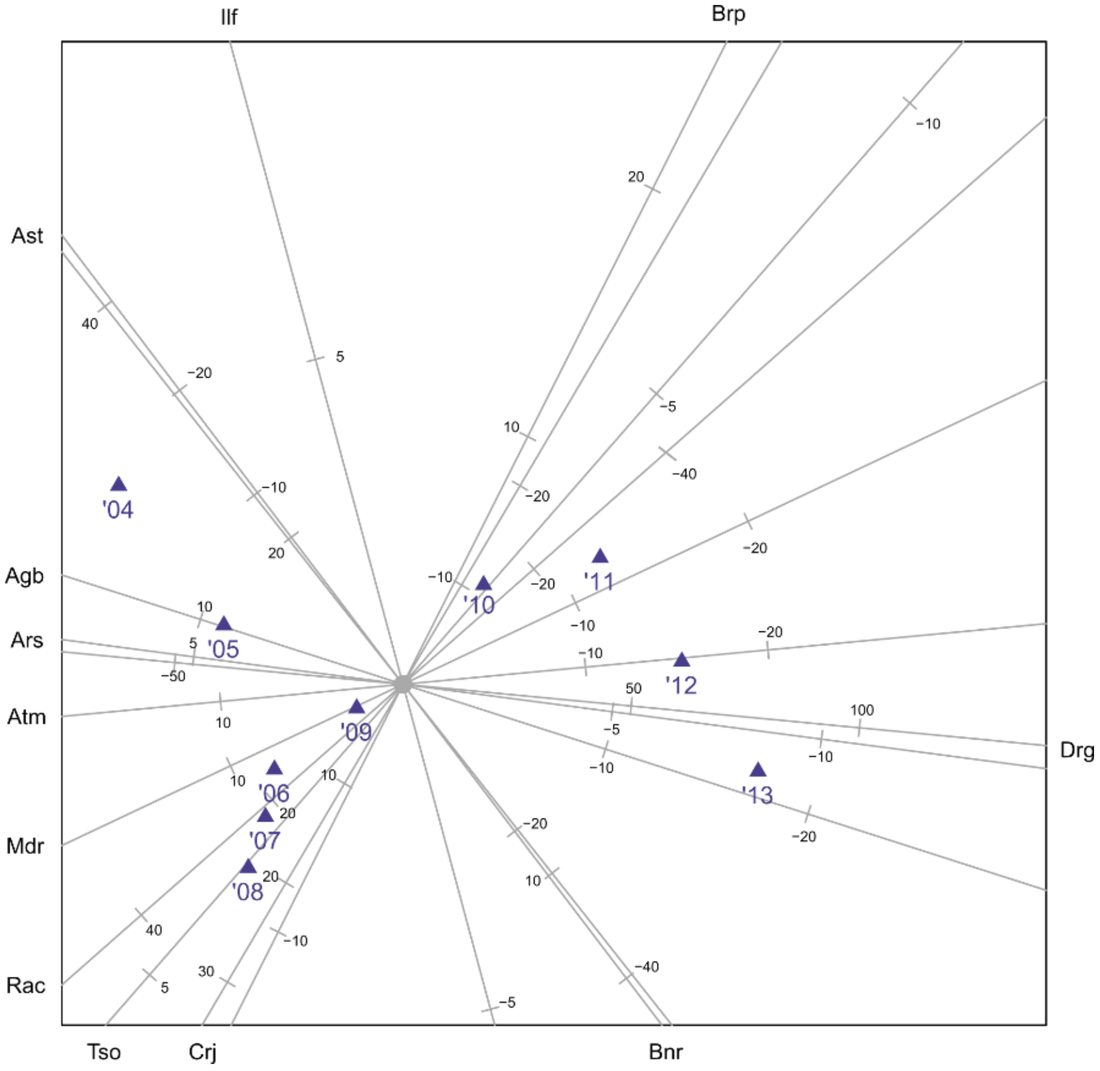

***Figure 6-26**: A CA biplot developed from the Kwa-Zulu Natal yearly crime data.*

***Table 6-28**: Cumulative quality of display for the Kwa-Zulu Natal yearly crime data, by number of compositional dimensions.*

Dimension	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9
Quality	86	95	99	99	100	100	100	100	100

Table 6-29*: Cumulative sample predictivities for the CA biplot developed from the yearly Kwa-Zulu Natal crime data.*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Dim_1	0.85	0.91	0.75	0.73	0.64	0.35	0.46	0.86	0.98	0.95
Dim_2	0.98	0.95	0.86	0.95	0.93	0.38	0.68	0.97	0.98	0.97
Dim_3	1.00	0.99	0.96	0.95	0.97	0.93	0.96	0.99	0.98	1.00
Dim_4	1.00	0.99	0.98	0.97	0.99	0.94	0.99	1.00	1.00	1.00
Dim_5	1.00	1.00	1.00	0.99	0.99	0.95	0.99	1.00	1.00	1.00
Dim_6	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Dim_9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6-30*: Cumulative axis predictivities for the CA biplot developed from the yearly Kwa-Zulu Natal crime data.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Dim_1	0.88	0.08	0.92	0.87	0.62	0.76	0.92	0.34	0.22	0.04	1.00	0.47
Dim_2	0.94	0.12	0.92	0.90	0.97	0.96	0.93	0.52	0.50	0.23	1.00	0.90
Dim_3	0.99	0.84	1.00	0.90	0.99	1.00	0.93	0.94	0.83	0.50	1.00	0.98
Dim_4	0.99	0.89	1.00	0.97	0.99	1.00	0.97	0.97	0.97	0.83	1.00	0.98
Dim_5	0.99	0.92	1.00	0.99	1.00	1.00	0.99	1.00	0.99	0.83	1.00	0.98
Dim_6	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.84	1.00	1.00
Dim_7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
Dim_8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
Dim_9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The final progression is that of the Eastern Cape (Figure 6-27). Similarly to before, the accompanying Table 6-31 provides the predictivities for the Eastern Cape profile points for the ten year period. Additionally, Table 6-32 provides the absolute percentage difference between the actual and predicted values for the Eastern Cape's interpolated points. From the inspection of Table 6-31, it appears that all the profile points are adequately represented in Figure 6-27. As opposed to the previous displays in this subsection, there is a minimal amount of variation in the Eastern Cape's profile point over the reported period. This is to be expected as the Eastern Cape displayed a relatively constant profile in Figure 4-10. Albeit that the predictivities are high, the predictions on the attempted murder assault, burglary at non-residential premises, burglary at residential premises, and illegal firearm axes are not accurate. Hence, the predictions on these axes should be interpreted with caution.

The Eastern Cape's profile point experiences an upward movement at an approximate 45° angle. As the Eastern Cape migrates horizontally towards the centroid, there is a shift from the negative ends of the robbery with aggravating circumstances and carjacking axes, towards their respective positive ends. From Figure 4-10 it was evident that there was a large growth in the robbery with aggravating circumstances, followed by minimal growth in carjacking portions of the stacked bars.

Additionally, there is a large decrease is assault, followed by a smaller decrease in assault with the intent to cause grievous bodily harm. Thus it appears that the progression of the Eastern Cape's profile point over the reported period is mostly due to the large increase in robbery with aggravating circumstances, and the large decrease assault with the intent to cause grievous bodily harm.

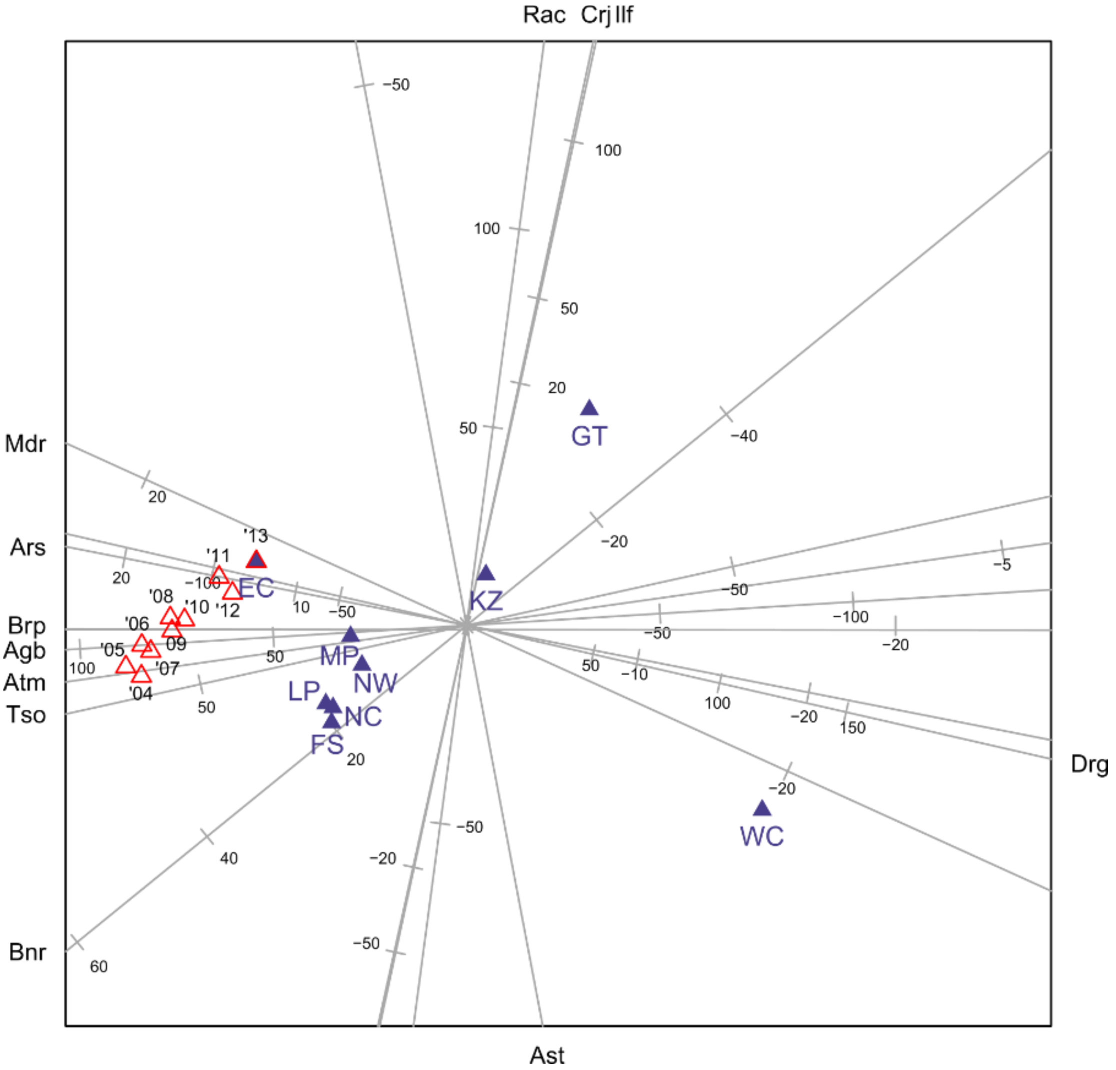

***Figure 6-27**: A two-dimensional CA biplot of the 2013 South African crime data set with the Eastern Cape's yearly progression.*

***Table 6-31**: Predictivities calculated for the interpolated Eastern Cape points in Figure 6-27.*

Year	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Predictivity	0.73	0.82	0.79	0.82	0.89	0.88	0.90	0.89	0.90	0.83

***Table 6-32**: Absolute percentage difference between the actual and two-dimensional approximations of $\boldsymbol{R}_{new}^{-\frac{1}{2}}(\boldsymbol{X}_{new} - \boldsymbol{E}_{new})\boldsymbol{C}^{-\frac{1}{2}}$ for the Eastern Cape, for the period 2004-2013.*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
2004	46	378	86	32	102	42	35	231	31	464	6	8
2005	58	66	76	23	108	43	26	236	22	31	3	9
2006	61	104	58	25	116	51	30	211	31	0	3	5
2007	62	85	67	26	116	47	31	259	5	26	0	13
2008	59	31	61	21	506	59	15	301	12	676	3	8
2009	58	54	48	25	227	48	26	340	5	24	3	11
2010	59	26	468	22	4536	39	29	377	53	91	5	36
2011	63	14	498	23	25	25	21	290	127	85	8	91
2012	68	5	17	17	61	238	26	553	1436	475	13	69
2013	70	21	40	19	66	31	41	491	200	85	21	132
Average	60	79	142	23	586	62	28	329	192	196	7	38

6.4 Summary and conclusions

This chapter aimed to analyse the relationships between the two variables of the South African crime data, namely provinces and crimes. This was achieved by utilizing correspondence analysis (CA), and the correspondence analysis biplot (CA biplot). Both techniques provided methods for visually representing the relationships between the variables. However, the CA biplot proved to have some advantages over the traditional CA displays, as it provided greater interpretability. This improvement was achieved by the use of Lambda-scaling and calibrated axes.

Evidence was found that at the beginning of the reported period (2004), there were only two main contributing provinces to the total inertia. The two largest contributors were Gauteng and the Western Cape, together comprising more than 50% of the total inertia contribution. Similarly, there were three crimes which made up the majority of contribution to total inertia. The aforementioned crimes were: assault with the intent to cause grievous bodily harm, robbery with aggravating circumstances, and drug-related crime. By analysing the individual cells of the 2004 South African crime table, it was revealed that the Western Cape's association to drug-related crime, and Gauteng's association to robbery with aggravating circumstances were the only interactions which contributed largely to the total inertia of the table.

From the results of Section 6.1.2, it is clear that there is an initial increase in the total inertia over the first seven years of the reported period. Where after, the total inertia quickly decreases to a lower value than that of the 2004 South African crime data. This suggests that the inertia of the South African crime data may behave in a cyclical manner. As one or many provinces deviate away from the average profile, a new equilibrium is eventually established. This equilibrium represents a convergence of all the provinces' profile points to a standard profile. This convergence results in a decrease in the total inertia of the South African crime data.

Over the 2004-2013 reported period, assault with the intent to cause grievous bodily harm, and drug-related crimes are consistently large contributors to the total inertia. At the beginning of the reporting period, robbery with aggravating circumstances was also a large contributor. However, after 2009 robbery with aggravating circumstances is no longer regarded as a main contributor. Additionally, total sexual offenses becomes the third largest contributor by 2013. Similarly, Gauteng and the Western Cape remain the two largest province contributors to the total inertia. However, by 2013, the Eastern Cape and the Free State are also classified as large contributors. From the analysis of the 2013 South African crime data, there appeared to be a single cell which made a considerable contribution to the total inertia, that being the interaction of the Western Cape and drug-related crime.

Throughout the entire reported period, the analysis continually highlighted the Western Cape as the largest contributor to the total inertia, specifically its association to drug-related crime. This was reaffirmed by the prominent progression of the Western Cape’s profile point along the drug-related crime axis in Figure 6-24.

Chapter 7 continues the analysis on the South African crime data by investigating the applicability of compositional data and the extension of biplots relating to such methods.

Chapter 7: Log-Ratio Biplots

The techniques utilized in Chapter 6 assumed that the data are on interval scales. That is to say that when two values are compared, the difference is expressed as an interval between the two. Alternatively, the log-ratio biplot is concerned with variables which are measured on ratio scales. By applying logarithmic transformations, these ratios can then be expressed as multiplicative differences (Greenacre, 2010:70).

Correspondence analysis and the CA biplot have already presented possible analytical methods for comparing the various crime frequencies. By converting the table into a set of column profiles, a comparison could be conducted on the proportional crime frequencies between provinces. An alternative measure could be to compare the ratios of crime within a province, and between provinces. However, due to the fact that each province has a different population size, each of the province's frequency measures have different scales. It then follows that a direct comparison of ratios would not suffice either. Rather, a method for transforming the data is needed in order to make the values comparable. Log-ratio biplots provide a method for analysing these ratios in a multiplicative context, as interval differences between the logarithms of the data.

Although log-ratio analysis (LRA) was originally developed in the analysis of compositional data, it is equally applicable to contingency tables, as long as there are no zero values within the table (Greenacre, 2009:70). The CA profiles of a contingency table are in fact compositional data, as a profile's elements sum to unity. LRA is thus applicable to both the profiles of the data, as well as raw counts too. In LRA, similarly to CA, values are compared between each row and between each column of the table. For a matrix of size $n \times p$ there are $½n(n-1)$ unique ratios between rows, and $½p(p-1)$ unique ratios between columns to be considered (Greenacre, 2010:71).

7.1 The Analysis of Compositional Data

Compositional data arise from various fields of study. Such data are often found in scientific studies and samples. However, as mentioned previously, the profiles of a two-way table mimic the form of compositional data. Hence, such analysis can be extended to the study of two-way tables. Aitchison (1986) provided an in depth study as to why regular statistical methods are inappropriate for the study of compositional data. In his book, Aitchison introduced the concept of sub-compositions and sub-compositional coherence, the role of which are crucial in the study of compositional data.

The concept of sub-compositions and sub-compositional coherence are best explained with the use of a short example. Assume there are two analysts, analysts A and B, who are conducting a study. They are both working with the same sample. However, analyst A has access to much more accurate analytical tools than that of analyst B. Due to this, analyst A will be able to conduct a more in depth analysis. Assume analyst A can view a composition in D parts, analyst B will then only be able to a view a composition of D^* parts due to poorer tools, such that $D > D^*$. Hence, analyst B is only working with a sub-composition. It is then required that any inference that either analyst make about the common parts to their compositions must be the same. This agreement of results is what is defined as *sub-compositional coherence*.

Aitchison (1986) proved that regular product moment correlations and principal components based on covariances calculated on raw compositional data do not have sub-compositional coherence. Therefore regular analysis would not perform well in the study of compositional data. Due to the fact that composition $\boldsymbol{x}$ can be determined by D ratios such as $\frac{x_i}{x_k}$ $(i = 1, \dots, D)$, and that ratios are invariant under the formation of sub-compositions, Aitchison (1986) suggested that the study of compositional data should be focused on relative values rather than absolute values. This led to the analysis of log-ratios as it proved easier to manipulate than non-transformed ratios.

Let $\boldsymbol{X}$ be a compositional data matrix of size $n \times p$, whose row elements $(x_1, \dots, x_p)$ sum to unity. A biplot which is to be developed on such data should then be produced in terms of log-ratios of the data. Aitchison (1986) stated that there are three equivalent ways of considering ratios within compositional vectors of length p:

1. The $\frac{1}{2}p(p-1)$ ratios $\frac{x_j}{x_{j'}}$ between pairs of components, where it is assumed that $j < j'$.
2. The $p-1$ ratios $\frac{x_j}{x_p}$ between the first $p-1$ components.
3. The p ratios $\frac{x_j}{g(\boldsymbol{x})}$ between the components and their geometric average $g(\boldsymbol{x}) = (x_1 x_2 \dots x_p)^{\frac{1}{p}}$.

The motivation for utilising the log-ratios is that on the logarithmic scale, the above ratios are respective differences:

1. $\log(x_j) - \log(x_{j'})$.
2. $\log(x_j) - \log(x_p)$.
3. $\log(x_j) - \left(\frac{1}{p}\right)\Sigma_j \log(x_j)$, the deviations from the mean.

Due to the fact that $\log(x_j) - \log(x_p)$ is not symmetric with respect to all components, Aitchison and Greenacre (2002) suggested that only the pairwise log-ratios $\log(x_j) - \log(x_{j'})$ and the centred log-ratios $\log(x_j) - \left(\frac{1}{p}\right)\Sigma_j \log(x_j)$ are of interest. Let $\boldsymbol{L} = \log(\boldsymbol{X})$, with elements $l_{ij} = \log(x_{ij})$. Additionally, let the dot subscripts in $l_{i.}$, $l_{.j}$, and $l_{..}$ denote the averages over the columns and rows of the table, as well as the overall average respectively. It then follows that the pairwise log-ratios and centred log-ratios are denoted as $l_{ij} - l_{ij'}$ and $l_{ij} - l_{..}$ respectively. Let $\boldsymbol{T}$ be an $n \times \frac{1}{2}p(p-1)$ matrix of pairwise log-ratios with general element $t_{i,jj'} = l_{ij} - l_{ij'}$, $j < j'$. Although the focus of the analysis is on matrix $\boldsymbol{T}$, Aitchison and Greenacre (2002) provide a method for obtaining all the results of $\boldsymbol{T}$ with the use of a smaller matrix of centred log-ratios which has only p columns. Let this smaller matrix of log-ratios be denoted as: $\boldsymbol{Z} = [z_{ij}]$, where $z_{ij} = l_{ij} - l_{i.} - l_{.j} + l_{..}$. This result means that only the smaller matrix $\boldsymbol{Z}$ is required in order to produce a biplot which accounts for all the results of the larger matrix $\boldsymbol{T}$. Aitchison and Greenacre (2002) named a biplot produced from data matrix $\boldsymbol{Z}$ a *relative variation biplot*.

7.2 Log-Ratio Biplot Algorithm

The log-ratio analysis relies on a double-centring of the log-transformed data matrix $\boldsymbol{X}$, and further weighting of the rows and columns proportional to the margins of $\boldsymbol{X}$. Fortunately, Greenacre (2010) provides a neat and simple algorithm for producing the log-ratio biplot.

Let the row and column sums of $\boldsymbol{X}$, relative to the grand total of $\boldsymbol{X}$ ($t = \sum_i \sum_j x_{ij}$), be denoted by $\boldsymbol{r}$ and $\boldsymbol{c}$ respectively:

$$\boldsymbol{r} = \left(\frac{1}{t}\right)\boldsymbol{X}\mathbf{1},\ \boldsymbol{c} = \left(\frac{1}{t}\right)\boldsymbol{X}'\mathbf{1}. \tag{7-1}$$

The weighted double-centring of the elements of $\log(\boldsymbol{X})$ is denoted as:

$$\boldsymbol{Y} = \boldsymbol{R}^{½}(\boldsymbol{I} - \mathbf{1}\boldsymbol{r}')\boldsymbol{L}(\boldsymbol{I} - \mathbf{1}\boldsymbol{c}')'\boldsymbol{C}^{½}, \tag{7-2}$$

where $\boldsymbol{L} = \log(\boldsymbol{X})$. Following the methods of previous biplots, the low-dimensional approximation of the data is determined through the singular-value decomposition of the weighted matrix $\boldsymbol{Y}$. The SVD of $\boldsymbol{Y}$ is then denoted as:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'. \tag{7-3}$$

To obtain the unweighted log-ratio biplot, both $\boldsymbol{r}$ and $\boldsymbol{c}$ must be replaced in the above algorithm by $\left(\frac{1}{n}\right)\mathbf{1}$ and $\left(\frac{1}{p}\right)\mathbf{1}$ respectively. The unweighted double-centring of the elements of $\log(\boldsymbol{X})$ is thus denoted as:

$$\boldsymbol{Y} = (\boldsymbol{I} - (1/n)\mathbf{1}\mathbf{1}')\boldsymbol{L}(\boldsymbol{I} - (1/p)\mathbf{1}\mathbf{1}')'. \tag{7-4}$$

Similarly to the CA and CA biplot, there are two possible sets of plotting points for the log-ratio biplot. The log-ratio biplot can have either the rows or the columns of the data matrix $\boldsymbol{X}$ represented in principal coordinates, and the second set in standard coordinates. The calculation of the coordinates are as follows:

$$\textit{Principal coordinates of rows}: \boldsymbol{F} = \boldsymbol{R}^{-½}\boldsymbol{U}\boldsymbol{\Sigma} \tag{7-5}$$

$$\textit{Principal coordinates of columns}: \boldsymbol{G} = \boldsymbol{C}^{-½}\boldsymbol{V}\boldsymbol{\Sigma} \tag{7-6}$$

$$\textit{Standard coordinates of rows}: \boldsymbol{\Phi} = \boldsymbol{R}^{-½}\boldsymbol{U} \tag{7-7}$$

$$\textit{Standard coordinates of columns}: \boldsymbol{\Gamma} = \boldsymbol{C}^{-½}\boldsymbol{V} \tag{7-8}$$

The biplot which preserves the row-metric plots $\boldsymbol{F}$ and $\boldsymbol{\Gamma}$, and depicts the log-ratios of the rows across columns of matrix $\boldsymbol{X}$. Similarly, the biplot which preserves the column-metric plots $\boldsymbol{G}$ and $\boldsymbol{\Phi}$, and depicts the log-ratio of columns across the rows of matrix $\boldsymbol{X}$. The points represented by the standard coordinates represent all the log-ratios between pairs of points. The vector connecting two standard coordinate points is known as the *link* vector.

The positions of the principal coordinates represent the log-ratio distances between them. The distances between these points are in fact a weighted Euclidean distance. Assuming a row-principal analysis, the squared distance between rows i and i' is:

$$d_{ii'}^2 = \sum\sum_{j<j'} c_j c_{j'} \left(\log\left(\frac{x_{ij}}{x_{ij'}}\right) - \log\left(\frac{x_{i'j}}{x_{i'j'}}\right)\right)^2 . \tag{7-9}$$

Equation 7-9 can be written in terms of the logarithms of odds-ratios across all pairs of samples as:

$$d_{ii'}^2 = \sum\sum_{j<j'} c_j c_{j'} \left(\log\left(\frac{x_{ij}}{x_{i'j}}\frac{x_{i'j'}}{x_{ij'}}\right)\right)^2 . \tag{7-10}$$

In equation 7-10 the squared terms have been averaged over the samples, so that the distances do not depend on sample size. This measure of distance is then comparable with the χ^2 distance of CA. The total variance of data matrix $\boldsymbol{X}$ is measured by the sum of squares of $\boldsymbol{Y}$, which can be evaluated by the squared distances in terms of the odds-ratio as:

$$\sum\sum_{i<i'}\sum\sum_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log\left(\frac{x_{ij}}{x_{i'j}}\frac{x_{i'j'}}{x_{ij'}}\right)\right)^2 . \tag{7-11}$$

Just as in CA, it is desirable to find a low-dimensional space which accounts for as much of the total variance as possible. The low-dimensional sub-space is found through the use of the singular value decomposition. The proportion of total variation accounted for in the sub-space is again used as the measure of quality of the display.

7.3 Relation to Correspondence Analysis

An important property of the log-ratio analysis is its connection to correspondence analysis. It was previously assumed that there was no apparent link between these two methods until Greenacre showed that the two do in fact belong to the same family of methods (Greenacre, 2009). As it so happens, correspondence analysis and both the unweighted and weighted log-ratio biplot can be connected via the power transformation of the original data matrix (Greenacre, 2009). This connection is heavily reliant of the Box-Cox transformation which states that $f(x) = \left(\frac{1}{\alpha}\right)(x^{\alpha} - 1)$ converges to $\log(x)$ as $\alpha \rightarrow 0$. This transformation was introduced by Box & Cox in their 1964 paper (Box & Cox, 1964).

In order to make the connection easier to demonstrate, the correspondence analysis algorithm is provided for below. However, unlike in Chapter 5, the algorithm is argued via the *correspondence matrix* $\boldsymbol{P} = \frac{1}{t}\boldsymbol{X}$, where t is the grand total of $\boldsymbol{X}$ (Greenacre, 2009:3108). As previously mentioned, the CA algorithm can be argued via $\boldsymbol{X}$ or $\frac{1}{t}\boldsymbol{X}$. The choice to argue by either option has no material effect on the optimal solution.

Let the correspondence matrix be denoted as:

$$\boldsymbol{P} = \frac{1}{t}\boldsymbol{X}. \tag{7-12}$$

It then follows that the calculated matrix of standardized residuals is given by:

$$\boldsymbol{S} = \boldsymbol{R}^{-½}(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}')\boldsymbol{C}^{-½} \tag{7-13}$$

The calculations of the low-dimensional approximation of the profile and vertex points are then calculated from $\boldsymbol{S}$ just as in Chapter 5. In order to make a direct comparison between CA and LRA, it proves beneficial to write equation 7-13 in terms of the matrix of contingency ratios $\boldsymbol{Q} = \boldsymbol{R}^{-1}\boldsymbol{P}\boldsymbol{C}^{-1}$:

$$\boldsymbol{S}^* = \boldsymbol{R}^{½}(\boldsymbol{I} - \mathbf{1}\boldsymbol{r}')(\boldsymbol{R}^{-1}\boldsymbol{P}\boldsymbol{C}^{-1})(\boldsymbol{I} - \mathbf{1}\boldsymbol{c}')'\boldsymbol{C}^{½}, \tag{7-14}$$

Greenacre (2007). The pre- and post-multiplication of $\boldsymbol{Q}$ by the respective centring matrices $(\boldsymbol{I} - \mathbf{1}\boldsymbol{r}^T)$ and $(\boldsymbol{I} - \mathbf{1}\boldsymbol{c}')$ creates a matrix of weighted double-centered contingency ratios. It should be evident to the reader that both the algorithms of CA and LRA are almost identical. It is only the pre-processing of the data matrix which differs slightly between the two methods. From equation 7-14 it is now possible to draw a direct comparison between CA and LRA, as both equation 7-2 and 7-14 are in terms of contingency ratios. Since the logarithm of the contingency ratios is $\log(x_{ij}) - \log(x) - \log(r_i) - \log(c_j)$, the double-centring removes the constant $\log(x)$ and main effects $\log(r_i)$ and $\log(c_j)$. Therefore the only differences between the initial matrices $\boldsymbol{Y}$ and $\boldsymbol{S}^*$ is that equation 7-14 operates on the contingency ratios, whereas equation 7-2 operates on the log-transformed contingency ratios.

The two forms of the CA model, 7-13 and 7-14, provide two methods for introducing the power transformations. In equation 7-13, a pre-transformation can be applied to matrix $\boldsymbol{P}$ of the form $p_{ij}(\alpha) = p_{ij}^{\alpha}$. The new correspondence model is then denoted as $\boldsymbol{S}(\alpha)$ which is calculated from the transformed matrix $\boldsymbol{P}$, and the CA algorithm is implemented as in all previous cases. For equation 7-14, a pre-transformation can be applied to the contingency ratios $\boldsymbol{Q}$, similarly of the form $p_{ij}(\alpha) = p_{ij}^{\alpha}$. The new correspondence model is then denoted as $\boldsymbol{S}^*(\alpha)$ which is calculated from the transformed contingency ratios $\boldsymbol{Q}$. In the case of $\boldsymbol{S}^*(\alpha)$, the masses r_i and c_j remain equal to their original values regardless of α. In both variants of the CA model, it is possible to standardize the analysis with different values of the power parameter α. This can be achieved by either dividing the singular values of $\boldsymbol{S}(\alpha)$ or $\boldsymbol{S}^*(\alpha)$ by α. Alternatively, standardizing can be achieved by dividing $\boldsymbol{S}(\alpha)$ by α before applying SVD, or by dividing $\boldsymbol{S}^*(\alpha)$ by α before double-cantering and applying the SVD.

The power transformation of equation 7-13 results in an analysis of contingency ratios of the form $(\frac{1}{\alpha})(q_{ij}^{\alpha} - 1)$. In this case the ratios, as well as the weights and double-centring, are all with respect to row and column masses which are all dependent on α. As $\alpha \to 0$, the masses tend to constant values, namely $1/n$ and $1/p$ for the rows and columns respectively. Therefore the limiting case of the power transformation of equation 7-13 is the analysis of logarithms with constant masses, or unweighted LRA.

Similarly to equation 7-13, the power transformation of equation 7-14 results in an analysis of contingency ratios of the form $(\frac{1}{\alpha})(q_{ij}^{\alpha} - 1)$. Due to the Box-Cox transformation, these contingency ratios converge to $\log(q_{ij})$ as $\alpha \to 0$. Therefore the limiting case of the power transformation of equation 7-14 converges to the weighted LRA as $\alpha \to 0$.

For a detailed review on LRA-see (Aitchison & Greenacre, 2002; Greenacre, 2009, 2010, 2011).

7.4 Interpreting the Log-Ratio Biplot

Unlike the biplots of previous chapter, the log-ratio biplot has a unique interpretation. Therefore this subsection is presented solely to provide the reader with the ability to interpret the displays. In the log-ratio biplot it is not the positions of the standard coordinate points which are of interest, but rather the link vectors which join any two of these points. These link vectors represent the pairwise log-ratios between the two points. A hypothetical example is provided below for means of explanation.

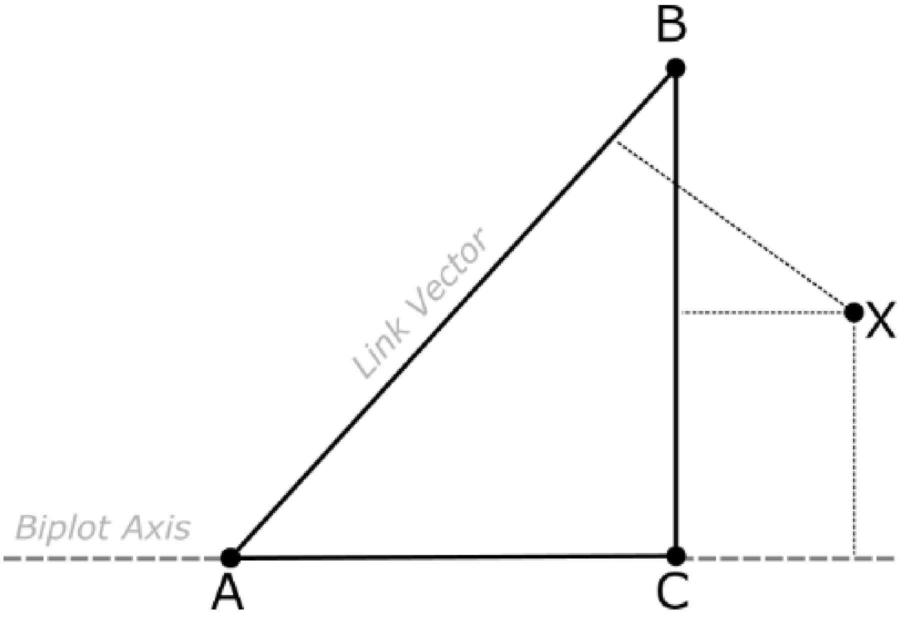

Figure 7-1*: An example of log-ratio biplot interpretation.*

Figure 7-1 displays three standard coordinates, A, B, and C, as well as the links which connect these points. Additionally, principal coordinate X is displayed in the image, along with its projections onto the respective link vectors. The link vector connecting A and B represents the logarithm of A/B. Similarly, the vector can be seen as connecting B and A and hence represents the logarithm of B/A. A biplot axis can be drawn through a link vector so that all observations can be projected upon it. The link vector between A and C requires such an axis, as the projection of X onto this link vector does not lie between points A and C. A biplot axis has been drawn through A and C so that the log-ratio of A/C can interpreted. Point X is then projected on the extreme positive end of the C/A axis, suggesting that point X has a large positive log-ratio of C/A or a large negative log-ratio of A/C.

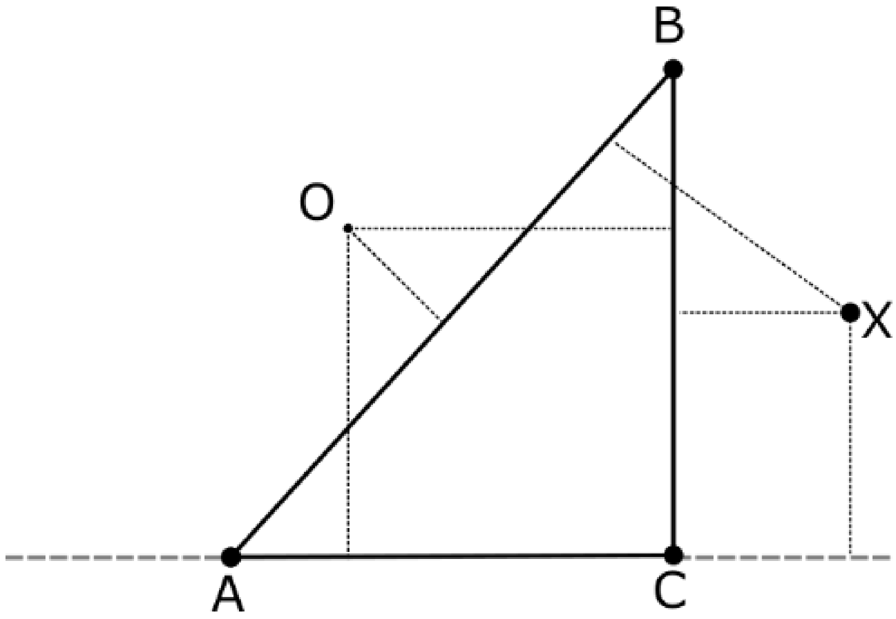

Figure 7-2*: Figure 7-1 with a representation of the origin (O) and its projection onto the respective link vectors.*

Similarly to the CA biplot, the origin represents the average observation. In order to compare point X to the average observation, the origin needs to be projected upon the respective axis of interest. Figure 7-2 represents the same scenario as in Figure 7-1, along with the origin which is represented by O. Additionally, the projections for the origin onto the respective link vectors have also been drawn. The projections represent the log-ratio of the average profile on the respective link vectors. With respect to the A/C link vector, it appears that the average observation has a positive log-ratio. Point X is thus positioned far from the average observation on the negative extreme of the A/C axis. Observation X's log-ratio for all other pairs of variables can be observed in a similar fashion.

7.5 Application of LRA to the South African Crime Data

In this section, the South African crime data which appeared in earlier chapters is subject to a log-ratio analysis. Subsequently, the results are then compared to the findings of Chapter 6. Prior to the implementation of the log-ratio analysis, it can be assumed that the results of the LRA biplots should closely match those of the CA maps. As stated by Greenacre & Primicerio (2013:183), "*when the variance in the data is small, then the CA solution will be close to the LRA solution*". The inertia value was used as the measure of variance in Chapter 6, and it was noted that the South African crime data had a relatively small variance. Following from this, it is expected that the results of the LRA biplots will concur with those of Chapter 6.

7.5.1 The Log-Ratio Biplot of the 2004 South African Crime Data

The weighted log-ratio biplot of the 2004 South African crime data is presented in Figure 7-3. The function `LogRatioBipl()` has been used to produce all the LRA biplots which are presented in this chapter. Additionally, the function code and its respective arguments are presented in the appendix. Figure 7-3 is produced by inputting the 2004 South African crime data into the `LogRatioBipl()` function and subsequently setting `principal=”column”` and `lambda=”True”`. Additionally, all succeeding LRA biplots are produced in a similar manner.

In Figure 7-3, the provinces are in principal coordinates and the various crimes are in standard coordinates. The difference between the weighted and unweighted analysis was stated algebraically in Section 7.2. This analysis makes sole use of weighted log-ratio biplots, the motivation for which is provided for by Greenacre & Primicerio (2013). Greenacre & Primicerio (2013) explain that the mean square of the log-ratio will be higher for rarer components, which can have larger ratios than those between components at a higher level. Similarly to the process of weighting in CA, the weighted log-ratio analysis compensates for the rarer components by assigning weights to each component proportional to its mean, such that the rarer components get smaller weights. Due to the large differences in frequencies between crime categories in the South African crime data, weighted LRA is a more suitable method of analysis.

When observing Figure 7-3 it is apparent that it closely resembles Figure 6-1. This is expected as the total variance from the LRA of the 2004 crime data is low (0.0907). Similarly for the CA counterpart, the total inertia was also very low (0.0856) for the 2004 South African crime data. In a case such as this, where the variance of the table is low and the biplots closely resemble one another, it can be said that the CA is close to being sub-compositionally coherent, and can be deemed appropriate for the analysis of compositional data. Additionally, the quality of Figure 7-3 is 85.26%, which is comparable to that of Figure 6-1 at 83.20%. Although both the CA and LRA of the 2004 South African crime data appear similar at first, due to the inner products of the log-ratio algorithm, Figure 7-3 can be subject to Lambda-scaling to further improve its interpretability. Figure 7-3 has been reproduced with Lambda-scaling in Figure 7-4 and there is a clear improvement in the interpretability of the graph. Although Figure 7-4 is much easier to read, some caution must be taken when analysing the LRA plots due to their unique interpretation.

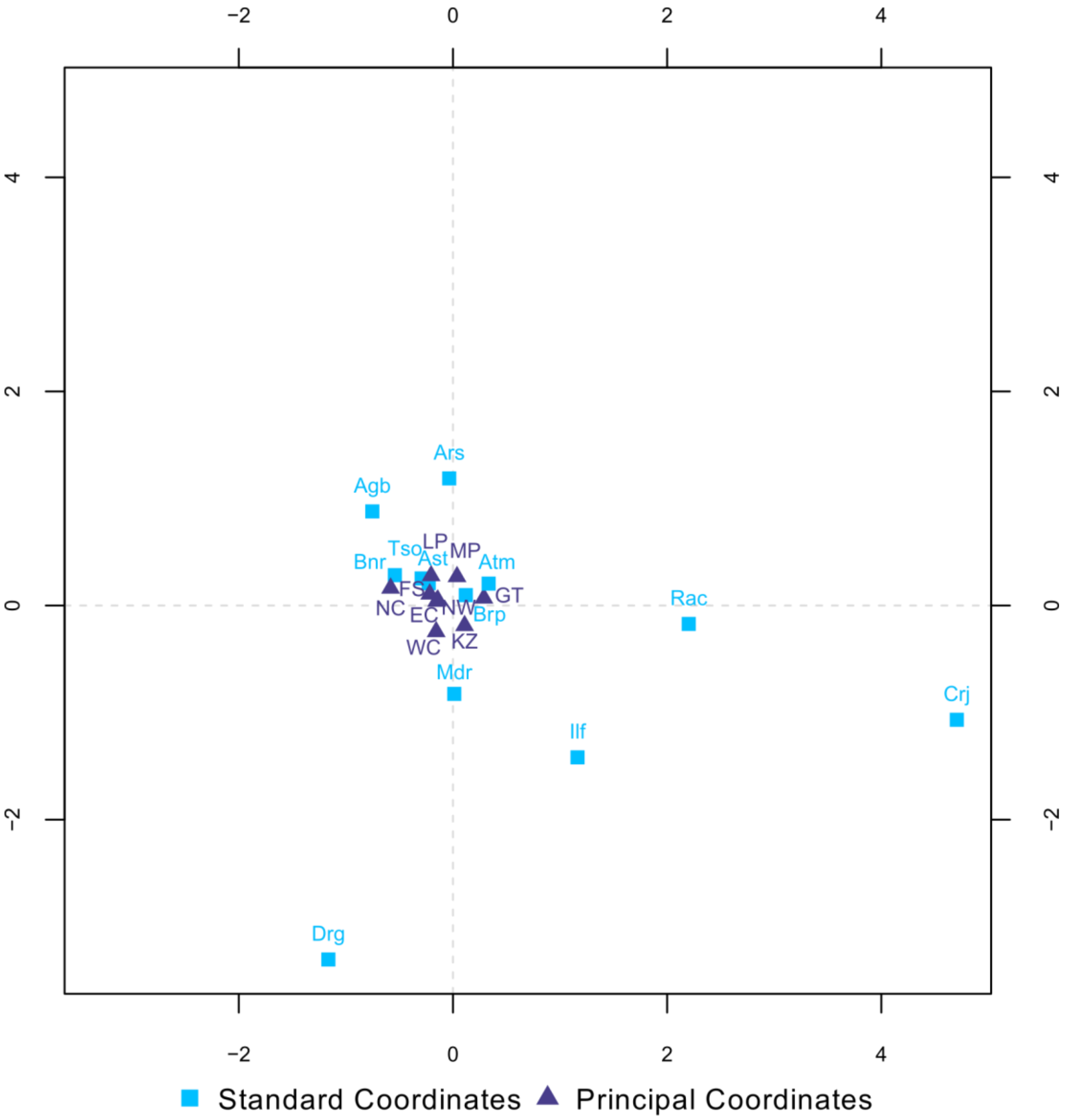

Figure 7-3*: Weighted column principal log-ratio biplot of the 2004 South African crime data without Lambda-scaling.*

As a result of the Lambda-scaling, the profile points have been drawn away from the centroid and their relative positioning made clearer. As noted earlier, it is not the distance between profile points and the standard coordinates which are important in LRA, but the position of the profile point's projection onto the link vector connecting two rays. In order to help visualize this concept, both the rays and a pair of link vectors have been added to the 2004 South African crime data in Figure 7-5. The lengths of the individual rays in Figure 7-5 provide an approximation to the standard deviation of the variables. Following from this it is then clear that drug-related crimes and carjacking have the largest standard deviations of all the crimes. Similarly, the length of the link vectors (dotted lines) provide an approximation of the standard deviation of the corresponding log-ratios. The standard deviations of the log-ratios between respective crime types has been provided for in Table 7-1. From Table 7-1 it is clear that the crimes which are positioned furthest away from the centroid have the largest standard deviations of the pairwise log-ratios between them and other crimes. For example, carjacking is positioned on the extreme right of the first dimension. Due to this, carjacking tends to have long link vectors between itself and any other ray, thus leading to larger standard deviations of the log-ratios. This fact holds true for other variables such as drug-related crimes which are also positioned far away from other variables.

The pair of link vectors drawn in Figure 7-5 plays multiple roles in the interpretation of the log-ratios. Firstly, as previously explained, the length of the link vectors provides an approximation to the standard deviation of the corresponding log-ratios. Secondly, the cosine of the angle between links estimates the correlations between ratios, such that:

$$\cos(\theta) \approx corr\left(\log\left(\frac{x_i}{x_j}\right), \log\left(\frac{x_k}{x_l}\right)\right). \tag{7-15}$$

Lastly, and most importantly, the position of an observation's perpendicular projection upon any link vector provides an approximation of the respective log-ratio for that particular observation.

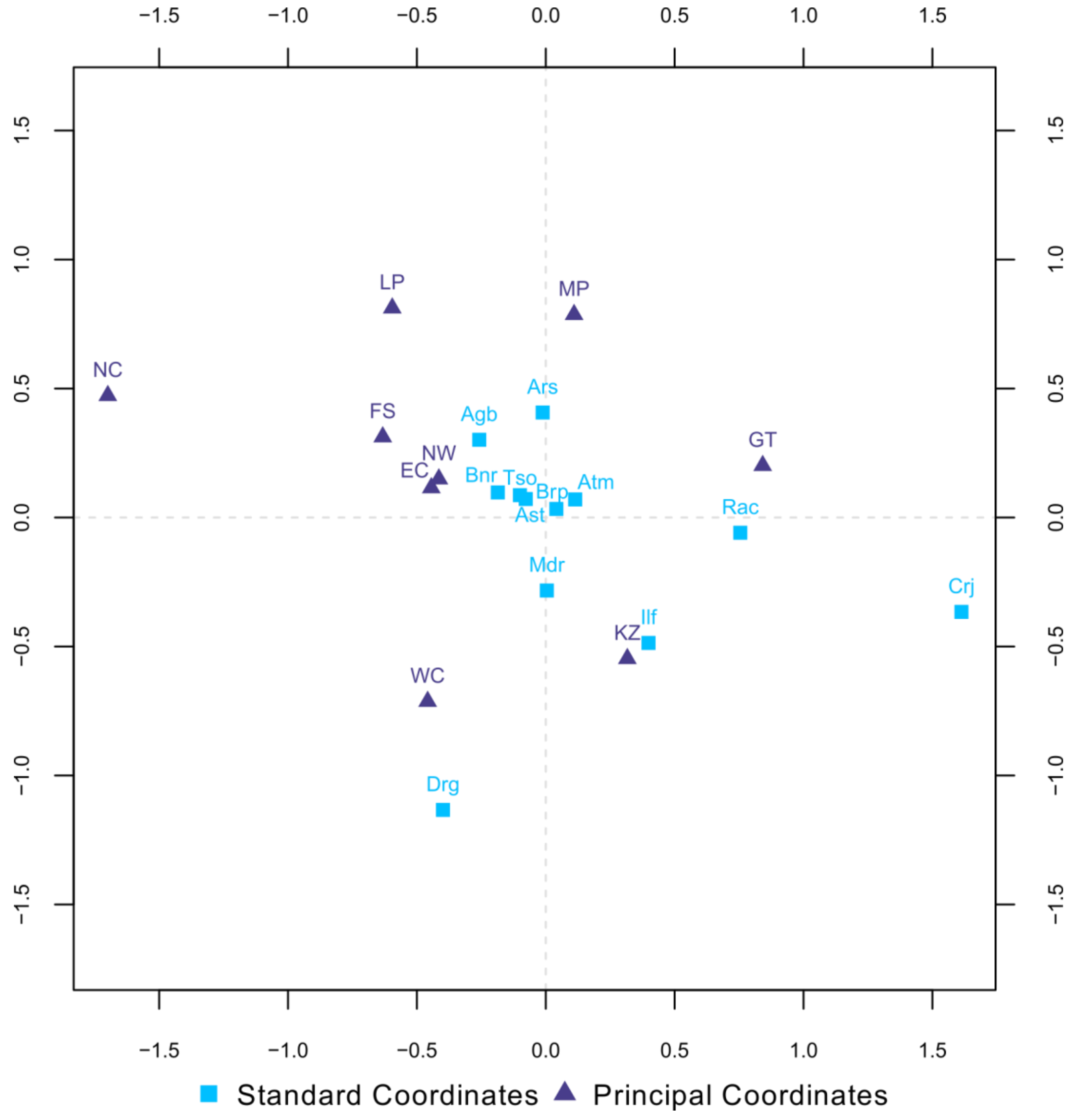

Figure 7-4*: Weighted column principal log-ratio biplot of the 2004 South African crime data with Lambda-scaling.*

***Table 7-1**: Standard deviations between log-ratios of the 2004 South African crime data*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Mdr	0.00	0.36	0.38	0.46	0.47	0.51	0.41	0.48	0.32	0.39	0.59	1.30
Tso	0.36	0.00	0.37	0.23	0.20	0.59	0.25	0.16	0.18	0.61	0.67	1.43
Atm	0.38	0.37	0.00	0.36	0.43	0.65	0.41	0.42	0.38	0.69	0.70	1.53
Agb	0.46	0.23	0.36	0.00	0.31	0.75	0.30	0.23	0.31	0.78	0.74	1.61
Ast	0.47	0.20	0.43	0.31	0.00	0.63	0.40	0.21	0.24	0.70	0.66	1.48
Rac	0.51	0.59	0.65	0.75	0.63	0.00	0.64	0.70	0.48	0.37	0.86	0.92
Ars	0.41	0.25	0.41	0.30	0.40	0.64	0.00	0.32	0.36	0.64	0.85	1.44
Bnr	0.48	0.16	0.42	0.23	0.21	0.70	0.32	0.00	0.31	0.72	0.67	1.54
Brp	0.32	0.18	0.38	0.31	0.24	0.48	0.36	0.31	0.00	0.54	0.65	1.34
Ilf	0.39	0.61	0.69	0.78	0.70	0.37	0.64	0.72	0.54	0.00	0.78	0.94
Drg	0.59	0.67	0.70	0.74	0.66	0.86	0.85	0.67	0.65	0.78	0.00	1.60
Crj	1.30	1.43	1.53	1.61	1.48	0.92	1.44	1.54	1.34	0.94	1.60	0.00
Sum	5.67	5.04	6.32	6.08	5.74	7.10	6.01	5.76	5.11	7.16	8.77	15.13

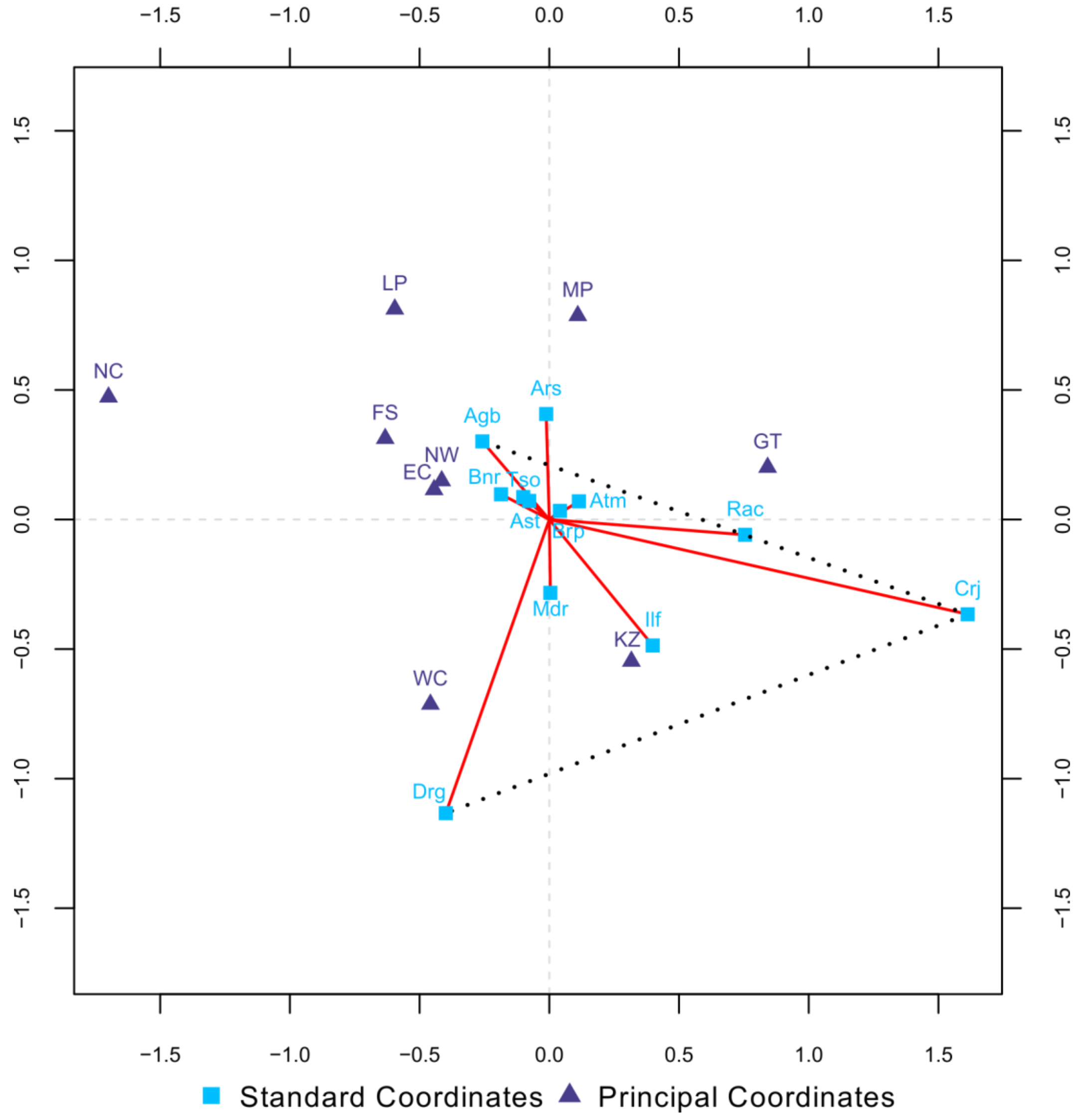

***Figure 7-5**: Weighted column principal log-ratio biplot of the 2004 South African crime data with Lambda-scaling and row vectors.*

The two link vectors in Figure 7-5 represent $\log\left(\frac{Crj}{Agb}\right)$ and $\log\left(\frac{Crj}{Drg}\right)$ respectively. The link between carjacking and assault with the intent to cause grievous bodily harm has the largest standard deviation of all the link vectors (1.61). Additionally, the link between drug-related crime and carjacking has the second largest standard deviation at 1.60. The links with the largest standard deviations have been drawn as it is expected that the greatest differentiation between provinces will be examined upon these links. Only a single pair of the total 66 possible unique links have been drawn. It is infeasible to draw all the links on a single diagram as the figure will quickly become incomprehensible.

By inspection of Figure 7-5, it can be determined that Gauteng, out of all the provinces, is positioned furthest towards the positive end of the $\log\left(\frac{Crj}{Drg}\right)$ link vector. This indicates that, compared to all other provinces, Gauteng has a much higher *Crj/Drg* ratio. The Northern Cape is situated on the extreme negative end of this link vector, indicating a high ratio of $\log\left(\frac{Drg}{Crj}\right)$ or a negative $\log\left(\frac{Crj}{Drg}\right)$ ratio. Interestingly, this relationship between the Northern Cape and drug-related crime was overlooked in both the CA and CA biplot due to the fact that both methods analyse raw frequencies and not ratios. The Northern Cape actually has a low number of reported cases of assault with the intent to cause grievous bodily harm, however, it has by far the lowest number of carjackings. Due to this, the Northern Cape has a high value on the *Drg/Crj* link. KwaZulu-Natal's projection places in between the two extreme provinces and is located approximately at the average profile point on the link between *Drg* and *Crj* (the projection of the origin onto the *Drg/Crj* link).

When analysing the link between assault with the intent to cause grievous bodily harm and carjacking, the Northern Cape is again positioned at the negative extreme of this link, and Gauteng is at the positive. As seen before, the Northern Cape's low value on the $\log\left(\frac{Crj}{Agb}\right)$ link does not come from its high *Agb* value, but rather from the discrepancy between the two variables. In fact the Northern Cape has the lowest frequencies out of all the provinces for both assault with the intent to cause grievous bodily harm and carjacking, but scores very high on the $\log\left(\frac{Abg}{Crj}\right)$ link. Gauteng on the other hand has the highest values for both assault with the intent to cause grievous bodily harm and carjacking. Although Gauteng has a carjacking frequency of more than 1000 times larger than that of the Northern Cape, it is drawn towards the centroid due to its large assault with the intent to cause grievous bodily harm frequency. The complexity to the interpretation of the log-ratio biplot is quickly becoming apparent, yet these unique features of the LRA provide another level to the analysis of the South African crime data.

An interesting feature of the log-ratio biplot is that if any of the rays (variables) fall into a straight line, then this reveals log-ratios of high correlation and a model summarizing this interdependence can be formulated from the relative lengths of their links (Aitchison & Greenacre, 2002:384). For example, *Agb*, *Rac*, and *Crj* all clearly fall onto a straight line in Figure 7-5, and thus the log-ratios formed from these three variables could be linearly related. In general, if three components A, B, and C lie approximately on a straight line with distances AB and BC equal to λ and μ respectively, then the log-contrast is said to be of the form:

$$\mu \log(A) + \lambda \log(C) - (\lambda + \mu) \log(B) = constant$$

(Aitchison & Greenacre, 2002:385).

A further interesting feature, which was previously mentioned, is that the angle cosines between two link vectors approximate the correlations between them. A subset of the correlation matrix between log-ratios is provided for in Table 7-2. Both the lowest and highest correlated pair of log-ratios from Table 7-2 have been drawn in Figure 7-6 and Figure 7-7 respectively. The least correlated pair of log-ratios (*Ars/Mdr* and *Rac/Mdr*) can be seen in Figure 7-6. The two link vectors between this pair of log-ratios are approximately orthogonal, and thus the correlation between them is close to zero. Likewise, the links of the highly correlated pair of log-ratios (*Agb/Rac* and *Bnr/Rac*) are drawn in Figure 7-7. These two links can be seen to be running near parallel to one another and hence the high correlation. Due to the vast number of link pairs that exist, the analysis of the correlation between link vectors is difficult to analyse. It is impractical to draw all pairs of link vectors on a LRA biplot and the correlation matrix is far too large to provide in a report. In such cases it proves easier to analyse the pairwise correlations between link vectors by inspection of the figures instead.

***Table 7-2**: Subset of the correlation matrix between log-ratios for the 2004 South African crime data*

	Bnr/Rac	**Ars/Mdr**	**Agb/Ilf**	**Rac/Mdr**	**Agb/Rac**	**Rac/Crj**
Bnr/Rac	1.0000	0.4943	0.8177	-0.7341	0.9499	0.8116
Ars/Mdr	0.4943	1.0000	0.6080	0.0242	0.4674	0.2912
Agb/Ilf	0.8177	0.6080	1.0000	-0.4489	0.8821	0.9003
Rac/Mdr	-0.7341	0.0242	-0.4489	1.0000	-0.7911	-0.6258
Agb/Rac	0.9499	0.4674	0.8821	-0.7911	1.0000	0.8689
Rac/Crj	0.8116	0.2912	0.9003	-0.6258	0.8689	1.0000

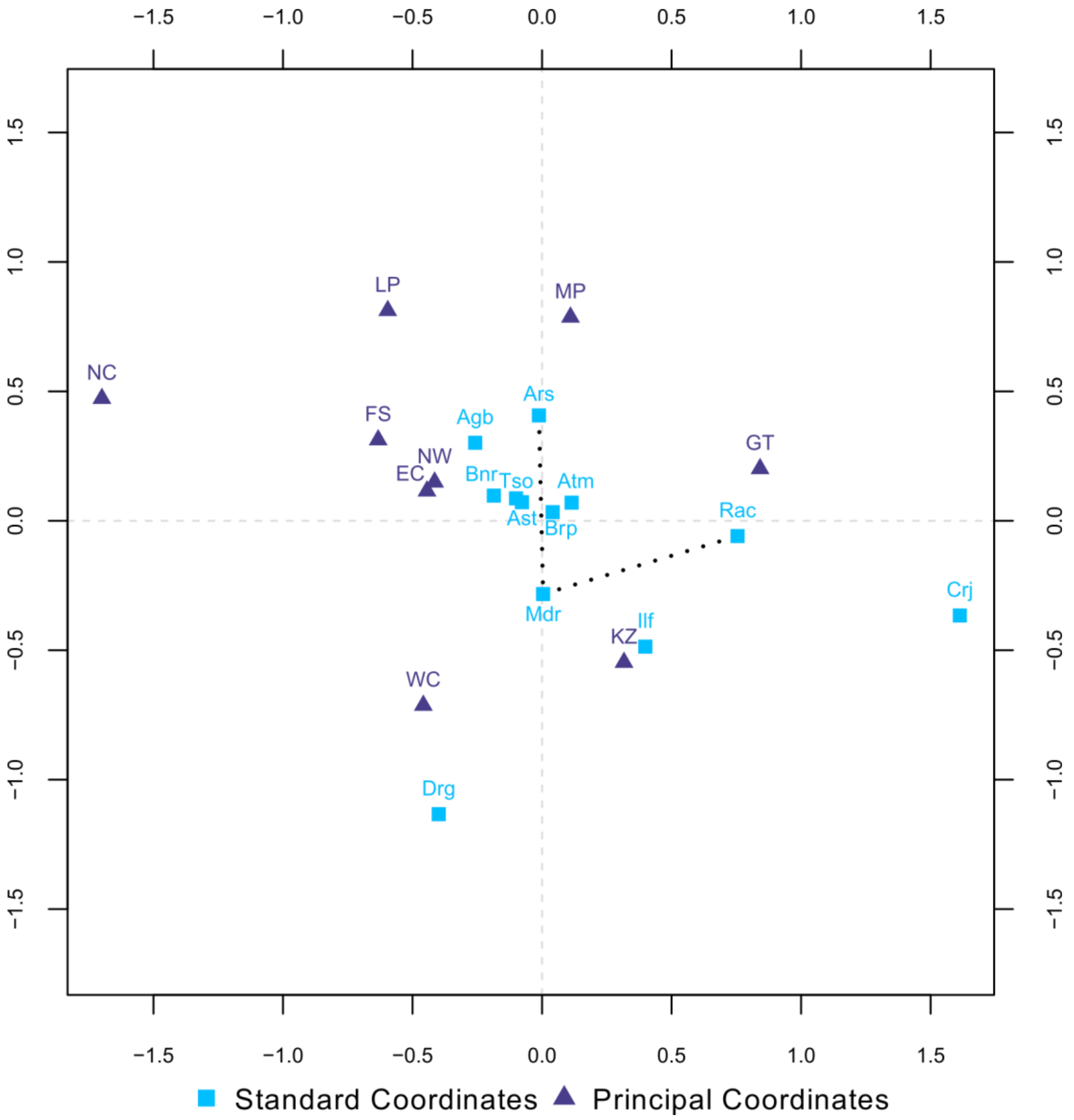

Figure 7-6*: Weighted column principal log-ratio biplot of the 2004 South African crime data with Lambda-scaling, displaying the Ars/Mdr and Rac/Mdr link vectors.*

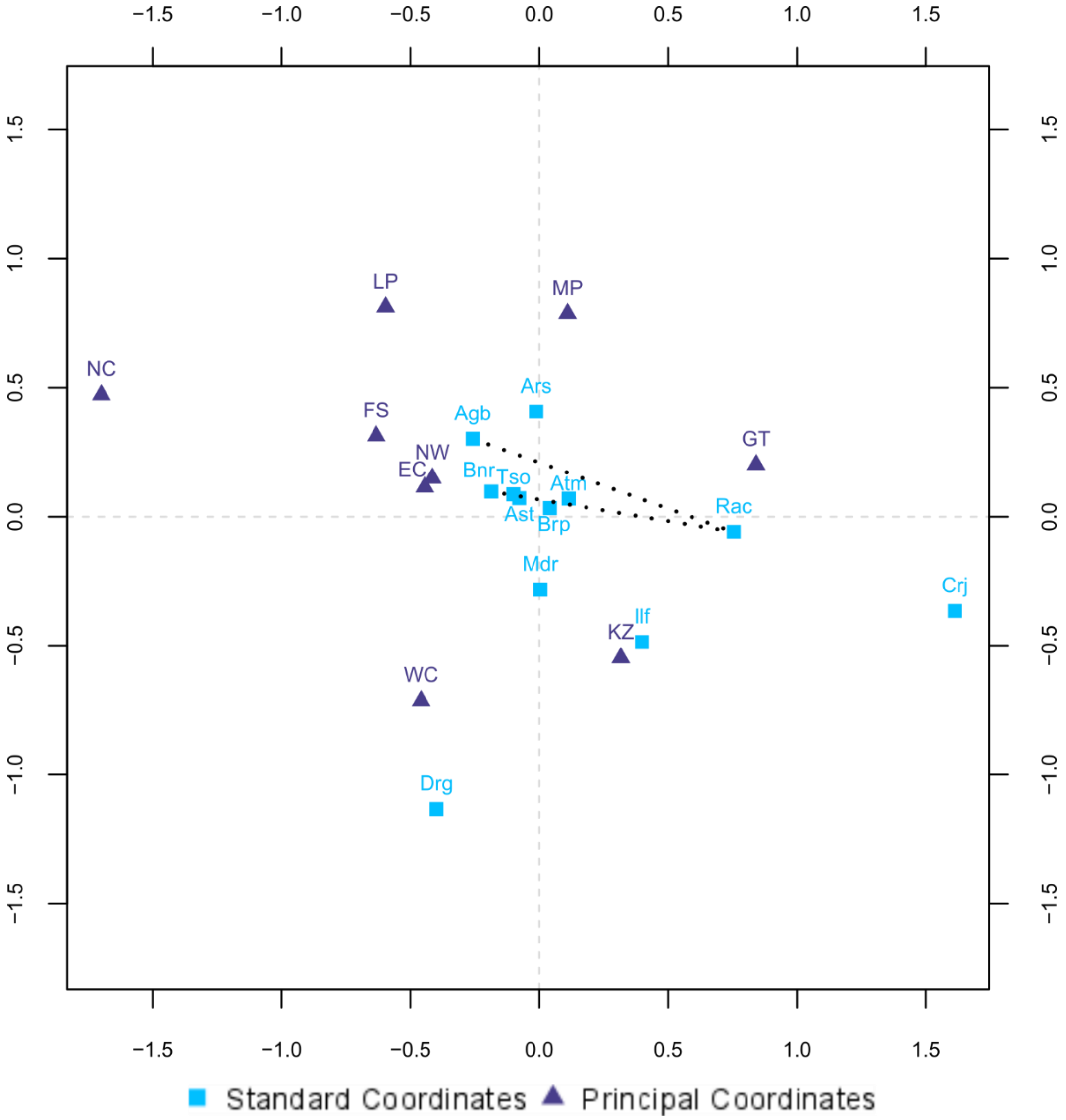

***Figure 7-7**: Weighted column principal log-ratio biplot of the 2004 South African crime data with Lambda-scaling, displaying the Agb/Rac and Bnr/Rac link vectors.*

7.5.2 Comparison of the CA and LRA results

In order to compare the results of the LRA and CA solutions for the South African crime data over the 2004-2013 period, a summary table has been provided for below. Table 7-3 contains the quality of the 2 dimensional approximations, as well as the respective measures of variance for both the CA and LRA solutions for the respective South African crime tables. Additionally, the absolute differences between the quality and variance for the two different techniques has been calculated in Table 7-3. The largest differences for these two measures have also been highlighted in the table.

***Table 7-3**: Summary table of the difference between CA and LRA results*

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
CA Quality (%)	83.20	84.58	87.55	86.91	87.54	88.32	89.05	87.18	85.94	80.41
LRA Quality (%)	85.26	86.45	88.49	87.10	86.89	88.28	88.85	87.36	86.50	85.43
Abs Diff	2.06	1.87	0.93	0.20	0.65	0.04	0.20	0.18	0.56	5.02
CA Inertia	0.09	0.09	0.11	0.11	0.12	0.12	0.13	0.11	0.09	0.08
LRA Variance	0.09	0.10	0.11	0.11	0.12	0.12	0.12	0.10	0.08	0.08
Abs Diff	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00

In general, it can be determined from Table 7-3 that the performance measure of both techniques match up closely. At worst, the two techniques differ by approximately 5% of the total variance accounted for in the two dimensional approximation. The largest difference in terms of quality came from the 2013 data set, and second to that was the 2004 data set. Additionally, and perhaps more importantly, is that the measures of variance for the two methods are approximately equal. The largest differences occur for the years 2010, 2011, and 2012. That is to say that for the entire reported period the results of the CA will be close to the LRA solution and can thus be considered sub-compositionally coherent.

The LRA of the 2004 South African crime data set have already been presented in the previous subsection. It was clear that both the CA and LRA solutions of this particular data set matched closely. Chapter 6 presented the 2004, 2005, 2010, 2012, and 2013 CA solutions. As these have been highlighted in Table 7-3 as having larger differences in variance and quality, there is potential to highlight differences between the results of the two techniques for analysis conducted on the aforementioned years. Additionally, new information about the South African crime data may be gained by analysing the two solutions of these respective years. The following section provides an analysis on the LRA biplots of these respective years.

7.5.3 LRA Biplots of the 2004-2013 South African Crime Data

This section presents the LRA biplots for the South African crime data for the years 2004, 2005, 2010, 2012, and 2013. Additionally, the standard deviation matrices of the crime type log-ratios are provided for each individual LRA. The two columns with the largest standard deviation sums have been highlighted in each one of the standard deviation matrices. It is the crimes of the respective highlighted columns which will receive focus in the analysis. The analysis of the following LRA biplots is limited to general patterns and groupings. As each profile point can be projected upon a total of ½ × 12 × (12-1) = 66 unique link vectors, the analysis can quickly become overbearing. In total there are 9 × 66 = 594 log-ratios for each of the following LRA biplots. It is infeasible to attempt to interpret all possible ratios for each of the LRAs, thus a broader analysis is much more suitable for this section. To accommodate for the high number of link vectors that can be drawn in the LRA biplots, it is beneficial to have interactive software which allows for the drawing of only selected links. The

`LogRatioBipl()` function allows for this. The arguments `draw.line` and `line.selct` are used to specify up to two link vectors to be drawn in the following LRA biplots.

The LRA of the 2005 South African crime data appears in Figure 7-8. The figure has been reflected about the vertical axis and subsequently rotated -90° in order to match Figure 6-8 as closely as possible. The reflection and rotation is achieved by the use of the `reflect` and `rotate.degrees` arguments of the `LogRatioBipl()` function. Just as was the case with the CA and CA biplot, it is possible to freely reflect and rotate the diagrams without affecting the integrity of the graph. The two largest standard deviations in Table 7-4 are for the log-ratios *Crj/Agb* and *Crj/Bnr*. The respective link vectors for these ratios have been added to Figure 7-8. As expected, Figure 7-8 matches very closely to the CA solution in Figure 6-8. At first glance the LRA solutions of the 2004 and 2005 crime data differ considerably. However, Figure 7-5 can be reflected about the vertical axis and subsequently rotated -60° to produce Figure 7-9. When comparing Figure 7-8 and Figure 7-9, it is difficult to distinguish the two apart. As was the case in Chapter 6, there is very little difference between the LRA biplots of the 2004 and 2005 South African crime data.

Although it is the links between crime types which are of importance in a LRA biplot, the points themselves also have a role to play in the interpretation. As stated in the previous section, crime points which are positioned far away from all other points will, in general, have long link vectors between themselves and other crime types, and it may therefore be easier to differentiate provinces with respect to such links. In Figure 7-8 there are two such points, *Crj* and *Drg*. As these points are situated on the extremes of the 2nd and 1st dimensions, any links involving these two crimes will be long, and therefore have large standard deviations for the respective log-ratios. It then follows that Gauteng and KwaZulu-Natal are positioned the highest (in the direction of *Crj*) on a link vector between *Crj* and any other crime type. Thus, by association, Gauteng and KwaZulu-Natal have a much higher proportion of carjacking to any other crime out of all the South African provinces. Similarly, the Western Cape and KwaZulu-Natal are separated from all other provinces by being positioned the highest (in the direction of *Drg*) on any link vector between *Drg* and any other crime type, with the exception of *Rac*, *Ilf*, and *Crj*. Therefore, in general, the Western Cape and KwaZulu-Natal have much larger ratios of drug-related crimes to any other crime type out of all the South African provinces.

***Table 7-4**: Standard deviations between the log-ratios of the 2005 South African Crime Data*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Mdr	0.00	0.35	0.43	0.49	0.48	0.53	0.49	0.49	0.37	0.45	0.59	1.47
Tso	0.35	0.00	0.40	0.22	0.25	0.56	0.25	0.20	0.18	0.63	0.70	1.57
Atm	0.43	0.40	0.00	0.38	0.46	0.61	0.44	0.46	0.43	0.74	0.77	1.67
Agb	0.49	0.22	0.38	0.00	0.32	0.72	0.30	0.26	0.32	0.83	0.81	1.76
Ast	0.48	0.25	0.46	0.32	0.00	0.58	0.42	0.24	0.23	0.70	0.67	1.61
Rac	0.53	0.56	0.61	0.72	0.58	0.00	0.62	0.65	0.44	0.40	0.77	1.11
Ars	0.49	0.25	0.44	0.30	0.42	0.62	0.00	0.35	0.35	0.72	0.93	1.58
Bnr	0.49	0.20	0.46	0.26	0.24	0.65	0.35	0.00	0.31	0.74	0.70	1.67
Brp	0.37	0.18	0.43	0.32	0.23	0.44	0.35	0.31	0.00	0.57	0.68	1.47
Ilf	0.45	0.63	0.74	0.83	0.70	0.40	0.72	0.74	0.57	0.00	0.71	1.05
Drg	0.59	0.70	0.77	0.81	0.67	0.77	0.93	0.70	0.68	0.71	0.00	1.64
Crj	1.47	1.57	1.67	1.76	1.61	1.11	1.58	1.67	1.47	1.05	1.64	0.00
Sum	6.13	5.31	6.79	6.43	5.96	6.98	6.47	6.07	5.35	7.53	8.98	16.61

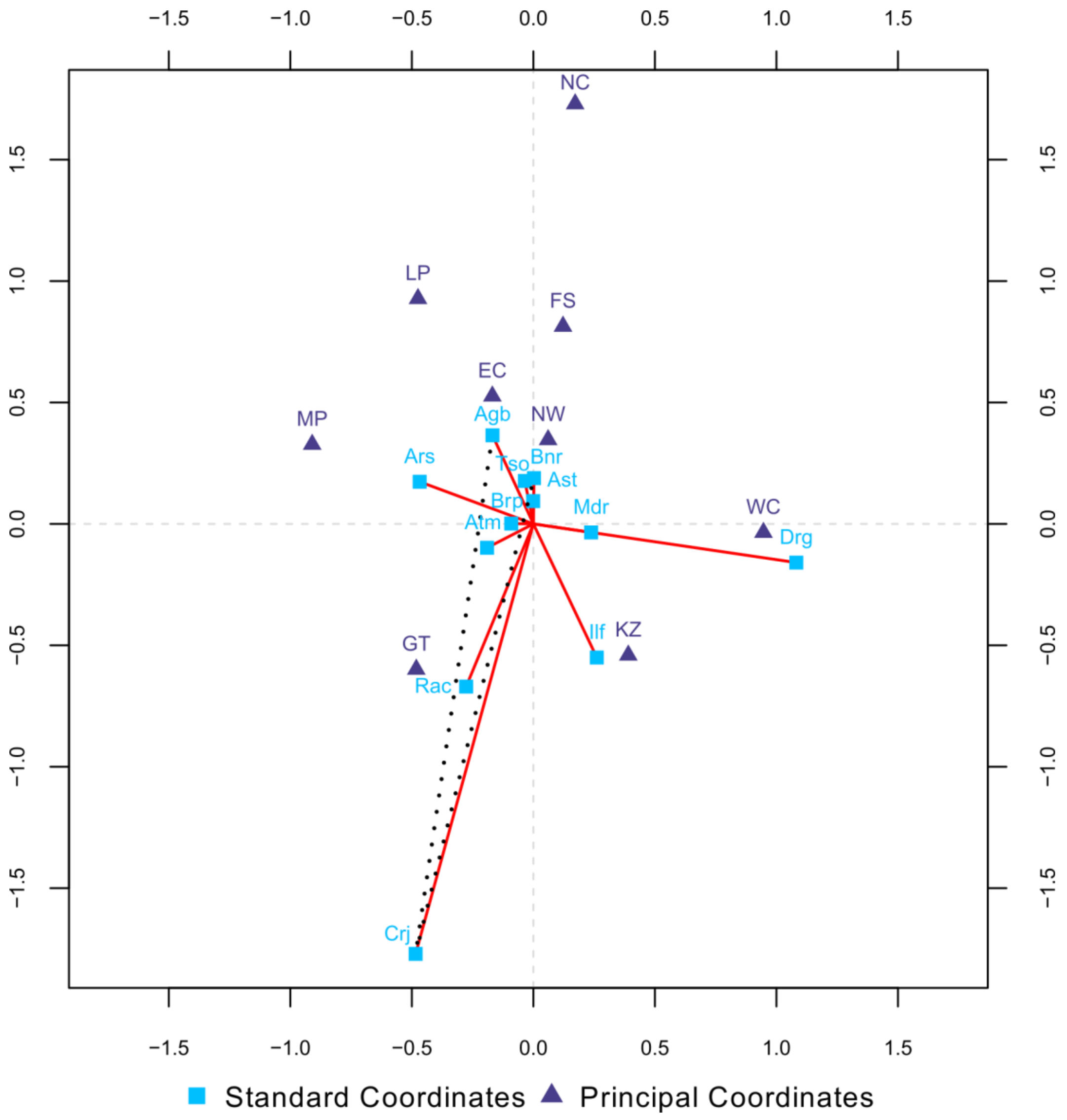

***Figure 7-8**: Weighted column principal log-ratio biplot of the 2005 South African crime data with Lambda-scaling. Additionally, the figure has been reflected about the vertical axis and subsequently rotated by -90°.*

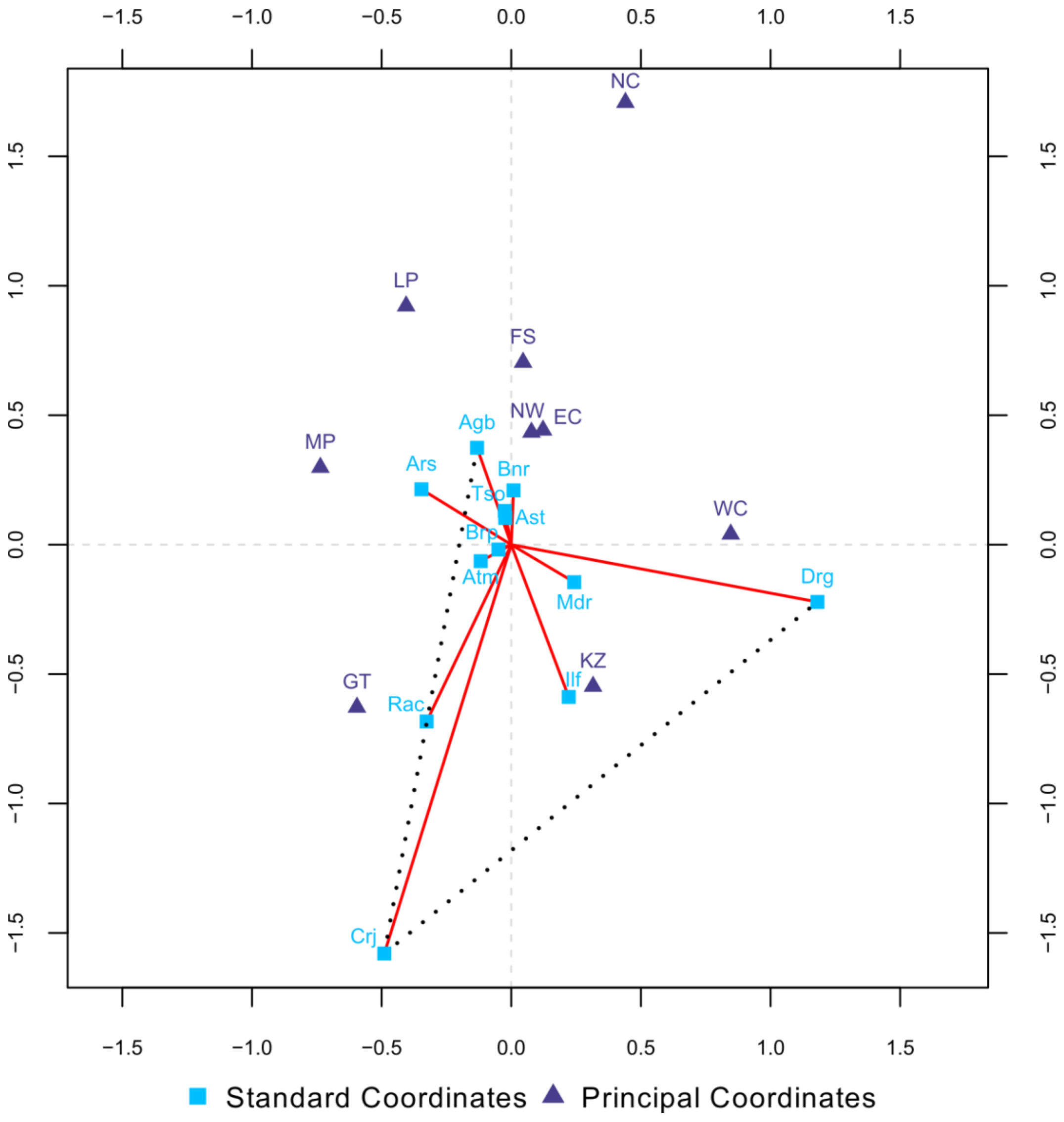

Figure 7-9*: Weighted column principal log-ratio biplot of the 2004 South African crime data with Lambda-scaling. Additionally, the figure has been reflected about the vertical axis and subsequently rotated by -60°.*

The standard deviation calculations for the 2010 South African crime data are presented in Table 7-5. The associated LRA biplot of the 2010 crime data is represented in Figure 7-10. In order to mimic Figure 6-10, Figure 7-10 has been reflected about the horizontal axis. As expected of low variance LRAs, Figure 7-10 does indeed correspond with the CA map of the 2010 data. Additionally, there is minimal change between the LRA biplots of the 2005 and 2010 crime data. The most noticeable difference is that the *Ilf* ray has now lengthened and thus its standard deviation has also increased. The standard deviations of *Drg* and *Ilf* are now approximately equal. The rays of *Rac* and *Crj* are running parallel, indicating high correlation between these two variables. Similarly, *Tso*, *Bnr*, and *Mdr* have shifted such that their corresponding rays are now all highly correlated.

The Northern Cape's profile point has shifted towards the centroid, emphasising a reduction in the distance between itself and other provinces. On the other hand, the Western Cape has become even more distinguished from other provinces by shifting further along the 1st dimension. Gauteng, KwaZulu-Natal and the Northern Cape remain distinct from other provinces with respect to the *Crj/Agb* link. The Western Cape and KwaZulu-Natal also remain separated from other provinces in the positive direction of the *Drg/Ars* link. Additionally, the *Ilf/Agb* link provides clear separation of Gauteng, KwaZulu-Natal, and the Western Cape from all other provinces.

Table 7-5*: Standard deviations between the log-ratios of the 2010 South African Crime Data*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Mdr	0.00	0.30	0.34	0.39	0.39	0.46	0.38	0.46	0.35	0.65	0.67	1.06
Tso	0.30	0.00	0.34	0.20	0.34	0.53	0.19	0.19	0.24	0.80	0.72	1.13
Atm	0.34	0.34	0.00	0.33	0.33	0.55	0.42	0.40	0.37	0.78	0.68	1.15
Agb	0.39	0.20	0.33	0.00	0.37	0.63	0.25	0.27	0.34	0.95	0.82	1.25
Ast	0.39	0.34	0.33	0.37	0.00	0.48	0.48	0.32	0.28	0.78	0.70	1.09
Rac	0.46	0.53	0.55	0.63	0.48	0.00	0.56	0.57	0.37	0.49	0.80	0.68
Ars	0.38	0.19	0.42	0.25	0.48	0.56	0.00	0.31	0.34	0.84	0.85	1.12
Bnr	0.46	0.19	0.40	0.27	0.32	0.57	0.31	0.00	0.22	0.85	0.72	1.16
Brp	0.35	0.24	0.37	0.34	0.28	0.37	0.34	0.22	0.00	0.68	0.67	1.00
Ilf	0.65	0.80	0.78	0.95	0.78	0.49	0.84	0.85	0.68	0.00	0.76	0.66
Drg	0.67	0.72	0.68	0.82	0.70	0.80	0.85	0.72	0.67	0.76	0.00	1.32
Crj	1.06	1.13	1.15	1.25	1.09	0.68	1.12	1.16	1.00	0.66	1.32	0.00
Sum	5.45	4.99	5.70	5.81	5.59	6.13	5.76	5.47	4.87	8.24	8.73	11.63

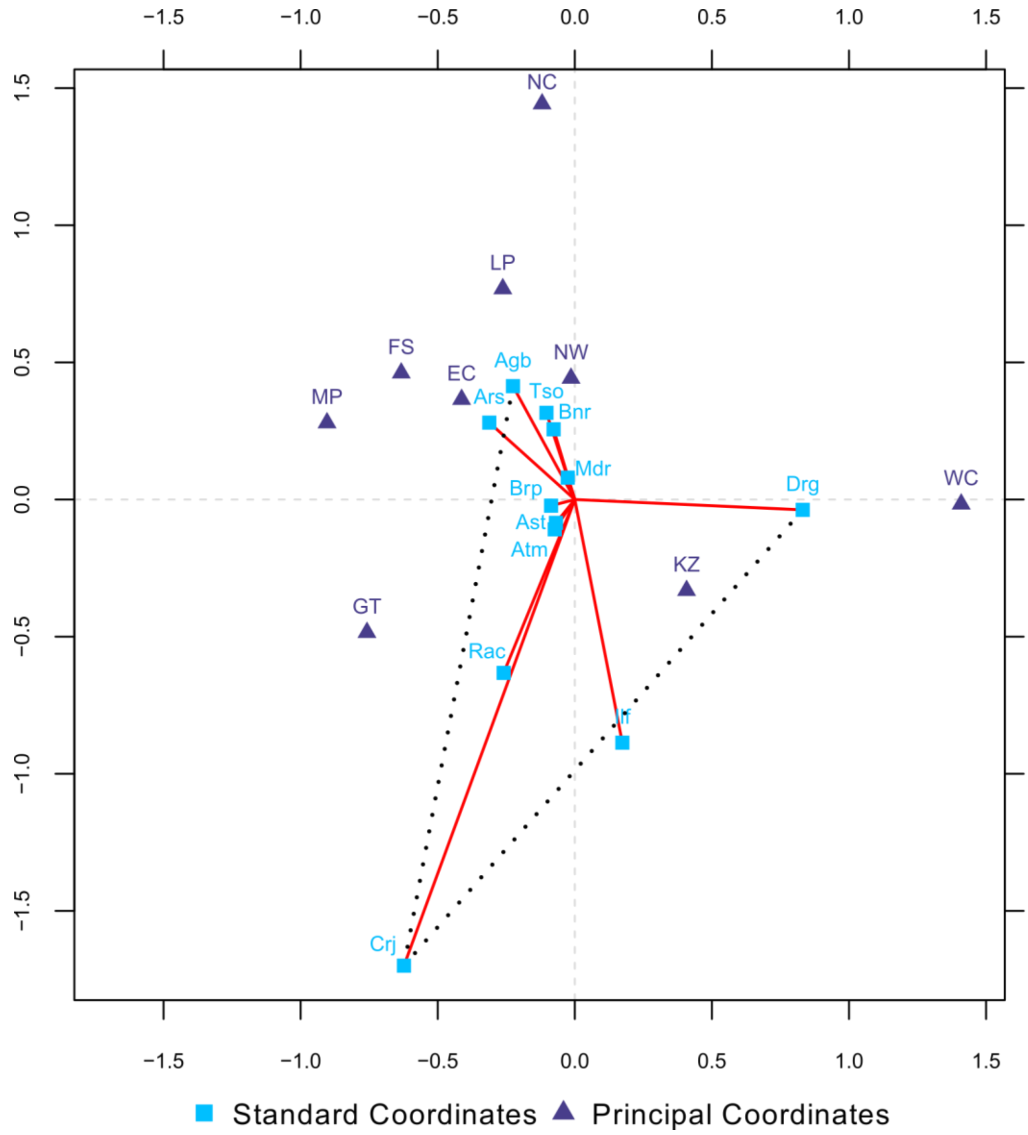

***Figure 7-10**: Weighted column principal log-ratio biplot of the 2010 South African crime data with Lambda-scaling. Additionally, the figure has been reflected about the horizontal axis.*

The standard deviation matrix of the log ratios and the LRA biplot of the 2012 South African crime data appear in Table 7-6 and Figure 7-11 respectively. From inspection of the standard deviations of the log-ratios from 2012 crime data in Table 7-6, illegal firearms has surpassed drug-related crime with respect to the standard deviation between log-ratios. Apart from a slight rotation of the plot there are minimal differences between the LRA biplot of the 2010 and 2012 South African crime data. However, it can still be said that the links for *Crj/Agb* and *Drg/Ars* separate the groups of provinces well. In addition to these two links, the *Ilf/Agb* link makes clear distinctions between Gauteng, KwaZulu-Natal, the Western Cape and all other provinces. There have been some minor shifts in some of the provinces, but a large shift inwards and in the direction of *Agb* for the Northern Cape.

***Table 7-6**: Standard deviations between the log-ratios of the 2012 South African Crime Data*

	Mdr	**Tso**	**Atm**	**Agb**	**Ast**	**Rac**	**Ars**	**Bnr**	**Brp**	**Ilf**	**Drg**	**Crj**
Mdr	0.00	0.37	0.27	0.37	0.43	0.40	0.40	0.43	0.37	0.69	0.59	0.87
Tso	0.37	0.00	0.40	0.23	0.42	0.49	0.16	0.20	0.32	0.80	0.69	0.96
Atm	0.27	0.40	0.00	0.35	0.29	0.40	0.44	0.34	0.28	0.71	0.44	0.89
Agb	0.37	0.23	0.35	0.00	0.39	0.54	0.23	0.23	0.35	0.91	0.73	1.04
Ast	0.43	0.42	0.29	0.39	0.00	0.42	0.48	0.32	0.27	0.76	0.54	0.89
Rac	0.40	0.49	0.40	0.54	0.42	0.00	0.55	0.47	0.28	0.45	0.51	0.54
Ars	0.40	0.16	0.44	0.23	0.48	0.55	0.00	0.28	0.40	0.86	0.74	1.00
Bnr	0.43	0.20	0.34	0.23	0.32	0.47	0.28	0.00	0.22	0.80	0.61	0.97
Brp	0.37	0.32	0.28	0.35	0.27	0.28	0.40	0.22	0.00	0.62	0.50	0.79
Ilf	0.69	0.80	0.71	0.91	0.76	0.45	0.86	0.80	0.62	0.00	0.60	0.43
Drg	0.59	0.69	0.44	0.73	0.54	0.51	0.74	0.61	0.50	0.60	0.00	0.86
Crj	0.87	0.96	0.89	1.04	0.89	0.54	1.00	0.97	0.79	0.43	0.86	0.00
Sum	5.18	5.03	4.81	5.39	5.22	5.06	5.56	4.85	4.41	7.63	6.82	9.24

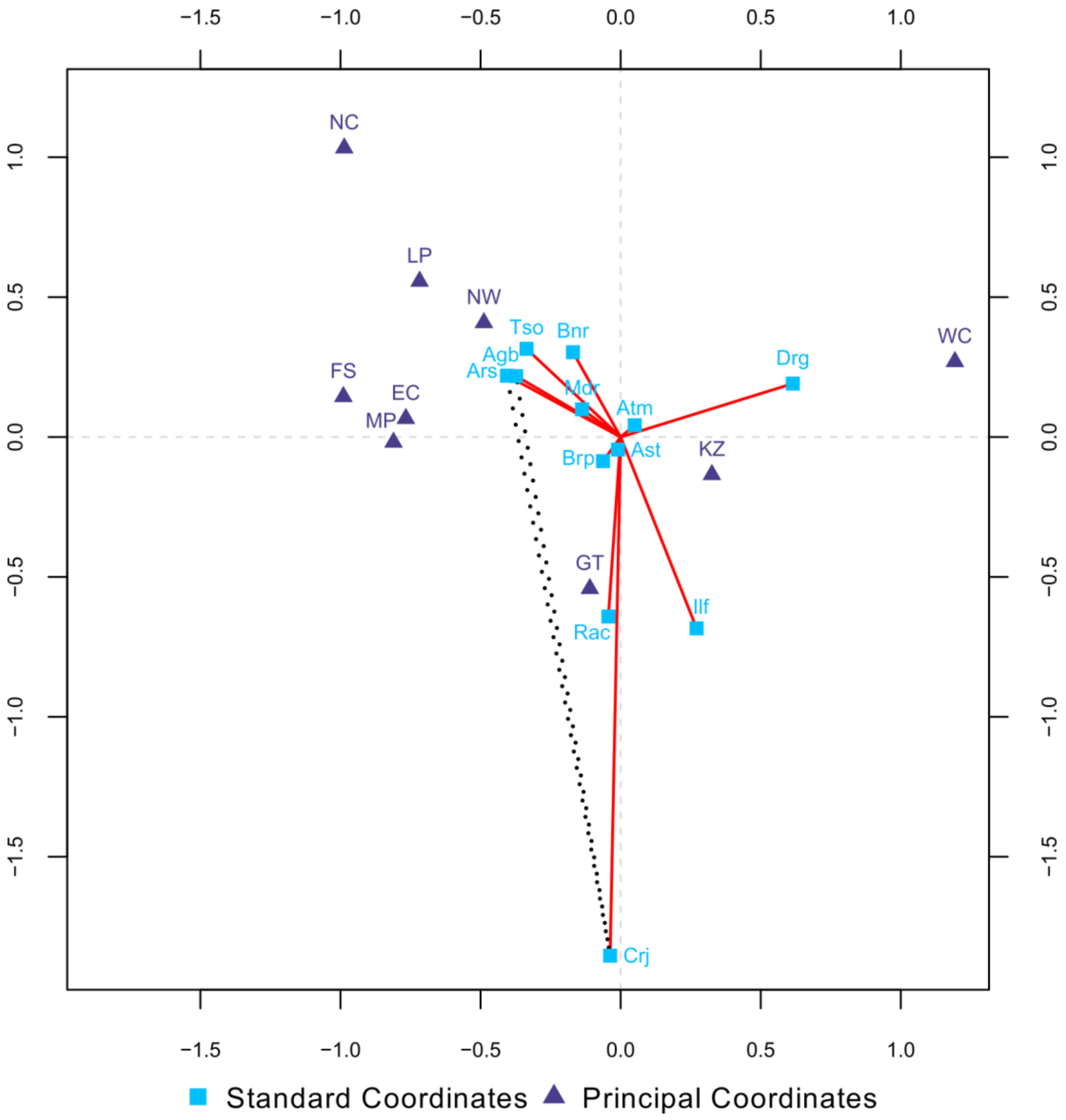

***Figure 7-11**: Weighted column principal log-ratio biplot of the 2012 South African crime data with Lambda-scaling.*

The final LRA to be presented is that of the 2013 South African crime data. The LRA of the 2013 South African crime data and its respective log-ratio standard deviation matrix are presented in Figure 7-12 and Table 7-7. As was the case in Chapter 6, the first substantial difference in the layouts of the LRA biplot occurs for the 2013 data. Again, the LRA of the 2013 data corresponds closely to its CA map counterpart in Figure 6-12. From inspection of Table 7-7, it can be deduced that the standard deviation with respect to log-ratios involving carjacking has increased, further distinguishing itself from the other crime types. Additionally, the standard deviation of log-ratios involving illegal firearms has well surpassed that of drug-related crimes. Although the *Crj/Agb* link has the largest standard deviation, the clearest distinction between provinces can be observed with respect to the *Drg/Ars* link. From this link, a clear distinction is made between Gauteng, KwaZulu-Natal, and the Western Cape whom all have much larger *Drg/Ars* ratios than all other provinces. The same can be said for the *Drg/Mdr*, *Drg/Agb*, and *Drg/Tso* links.

Table 7-7*: Standard deviations between the log-ratios of the 2013 South African Crime Data*

	Mdr	Tso	Atm	Agb	Ast	Rac	Ars	Bnr	Brp	Ilf	Drg	Crj
Mdr	0.00	0.38	0.28	0.35	0.45	0.39	0.37	0.42	0.37	0.56	0.57	0.88
Tso	0.38	0.00	0.41	0.26	0.48	0.51	0.16	0.25	0.34	0.73	0.71	1.00
Atm	0.28	0.41	0.00	0.31	0.35	0.41	0.43	0.35	0.29	0.65	0.46	0.93
Agb	0.35	0.26	0.31	0.00	0.40	0.54	0.26	0.26	0.34	0.81	0.70	1.06
Ast	0.45	0.48	0.35	0.40	0.00	0.45	0.52	0.34	0.30	0.72	0.50	0.93
Rac	0.39	0.51	0.41	0.54	0.45	0.00	0.50	0.47	0.32	0.39	0.38	0.55
Ars	0.37	0.16	0.43	0.26	0.52	0.50	0.00	0.31	0.39	0.76	0.72	0.98
Bnr	0.42	0.25	0.35	0.26	0.34	0.47	0.31	0.00	0.19	0.73	0.59	0.98
Brp	0.37	0.34	0.29	0.34	0.30	0.32	0.39	0.19	0.00	0.57	0.46	0.83
Ilf	0.56	0.73	0.65	0.81	0.72	0.39	0.76	0.73	0.57	0.00	0.54	0.49
Drg	0.57	0.71	0.46	0.70	0.50	0.38	0.72	0.59	0.46	0.54	0.00	0.74
Crj	0.88	1.00	0.93	1.06	0.93	0.55	0.98	0.98	0.83	0.49	0.74	0.00
Sum	5.02	5.24	4.86	5.30	5.42	4.92	5.39	4.89	4.39	6.95	6.37	9.37

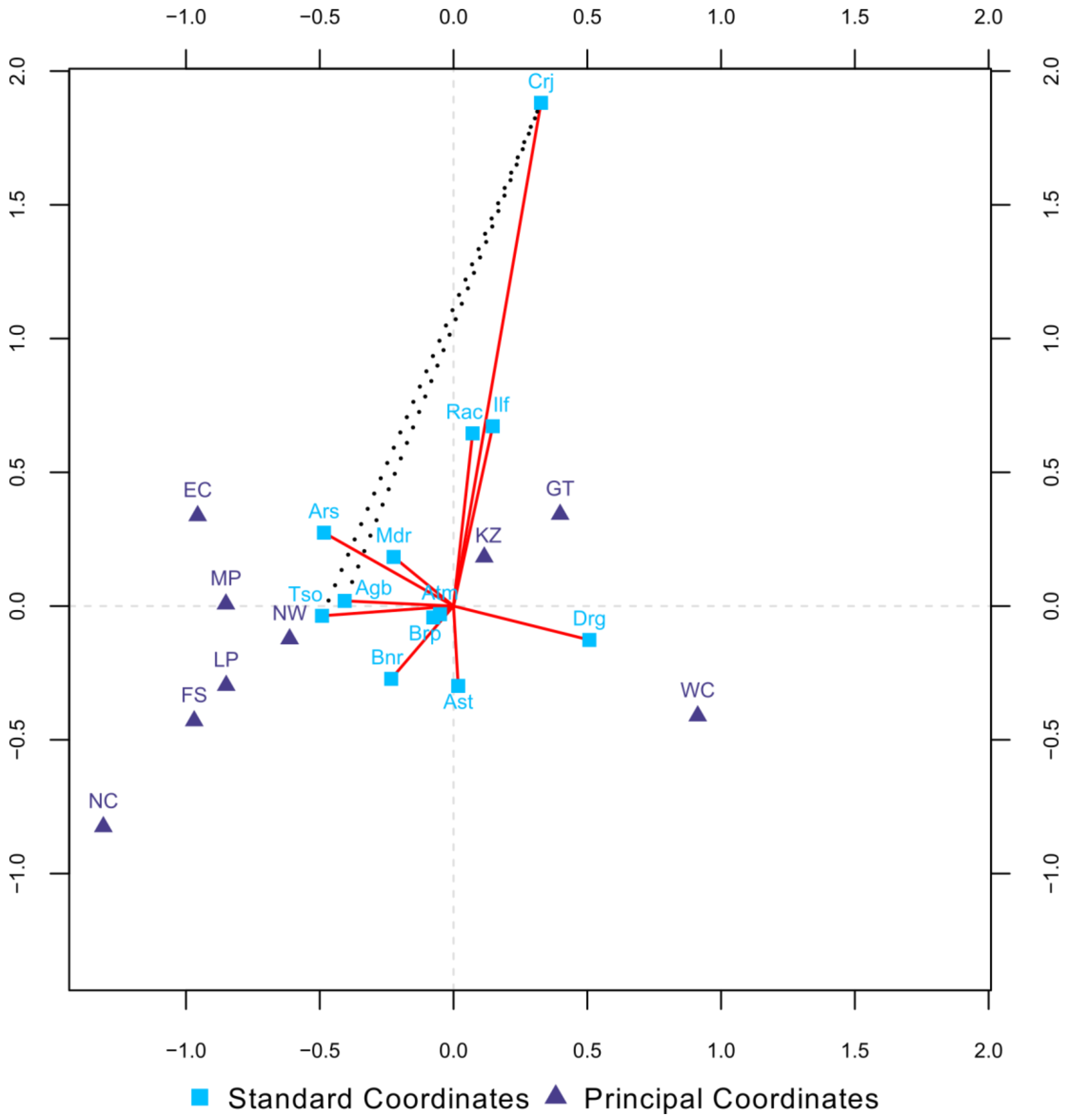

***Figure 7-12**: Weighted column principal log-ratio biplot of the 2013 South African crime data with Lambda-scaling.*

7.6 Summary and Conclusions

Due to the low variance in the South African crime data over the 2004-2013 period, the log-ratio analysis drew similar conclusions to that of Chapter 6. However, the LRA biplots provided additional insight into the South African crime problem. By analysing the ratios of crimes within provinces, associations that were overlooked in previous methods are brought to light in Chapter 7. The Northern Cape is repetitively mentioned for its notably high ratio values with respect to *Drg/Crj* and *Agb/Crj*. This result was previously overlooked due to the low frequencies in the Northern Cape for the aforementioned crimes. The case of the Northern Cape is a prime example of when methods such as the LRA biplot can provide additional information about the data, where they were previously overlooked.

The LRA biplot proved to have a complex interpretation. Unlike the CA maps, the inter-point distances are not a good measure of association. Due to the fact that it is the orthogonal projection of the observations onto the link vectors that is of importance, often provinces which are seemingly far away from one another, are located close together once projected orthogonally onto the link vector. Thus, regardless of how far away two observations are from one another, if a link vector intersects the line between them perpendicularly, both observations will have similar log ratios with respect to such a link. This result holds for the CA biplot too.

The LRA biplot proved useful compared to the CA maps, as it allowed for Lambda-scaling of the points, as well as numerous axes to be drawn between the vectors. Conversely, it is difficult to evaluate the benefit of the LRA biplot over the CA biplot in low variance data sets, as both methods have similar results. Both methods allow for Lambda-scaling, along with calibrated biplot axes. However, due to the vast number of possible axes in LRA, it is not advisable to draw all possible biplot axes in a single figure. Thus, the benefit of this feature can be lost in wide data sets. The CA biplot on the other hand will only have p calibrated axes, and it is common practice to display all axes in a single figure. Additionally, unlike in the LRA biplot, every axis passes through the origin and thus the average profile is always located at the centre of the axis. Whereas with the LRA biplot, it is necessary to orthogonally project the origin onto a link vector to determine the position of the average profile.

It can thus be concluded that, even in the case of data sets with small variances, the LRA biplot proves beneficial in the analysis. However, it is expected that the benefit to using the LRA biplot will increase for high variance data sets. When the variance in the data set is low, the CA map and CA biplot provide an adequate analysis of the data, and the use of LRA biplots is not essential. However, when the variance in the data is high, it is advisable to make use of either one of the CA methods and the LRA biplot. Due to reasons explained in Chapter 6, it is advisable to use the CA biplot over the CA map, when the analysis paired with the LRA biplot.

Chapter 8: Discussion and Conclusions

This study aimed to explore the application of geometric data analysis as a useful method of investigating crime statistics in South Africa. The investigation began with an overview of homicide and violent crime in both South Africa and across the globe, and further progressed to investigate the issues associated with the collection of such data. The analysis of the South African crime data was split into three distinct categories: the univariate analysis, the bivariate analysis, and the analysis of compositional data.

The univariate analysis provided an overview of the South African crime data for the 2004-2013 period, as well as a demonstration of what can be regarded as standard reporting methods for crime statistics. Several shortcomings of the univariate analysis were highlighted, which included the constraints on the number of variables which can be included in each figure, the lack of meaningful interpretation between points, and the large number of figures required to produce an in depth analysis. Additionally, the mosaic plot was also introduced in the univariate analysis as a possible solution to the aforementioned issues. However, said plots failed to provide a substantial improvement upon the remaining univariate techniques, and were still subject to the same limitations in the case of the South African crime data.

The bivariate analysis proposed two new reporting methods, correspondence analysis (Greenacre, 2007) and correspondence analysis biplots (Gower *et al.*, 2011). These two techniques provided a geometric setting where both the crimes and provinces could be represented in a single diagram and the relationships between both sets of variables could be analysed. The CA biplot proved to be a much needed improvement on the correspondence analysis maps, as it can display numerous metrics, provide multiple calibrated axes, and allows for greater manipulation of the figure itself.

The analysis of compositional data was addressed in Chapter 7. The log-ratio analysis (Aitchison & Greenacre, 2002) is the suggested method for analysing compositional data as such analysis is said to be sub-compositional coherent, and is thus appropriate for study of compositional data. The LRA biplot proved useful in the analysis of the South African crime data as it expressed differences on a ratio scale as multiplicative differences. The log-ratios then provided relative comparisons of the crimes within and between provinces, as well as details about the data not seen before.

From the use of the aforementioned analytical methods, a clear understanding of the South African crime data was achieved. The most relevant results from the univariate analysis are presented in a short list:

- It was clearly evident that Gauteng had the largest crime frequencies of all the provinces over the reported period.
- When the crime frequencies per capita were analysed, the Western Cape had a substantially higher rate than all other provinces.

- For a majority of the provinces, there appeared to be three crimes which experienced large decreases. These were: burglary at residential premises, common assault, and assault with the intent to cause grievous bodily harm. The exceptions of which were KwaZulu-Natal and Limpopo, which experiences slight increasing trends in burglary at residential premises.
- KwaZulu-Natal and Gauteng are noted for having robbery with aggravating circumstances frequencies much larger than most other provinces.
- Additionally, KwaZulu-Natal and Gauteng appear to share a small, albeit notable, association to carjacking.
- Over the past decade, South Africa experienced a generally downward trend in murder. However, there has been a spike in recent years. This appears to have been driven mostly by the large increase in reported cases of murder in the Western Cape. Gauteng and KwaZulu-Natal also experience increases in murder, however, smaller than that of the Western Cape.
- Over the past decade, South Africa experienced a considerable increase in drug-related crimes across all provinces. In particular, the Western Cape has, by far, the largest number of reported drug-related crimes. Additionally, Gauteng experienced a particularly steep increase in drug-related crime between 2010 and 2013.

From the application of the geometrical statistical methods, it was determined that most of the South African provinces do not diverge significantly from the average profile. However, the Western Cape and Gauteng are noted for their consistently large deviations away from the average province profile, specifically in the directions of drug-related crime and robbery with aggravating circumstances respectively. Baring KwaZulu-Natal, the remaining provinces form a tight cluster near the average profile. The KwaZulu-Natal however, consistently appears to be positioned between Gauteng and the Western Cape. Although, interestingly, the KwaZulu-Natal never makes substantial contributions to the total inertia over the reported period.

The most notable findings of this study concur with the analysis conducted in Chapter 7 of *Understanding Biplots* (Gower *et al.*, 2011). In Chapter 7 of their book, Gower et al. conduct an analysis on the 2001-2007 South African crime data. The most notable conclusions thereof was Gauteng's association to robbery with aggravating circumstances and carjacking, and Western Cape's high levels of drug-related crimes. Similarly to this study, the remaining provinces did not deviate away from the average profile.

The geometric data analysis proved extremely useful in the study of the South African crime data. Clear associations were seen between provinces and crimes. Additionally, the progression for an individual province could be clearly represented for the reported period in a single, highly informative figure. It is therefore the recommendation of the author that geometric data analysis be used in the study of crime statistics. Additionally, due to the fact that the bivariate geometric data analysis proved to be a great improvement upon traditional univariate techniques, it is suggested that further research into multi-way data analysis and its potential applications to the study of crime data be investigated, see (Kroonenberg, 2008).

Appendix A: The South African Crime Data Tables

This section presents the 10 contingency tables of the South African crime data for the 2004-2013 reporting period. The data used in this study is provided for by the South African Police Service (SAPS). Crime statistics are provided for on a yearly basis by the SAPS, and are readily available from www.saps.gov.za.

***Table A-1**: 2004/2005 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3352	902	3818	5001	733	1099	800	408	2680
Tso	8626	4972	16333	12122	5070	4674	4610	2212	10498
Atm	3046	1324	6661	5979	1032	1568	1065	1351	2490
Agb	40447	17998	54138	33898	17756	19978	18097	13188	33869
Ast	25763	25197	72484	37852	18148	15736	14612	9326	48739
Rac	9576	4532	57628	25207	3285	6947	5376	1095	13143
Ars	1417	503	1985	1470	762	579	506	242	720
Bnr	6415	4063	12986	8990	4994	2992	4288	2370	8950
Brp	33364	17802	77383	43122	13569	21216	15378	7353	46977
Ilf	1938	432	3974	4880	563	792	549	122	2247
Drg	9061	4063	10722	19290	1786	1714	4383	2550	30432
Crj	529	156	7230	2703	203	485	221	6	901

***Table A-2**: 2005/2006 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3669	872	3621	4910	688	882	760	393	2750
Tso	10312	4559	15676	11932	4671	4756	4546	1993	9631
Atm	2525	1042	5761	5302	823	1343	843	1058	1856
Agb	37398	16895	51371	31170	15946	18671	15505	11507	28479
Ast	22713	22417	63012	34059	14679	14196	11424	6827	38226
Rac	9202	4166	52437	24668	3020	6607	5550	1131	12945
Ars	1310	443	2003	1274	689	626	473	209	595
Bnr	6032	4107	13274	9005	4782	2932	4294	1997	7944
Brp	33134	17507	75243	40844	12839	20360	15506	6102	41000
Ilf	1540	387	3388	4320	447	579	451	89	2252
Drg	7511	5074	14202	23206	1977	1794	5053	2085	34788
Crj	523	96	7250	3079	167	515	226	4	965

***Table A-3**: 2006/2007 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3626	953	3884	5002	747	869	823	417	2881
Tso	9117	4386	15124	11649	4780	4631	4588	1957	8969
Atm	2150	998	5741	5245	829	1305	817	1011	2046
Agb	35503	15999	50084	30792	15752	18005	15114	10876	25905
Ast	21230	19883	58915	31739	13178	13522	10196	6311	35083
Rac	9262	4284	55329	26206	3216	6669	5270	1096	15226
Ars	1274	468	2099	1281	699	669	528	215	625
Bnr	5398	4110	14722	9441	4763	3490	4321	2075	10118
Brp	31699	16115	67886	39655	12427	19503	13732	5506	43142
Ilf	1445	341	3920	4614	422	585	425	76	2526
Drg	7231	462	12582	26228	2178	2068	5759	2114	41067
Crj	608	123	7337	3562	196	597	261	4	911

***Table A-4**: 2007/2008 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3526	879	3766	4702	696	835	825	422	2836
Tso	9087	4396	15398	11355	4528	4169	4513	1749	8623
Atm	2166	939	5313	4940	722	1271	825	775	1844
Agb	34571	16833	48076	30514	13670	16849	14778	9898	24915
Ast	19979	19885	58000	29306	11024	12202	9559	5431	32663
Rac	8951	4501	51280	24278	2447	5907	5218	1175	14555
Ars	1237	432	1864	1320	573	588	584	169	629
Bnr	5993	4418	15321	10211	5401	4273	4783	1956	10639
Brp	29628	15705	63799	37083	11857	18855	13626	4924	42376
Ilf	1441	311	3486	4325	473	524	480	90	2346
Drg	8003	4525	12742	24100	3198	1770	6610	2201	45985
Crj	604	156	7489	3889	203	664	268	5	923

***Table A-5**: 2008/2009 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3260	910	3963	4747	751	902	858	411	2346
Tso	9456	4523	18176	13279	4675	4695	5021	1917	8772
Atm	1996	922	5207	4922	701	1265	788	731	1766
Agb	31428	15884	48257	30119	13221	17062	14759	9961	23086
Ast	16895	19605	58566	29906	9753	11491	9336	5606	31680
Rac	9814	5164	51251	25856	2815	6952	5592	1219	12729
Ars	1056	393	1747	1200	584	639	525	178	524
Bnr	6212	5511	17563	11173	6343	5329	5370	2058	10450
Brp	28572	16202	69300	37650	12398	19839	14319	5416	42920
Ilf	1525	324	4040	4236	461	589	482	74	2314
Drg	8437	4561	13574	23819	3316	1642	7109	1933	52781
Crj	706	255	7662	4062	289	984	252	7	698

***Table A-6**: 2009/2010 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3218	910	3444	4224	762	878	743	381	2274
Tso	9047	4581	15645	13269	4905	4603	4759	1845	9678
Atm	1941	845	4800	4614	726	1227	839	711	1707
Agb	32247	15745	49082	30884	13321	15864	14556	9533	24061
Ast	17313	19016	58956	32980	8939	11203	8893	5574	34410
Rac	9677	4969	47289	23239	2968	6611	5422	1037	12543
Ars	1125	400	1597	1202	616	575	420	168	598
Bnr	6424	5197	17904	11314	6253	5539	5361	2237	11544
Brp	28384	15825	74902	40393	13980	19350	14893	5550	43300
Ilf	1426	332	4113	4968	461	621	390	50	2181
Drg	8946	5110	14729	28693	4837	2041	7704	2371	60409
Crj	606	316	7444	3715	251	709	273	13	575

***Table A-7**: 2010/2011 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3187	963	3257	3749	665	722	744	342	2311
Tso	9380	4838	13987	12793	4883	4442	4706	1868	9299
Atm	1718	770	4104	3915	652	820	703	649	2162
Agb	30804	15439	46600	30582	12934	14436	14082	9002	24723
Ast	16587	17934	54476	32271	8120	10340	7786	5099	33278
Rac	10448	4853	40052	19573	2766	5550	5080	891	12250
Ars	1113	374	1624	1141	563	405	503	178	632
Bnr	6515	5110	16757	10984	5876	5235	4987	2036	11582
Brp	27251	14927	70794	39550	13420	18115	14777	4995	43801
Ilf	1465	335	3665	5072	404	544	375	61	2551
Drg	9566	4209	16457	32457	4634	3178	7166	2418	70588
Crj	527	234	5936	2619	177	427	236	14	457

***Table A-8**: 2011/2012 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3278	963	3012	3422	735	729	802	368	2300
Tso	9239	4927	12419	12288	5686	4092	4972	1738	9153
Atm	1731	869	3474	3666	704	773	765	549	2328
Agb	29407	15079	43357	29608	14701	13123	14230	8432	24714
Ast	15290	18090	49226	31983	10518	9266	7667	5077	34553
Rac	12526	5351	35323	18469	3675	5720	5381	970	13788
Ars	1026	432	1539	1074	634	335	511	186	681
Bnr	6508	5188	16019	10958	6613	5490	5255	2253	11757
Brp	26941	15203	64714	41120	15255	18239	14595	4866	44598
Ilf	1462	330	3923	4696	415	705	479	56	2395
Drg	11654	4463	25949	37415	5254	4153	7678	2672	77069
Crj	644	283	5000	2229	163	369	236	9	542

***Table A-9**: 2012/2013 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3344	1023	2997	3629	702	696	876	412	2580
Tso	9567	5252	12288	12405	6467	4267	5521	1844	8776
Atm	1768	947	3609	3855	713	730	918	543	3280
Agb	27880	15385	40793	28897	13755	11737	14248	8679	24519
Ast	14273	17716	45115	30172	9596	8295	7234	4905	35603
Rac	11794	5809	35869	19972	3935	5237	5293	1241	16738
Ars	1015	398	1287	975	699	318	459	195	718
Bnr	7539	5665	15582	11971	6508	5416	5403	2362	13184
Brp	25902	17347	68544	45483	14877	18883	15755	5723	49599
Ilf	1538	436	3713	4444	498	802	469	65	2907
Drg	12877	6168	38159	42167	7530	5844	9157	2861	82062
Crj	695	284	4952	2427	225	360	230	28	789

***Table A-10**: 2013/2014 South African crime data as provided for by the SAPS*

	EC	FS	GT	KZ	LP	MP	NW	NC	WC
Mdr	3453	946	3333	3625	729	810	825	438	2909
Tso	9897	4814	11021	11875	6423	3953	4850	1754	8062
Atm	1858	911	3901	3866	753	772	1079	607	3363
Agb	27451	14531	41581	29040	12678	10803	13509	8734	24846
Ast	13392	17124	44748	26393	9078	7575	6783	4791	37273
Rac	13485	5358	42646	21040	5180	5284	5427	1405	19526
Ars	1107	368	1273	939	586	284	422	169	663
Bnr	7658	5194	16480	11206	7000	5316	4875	2382	13489
Brp	24750	16363	68139	44055	16503	18600	15434	6027	50589
Ilf	1843	483	3679	4586	492	939	494	94	2810
Drg	15063	8199	74713	45954	9609	7464	11015	3252	85463
Crj	775	259	6064	2274	252	365	242	29	961

Appendix B: Supplementary CA Maps

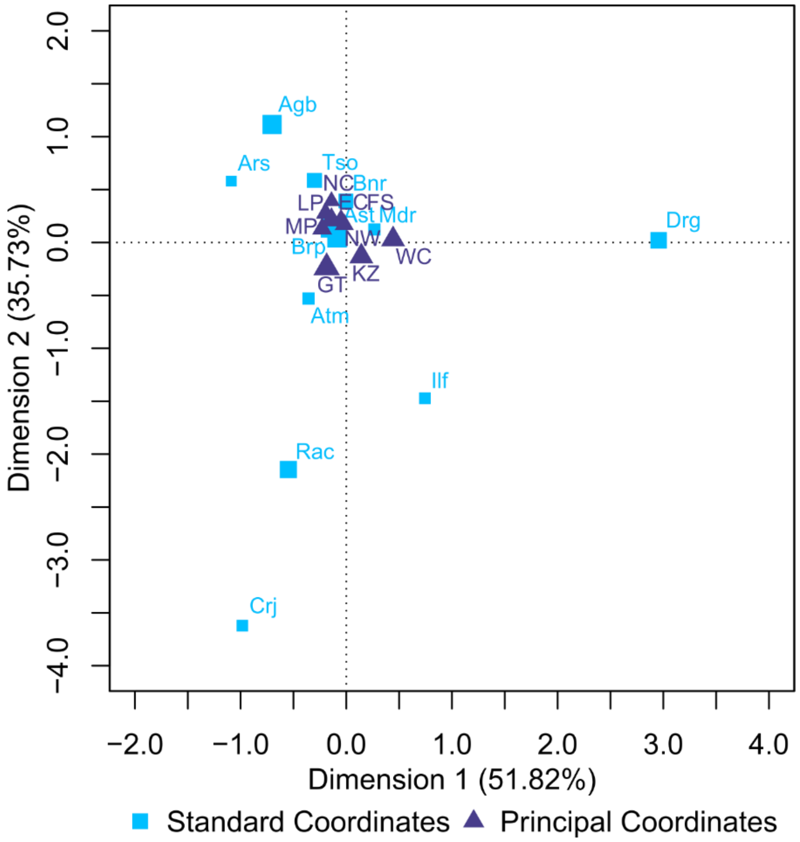

***Figure B-1**: An asymmetric CA display of the 2006 South African crime data.*

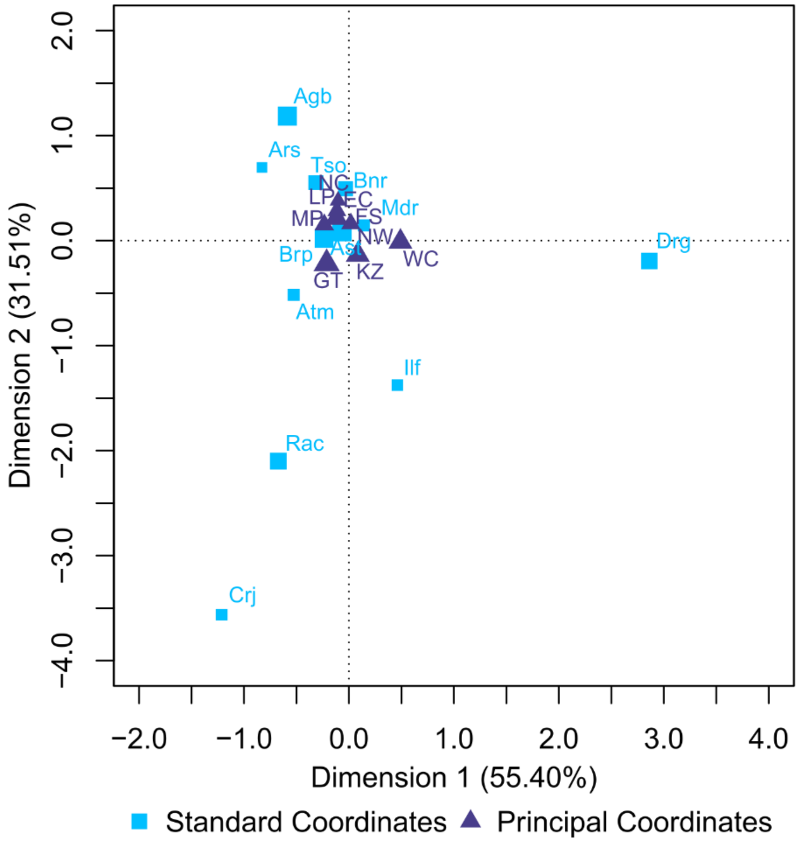

***Figure B-2**: An asymmetric CA display of the 2007 South African crime data.*

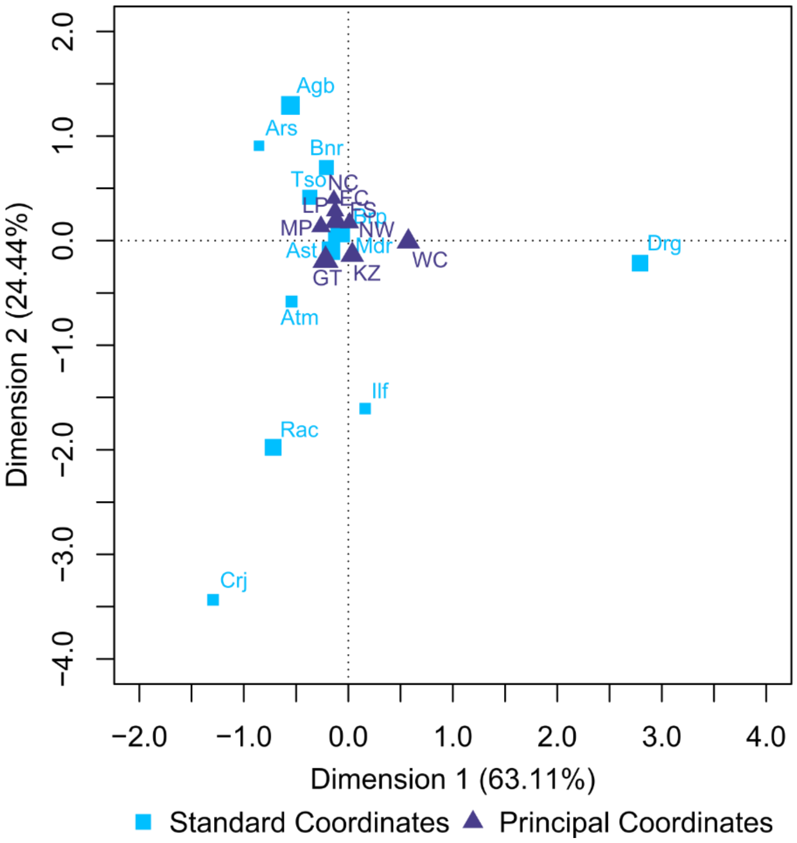

Figure B-3*: An asymmetric CA display of the 2008 South African crime data.*

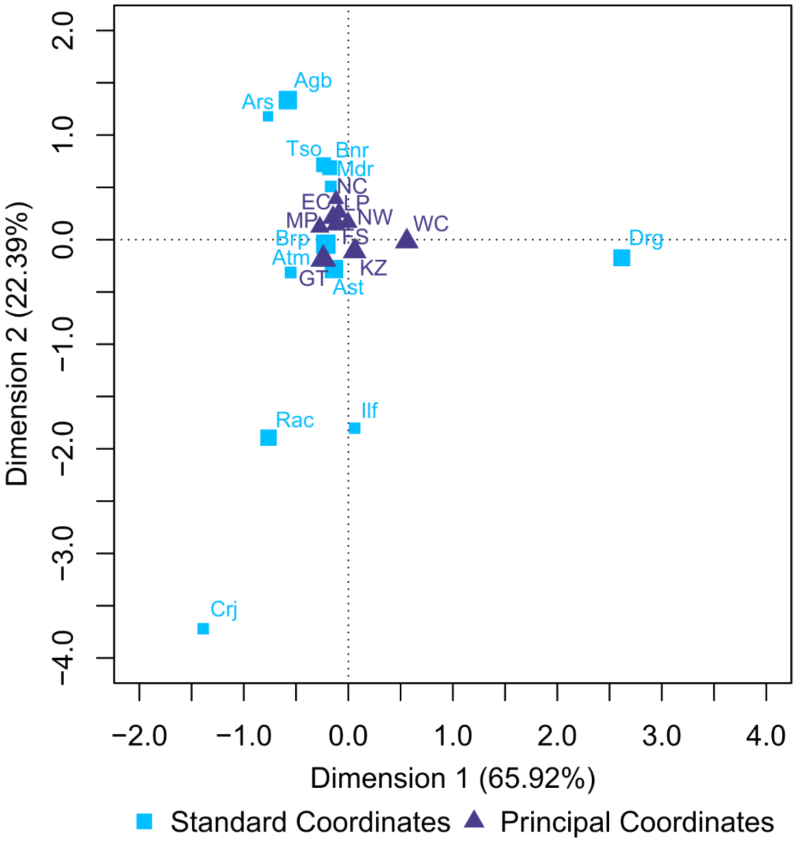

Figure B-4*: An asymmetric CA display of the 2009 South African crime data.*

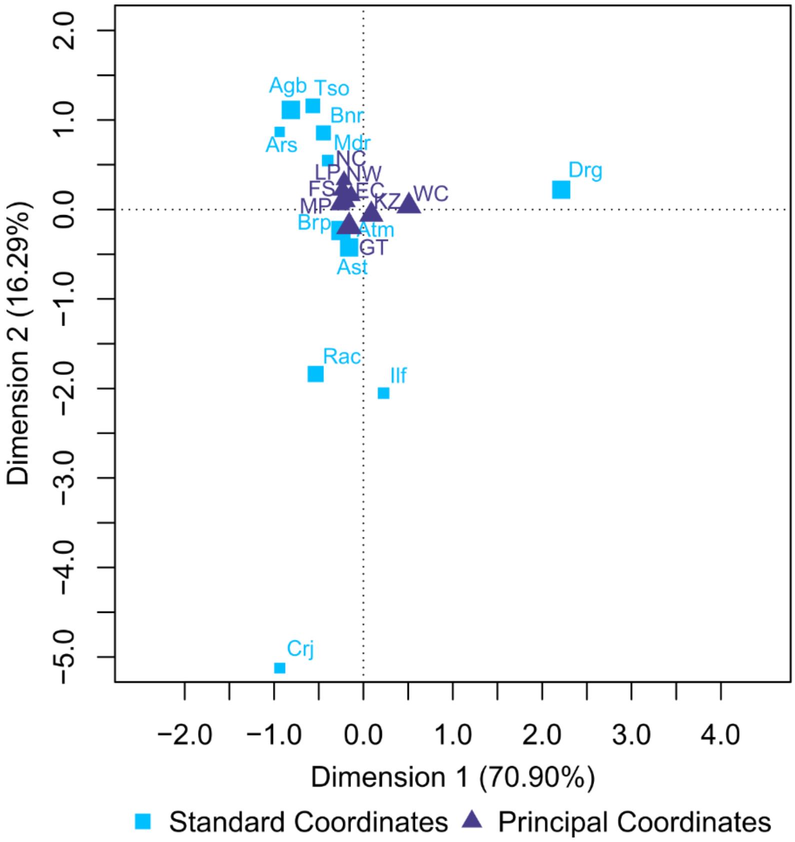

***Figure B-5**: An asymmetric CA display of the 2011 South African crime data.*

Appendix C: R Functions Utilised in Chapter 6

CA Rotated Function

Description

The CArotated function rotates and reflects the points of the two-dimensional correspondence analysis approximation from the ca function. These points are then subsequently used as the input matrix X in the plot.ca function. Both the ca and plot.ca functions form part of the ca package available on the CRAN repository (https://cran.r-project.org/web/packages/ca).

Arguments

X The output from a ca analysis, stored as an object in R.

reflect Defaults to FALSE. Possible values are “x” (reflect about X-axis) and “y” (reflect about Y-axis).

rotate.degrees Default is 0. A positive value results in an anti-clockwise rotation and a negative value results in a clockwise rotation.

Values

The function returns the X object, however both the row and column points are replaced with the new reflected and/or rotated coordinates.

R Code

```
CArotated<-function (X,reflect = c(FALSE,"x", "y"), rotate.degrees = 0)
{
rowcoords<-as.matrix(X$rowcoord)
colcoords<-as.matrix(X$colcoord)

reflect <- reflect[1]
radns <- pi * rotate.degrees/180

###REFLECTING POINTS###
if (reflect == FALSE)
    Ref.Matrix <- diag(2)
else {
    if (reflect == "x")
        Ref.Matrix <- diag(c(1, -1))
    else {
        if (reflect == "y")
          Ref.Matrix <- diag(c(-1, 1))
        else stop("Argument reflect can only set to be NULL, x or y. \n")
    }
}

###ROTATING POINTS###
R.Matrix <- matrix(c(cos(radns), -sin(radns), sin(radns),
            cos(radns)), ncol = 2)


X$rowcoord <- rowcoords[,1:2] %*% R.Matrix %*% Ref.Matrix
X$colcoord<- colcoords[,1:2] %*% R.Matrix %*% Ref.Matrix
return(X)
}
```

Attractions Plot Function

Description

The function `Attractions.Plot` is used to construct a figure of the associations between the rows and columns of a data matrix.

Arguments

`X` A data matrix of size $n \times p$, where $n > p$.

`CRIT` A cut-off value which determines which associations are used in the construction of the figure. A value greater than 0 creates a plot depicting only the attractions between the rows and columns of the data matrix. A value less than zero will include both repulsions and attractions.

`YLIM` A vector of length 2 specifying the limits of the vertical axis. Defaults to NULL.

R Code

```
function (X,CRIT=0.35,YLIM=NULL)
{
###############################
###Creating a CA of Matrix X###
###############################
CAOUT<-ca(X)
##Calculating the Row and Column Principal Coordinates
UD<-CAOUT$rowcoord%*%diag(CAOUT$sv)
VD<-CAOUT$colcoord%*%diag(CAOUT$sv)

##Standardizing the Row and Column Coordinates
Row.Proj<-as.matrix(UD[,1])
Col.Proj<-as.matrix(VD[,1])
rownames(Row.Proj)<-rownames(X)
rownames(Col.Proj)<-colnames(X)

##Order Row Coordinates according to value
Order.Row<-order(Row.Proj,decreasing=TRUE)
Ordered.Row<-as.matrix(Row.Proj[Order.Row,])

##Creating a matrix of plotting coordinates for the Rows
Row.Out<-matrix(seq(9,1),ncol=1)
rownames(Row.Out)<-rownames(Ordered.Row)
Row.Out<-cbind(1,Row.Out)

##Order Column Coordinates according to value
Order.Col<-order(Col.Proj,decreasing=TRUE)
Ordered.Col<-as.matrix(Col.Proj[Order.Col,])

##Order Column Coordinates according to value
Col.Out<-matrix(seq(12,1),ncol=1)/12*9
rownames(Col.Out)<-rownames(Ordered.Col)
Col.Out<-cbind(2,Col.Out)
##Plotting the standardized principal coordinates
plot(Row.Out,pch=19,xlim=c(0,3),ylim=YLIM,xlab="",ylab="")
text(Row.Out,labels=rownames(Row.Out),pos=2)
points(Col.Out,pch=19)
text(Col.Out,labels=rownames(Col.Out),pos=4)

###Calculating the Association Rate Matrix###
Association.Mat<-Association.mat(X)
Ordered.Mat1<-Association.Mat[rownames(Row.Out),]
Ordered.Mat2<-Ordered.Mat1[,rownames(Col.Out)]

###Drawing lines of Association###
for(i in 1:nrow(Ordered.Mat2)){

for(j in 1:ncol(Ordered.Mat2)){
if(Ordered.Mat2[i,j]>CRIT)
{segments(x0=Row.Out[i,1],y0=Row.Out[i,2],x1=Col.Out[j,1],y1=Col.Out[j,2],col="p
urple")}
}
}
return(list("Row.Out"=Row.Out,"Col.Out"=Col.Out))}
```

Association.mat Function

Description

The function `Association.mat` calculates the association matrix from data matrix ***X***.

Arguments

X	A numerical matrix or data frame.

R Code

```
function(X){

X=as.matrix(X)
Column.Sums<-apply(X,2,sum)
Row.Sums<-apply(X,1,sum)
Total.Sum<-sum(X)

Step.1<-Total.Sum*X
Step.2<-sweep(Step.1,1,Row.Sums,"/")
Step.3<-sweep(Step.2,2,Column.Sums,"/")
Final<-Step.3-1

return("Association Rate Matrix"=Final)}
```

Appendix D: R Functions Utilised in Chapter 7

Zoom Function

Description

The function `Zoom` is used to obtain the new set of X and Y limits for the new zoomed-in plotting region

Arguments

`Zoomval` Specifies zooming factor. Defaults to `NULL` for no zooming, a value less than unityspecifies zooming in, a value larger than unity specifies zooming out.

Values

The function returns a vector of length 4, containing the maximum and minimum values for the X and Y axes of the zoomed-in figure.

R Code

```
Zoom<-function (X)
{
Select.zoom.point   <-locator(1)
Current.limits      <-par("usr")
Xrange              <-Current.limits[2]-Current.limits[1]
Yrange              <-Current.limits[4]-Current.limits[3]

New.limits          <-c(Select.zoom.point$x, Select.zoom.point$x + X * Xrange,
Select.zoom.point$y, Select.zoom.point$y + X* Yrange)
return(New.limits)
}
```

Log Ratio Biplot Function

Description

The function `LogRatioBipl` is used to construct and manipulate the log-ratio biplots of Chapter 7.

Arguments

`X` A numerical matrix or data frame which provides the data for the construction of the log-ratio biplot.

`principal` A character string specifying if the rows or columns should be displayed in principal coordinates. Allowed options include `"row”` or `“column”`.

`Weight` A character string specifying between the weighted or unweighted log-ratio analysis. Allowed values include: `“TRUE”` (weighted LRA) or `“FALSE”` (unweighted LRA).

`type` A character string specifying between the asymmetric and symmetric log-ratio biplot. Allowed options include: `“asymmetric”` and `“symmetric”`.

lambda A logical argument specifying if lambda scaling should be implemented in the log-ratio biplot.

point.col A vector of length 2 specifying the colours of the row and column points.

point.pch A vector of length 2 specifying the plotting characters of the row and column points.

draw.line A logical argument specifying if link vectors are to be drawn.

line.selct If draw.line=”TRUE” then a vector of length 2 or 4 is specified to draw link vectors between the first 2 and last 2 specified variables. Links must be specified by name.

Zoomval Specifies zooming factor. Defaults to NULL for no zooming, a value less than unity specifies zooming in, a value larger than unity specifies zooming out.

atx The points at which the tick marks are to be drawn on the x-axis.

aty The points at which the tick marks are to be drawn on the y-axis.

Zatx The points at which the tick marks are to be drawn on the x-axis for the zoomed-in figure.

Zaty The points at which the tick marks are to be drawn on the y-axis for the zoomed-in figure.

reflect Defaults to FALSE. Possible values are “x” (reflect about X-axis) and “y” (reflect about Y-axis).

rotate.degrees Default is 0. A positive value results in an anti-clockwise rotation and a negative value results in a clockwise rotation.

vectors.as Defaults to FALSE. Possible value are “Lines” to draw the vectors as lines or “Arrows” to draw the vectors as arrows.

Values

A graph depicting the log-ratio biplot of the input data matrix X. Additionally, a possible second zoomed-in figure is also produce.

Total.Variance A measure of the total variance of the data matrix X

Quality The quality measure of the two dimensional approximation.

R Code

```
LogRatioBipl<-function (X, principal=c("row","column"),Weight=c("TRUE","FALSE"),
type=c("asymmetric","symmetric"),lambda=FALSE,point.col=c("deepskyblue","darksla
teblue"),
draw.line=FALSE,line.selct=NULL,zoomval=NULL,atx=NULL,aty=NULL,Zatx=NULL,Zaty=NU
LL,
reflect = c(FALSE,"x", "y"), rotate.degrees =
0,dim.biplot=2,vector.as.line=FALSE)
{
N<-as.matrix(X)
n<-sum(N)
NC<-ncol(N)
NR<-nrow(N)
principal<-principal[1]
Weight<-Weight[1]
type=type[1]
lambda=lambda[1]

    if(Weight=="TRUE"){
    R<-as.vector((1/n)*N%*%matrix(1,ncol=1,nrow=NC))
    C<-as.vector((1/n)*t(N)%*%matrix(1,ncol=1,nrow=NR))}
    else{ if(Weight=="FALSE"){
    R<-as.vector((1/NR)*matrix(1,ncol=1,nrow=NR))
    C<-as.vector((1/NC)*matrix(1,ncol=1,nrow=NC))} else{ stop(cat("Incorrect
Weighting Option\n"))}}

        P   <-  (1/n)*N
        Z   <- as.matrix(log(P))
        mc  <- t(Z) %*% as.vector(R)
        Z   <- Z - rep(1, nrow(P)) %*% t(mc)
        mr  <- Z %*% as.vector(C)
        Z   <- Z - mr %*% t(rep(1,ncol(P)))
        S   <- diag(sqrt(R)) %*% Z %*% diag(sqrt(C))
        svdS<- svd(S)

        eigensqrd<-svdS$d^2
        total.variance<-sum(S*S)
        TwoD.Appr<-sum(eigensqrd[1:2])/sum(eigensqrd)

#Principal coordinates of the rows (Crimes)
F<- diag(1/sqrt(R)) %*% svdS$u[,1:2] %*% diag(svdS$d[1:2])
rownames(F)<-rownames(X)
#standard coordinates of the columns (provinces)
T <- diag(1/sqrt(C)) %*% svdS$v[,1:2]
rownames(T)<-colnames(X)

#Principal coordinates of the columns (provinces)
G<-diag(1/sqrt(C)) %*% svdS$v[,1:2] %*% diag(svdS$d[1:2])
rownames(G)<-colnames(X)
#standard coordinates of the rows (Crimes)
Phi<-diag(1/sqrt(R)) %*% svdS$u[,1:2]
rownames(Phi)<-rownames(X)



windows()
par(pty="s")

###CREATING AN ASYMMETRIC ROW-PRINCIPAL PLOT###
if(type=="asymmetric"){
    if(principal=="row"){

        if (lambda) {
          lam.4 <- NR * sum(T * T)/(NC * sum(F *
                F))
          lam <- sqrt(sqrt(lam.4))
          F <- F * lam
          T<- T/lam}
####################ROTATING POINTS#####################
reflect <- reflect[1]
radns <- pi * rotate.degrees/180

if (reflect == FALSE)
    Ref.Matrix <- diag(2)
else {
    if (reflect == "x")
        Ref.Matrix <- diag(c(1, -1))
    else {
        if (reflect == "y")
          Ref.Matrix <- diag(c(-1, 1))
        else stop("Argument reflect can only set to be NULL, x or y. \n")
    }
}

R.Matrix <- matrix(c(cos(radns), -sin(radns), sin(radns),
            cos(radns)), ncol = 2)

    F.Rot<-F %*% R.Matrix %*% Ref.Matrix
    T.Rot<-T %*% R.Matrix %*% Ref.Matrix
#   return(T.Rot)
#######################################################

        Max<-max(rbind(F.Rot,T.Rot))
        Min<-min(rbind(F.Rot,T.Rot))

        plot(rbind(F.Rot,T.Rot), asp = 1, type = "n", xlab = "", ylab = "",
        xaxt = "n", yaxt = "n", cex.axis = 0.7,ylim=c(Min,Max),
        xlim=c(Min,Max),main="Row Principal Log-Ratio Biplot")

        abline(v=0,h=0,lty=2,col="gray87")
        axis(1,at=atx,cex.axis = 0.7)
        axis(2,at=aty,cex.axis = 0.7)
        axis(3,at=atx,cex.axis = 0.7)
        axis(4,at=aty,cex.axis = 0.7)

                        if(vector.as.line=="Arrows"){
                        for(i in 1:9){

arrows(x0=0,y0=0,x1=T.Rot[i,1],y1=T.Rot[i,2],col="red",lwd=1.5,length=0.15)}}

                        if(vector.as.line=="Lines"){
                        for(i in 1:9){

segments(x0=0,y0=0,x1=T.Rot[i,1],y1=T.Rot[i,2],col="red",lwd=1.5)}}

        points(F.Rot,pch=15,col=point.col[1])
        points(T.Rot,pch=17,col=point.col[2])

        text(F.Rot, labels = rownames(F), col = point.col[1], font = 1,
        cex = 0.7,pos=3)
        text(T.Rot, labels = rownames(T), col = point.col[2], font = 1,
        cex = 0.7,pos=3)

        if(draw.line==TRUE){
        if(is.null(line.selct)){ stop("If draw.line=TRUE, a pair of standard
        coordinate points must be selected by column names\n")}
        else{
        L1<-T.Rot[line.selct[1],]
        L2<-T.Rot[line.selct[2],]
        L<-rbind(L1,L2)
        points(L,type="l",lty=3)
                                    if(length(line.selct)>2){
                                    L3<-T.Rot[line.selct[3],]
                                    L4<-T.Rot[line.selct[4],]
                                    LL<-rbind(L3,L4)
                                    points(LL,type="l",lty=3)}
        }}

        if(!is.null(zoomval)){
        zoomval<-Zoom(zoomval)
        windows()
        par(pty="s")
        plot(rbind(F.Rot,T.Rot), asp = 1, type = "n", xlab = "", ylab = "",
        xaxt = "n", yaxt = "n", cex.axis = 0.7,ylim=zoomval[3:4],
        xlim=zoomval[1:2],main="Row Principal Log-Ratio Biplot Zoom")

        abline(v=0,h=0,lty=2,col="gray87")
        axis(1,at=Zatx,cex.axis = 0.7)
        axis(2,at=Zaty,cex.axis = 0.7)
        axis(3,at=Zatx,cex.axis = 0.7)
        axis(4,Z=Zaty,cex.axis = 0.7)

                                if(vector.as.line=="Arrows"){
                                for(i in 1:9){

arrows(x0=0,y0=0,x1=T.Rot[i,1],y1=T.Rot[i,2],col="red",lwd=1.5,length=0.15)}}

                                if(vector.as.line=="Lines"){
                                for(i in 1:9){

segments(x0=0,y0=0,x1=T.Rot[i,1],y1=T.Rot[i,2],col="red",lwd=1.5)}}

        points(F.Rot,pch=15,col=point.col[1])
        points(T.Rot,pch=17,col=point.col[2])

        text(F.Rot, labels = rownames(F), col = point.col[1], font = 1,
        cex = 0.7,pos=3)
        text(T.Rot, labels = rownames(T), col = point.col[2], font = 1,
        cex = 0.7,pos=3)

        if(draw.line==TRUE){
        windows()
        par(pty="s")
        if(is.null(line.selct)){ stop("If draw.line=TRUE, a pair of principal
coordinate
        points must be selected by variable name\n")}else{
        L1<-T.Rot[line.selct[1],]
        L2<-T.Rot[line.selct[2],]
        L<-rbind(L1,L2)
        points(L,type="l",lty=3)
                                    if(length(line.selct)>2){
                                    L3<-T.Rot[line.selct[3],]
                                    L4<-T.Rot[line.selct[4],]
                                    LL<-rbind(L3,L4)
                                    points(LL,type="l",lty=3)}
        }}}
}
                ###CREATING AN ASYMMETRIC COLUMN-PRINCIPAL PLOT###
                else {if(principal=="column"){

                                if (lambda) {
                                  lam.4 <- NR * sum(Phi * Phi)/(NC * sum(G *
                                        G))
                                  lam <- sqrt(sqrt(lam.4))
```

```
                          G <- G * lam
                          Phi<- Phi/lam}

#############################ROTATING
POINTS############################################################
                            reflect <- reflect[1]
                            radns <- pi * rotate.degrees/180

                            if (reflect == FALSE)
                                Ref.Matrix <- diag(2)
                            else {
                                if (reflect == "x")
                                    Ref.Matrix <- diag(c(1, -1))
                                else {
                                    if (reflect == "y")
                                    Ref.Matrix <- diag(c(-1, 1))
                                    else stop("Argument reflect can only set to
be NULL, x or y. \n")
                                }
                            }

                            R.Matrix <- matrix(c(cos(radns), -sin(radns),
sin(radns),
                                        cos(radns)), ncol = 2)


                            G.Rot<-G %*% R.Matrix %*% Ref.Matrix
                            Phi.Rot<-Phi %*% R.Matrix %*% Ref.Matrix
##################################################################################
#######################
                        Max<-max(rbind(G.Rot,Phi.Rot))
                        Min<-min(rbind(G.Rot,Phi.Rot))

                        plot(rbind(G.Rot,Phi.Rot), asp = 1, type = "n", xlab =
"", ylab = "",
                        xaxt = "n", yaxt = "n", cex.axis = 0.7,ylim=c(Min,Max),
                        xlim=c(Min,Max),main="Column Principal Log-Ratio
Biplot")

                        abline(v=0,h=0,lty=2,col="gray87")
                        axis(1,at=atx,cex.axis = 0.7)
                        axis(2,at=aty,cex.axis = 0.7)
                        axis(3,at=atx,cex.axis = 0.7)
                        axis(4,at=aty,cex.axis = 0.7)

                        if(vector.as.line=="Arrows"){
                        for(i in 1:12){

arrows(x0=0,y0=0,x1=Phi.Rot[i,1],y1=Phi.Rot[i,2],col="red",lwd=1.5,length=0.15)}
}

                        if(vector.as.line=="Lines"){
                        for(i in 1:12){

segments(x0=0,y0=0,x1=Phi.Rot[i,1],y1=Phi.Rot[i,2],col="red",lwd=1.5)}}

                        points(G.Rot,pch=15,col=point.col[1])
                        points(Phi.Rot,pch=17,col=point.col[2])

                        text(G.Rot, labels = rownames(G), col = point.col[1],
font = 1,
                        cex = 0.7,pos=3)
                        text(Phi.Rot, labels = rownames(Phi), col =
point.col[2], font = 1,
                        cex = 0.7,pos=3)

                        if(draw.line==TRUE){
                        if(is.null(line.selct)){ stop("If draw.line=TRUE, a pair
of principal coordinate
                        points must be selected by variable name\n")}else{
                        L1<-Phi.Rot[line.selct[1],]
                        L2<-Phi.Rot[line.selct[2],]
                        L<-rbind(L1,L2)
                        points(L,type="l",lty=3)
                                        if(length(line.selct)>2){
                                        L3<-Phi.Rot[line.selct[3],]
                                        L4<-Phi.Rot[line.selct[4],]
                                        LL<-rbind(L3,L4)
                                        points(LL,type="l",lty=3)}
                        }}

                            if(!is.null(zoomval)){
                            zoomval<-Zoom(zoomval)
                            windows()
                            par(pty="s")


                            plot(rbind(G.Rot,Phi.Rot), asp = 1, type = "n", xlab
= "", ylab = "",
                            xaxt = "n", yaxt = "n", cex.axis =
0.7,ylim=zoomval[3:4],
                            xlim=zoomval[1:2],main="Column Principal Log-Ratio
Biplot Zoom")

                            abline(v=0,h=0,lty=2,col="gray87")
                            axis(1,at=Zatx,cex.axis = 0.7)
                            axis(2,at=Zaty,cex.axis = 0.7)
                            axis(3,at=Zatx,cex.axis = 0.7)
                            axis(4,Z=Zaty,cex.axis = 0.7)

                            points(G.Rot,pch=15,col=point.col[1])
                            points(Phi.Rot,pch=17,col=point.col[2])

                            text(G.Rot, labels = rownames(G), col =
point.col[1], font = 1,
                            cex = 0.7,pos=3)
                            text(Phi.Rot, labels = rownames(Phi), col =
point.col[2], font = 1,
                            cex = 0.7,pos=3)

                        if(draw.line==TRUE){
                        if(is.null(line.selct)){ stop("If draw.line=TRUE, a pair
of principal coordinate
                        points must be selected by variable name\n")}else{
                        L1<-Phi.Rot[line.selct[1],]
                        L2<-Phi.Rot[line.selct[2],]
                        L<-rbind(L1,L2)
                        points(L,type="l",lty=3)
                                        if(length(line.selct)>2){
                                        L3<-Phi.Rot[line.selct[3],]
                                        L4<-Phi.Rot[line.selct[4],]
                                        LL<-rbind(L3,L4)
                                        points(LL,type="l",lty=3)}
                            }}}
                }
                else{stop(cat("Incorrect Principal Point Selection\n"))}}}

###CREATING A SYMMETRIC PLOT###
```

```
     else{if(type=="symmetric"){
    Max<-max(rbind(F,G))
    Min<-min(rbind(F,G))
     plot(rbind(F,G), asp = 1, type = "n", xlab = "", ylab = "",
     xaxt = "n", yaxt = "n", cex.axis =
0.7,xlim=c(Min,Max),ylim=c(Min,Max),main="Symmetric Log-Ratio Biplot")
        abline(v=0,h=0,lty=2,col="gray87")
        axis(1,at=atx,cex.axis = 0.7)
        axis(2,at=aty,cex.axis = 0.7)
        axis(3,at=atx,cex.axis = 0.7)
        axis(4,at=aty,cex.axis = 0.7)



        points(F,pch=15,col=point.col[1])
        points(G,pch=17,col=point.col[2])

        text(F, labels = rownames(F), col = point.col[1], font = 1,
        cex = 0.7,pos=3)
        text(G, labels = rownames(G), col = point.col[2], font = 1,
        cex = 0.7,pos=3)

            if(!is.null(zoomval)){
            zoomval<-Zoom(zoomval)
            windows()
            par(pty="s")
            plot(rbind(F,G), asp = 1, type = "n", xlab = "", ylab = "",
            xaxt = "n", yaxt = "n", cex.axis = 0.7,ylim=zoomval[3:4],
            xlim=zoomval[1:2],main="Symmetric Log-Ratio Biplot Zoom")

            abline(v=0,h=0,lty=2,col="gray87")
            axis(1,at=Zatx,cex.axis = 0.7)
            axis(2,at=Zaty,cex.axis = 0.7)
            axis(3,at=Zatx,cex.axis = 0.7)
            axis(4,at=Zaty,cex.axis = 0.7)

            points(F,pch=15,col=point.col[1])
            points(G,pch=17,col=point.col[2])

            text(F, labels = rownames(F), col = point.col[1], font = 1,
            cex = 0.7,pos=3)
            text(G, labels = rownames(G), col = point.col[2], font = 1,
            cex = 0.7,pos=3)

        }}
    }
return(list("Total.Variance"=total.variance,"Quality"=TwoD.Appr))
}
```

References

Aitchison, J. 1986. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. 44(2):139–177.

Aitchison, J. & Greenacre, M. 2002. Biplots of Compositional Data. Journal of the Royal Statistical Society. 51(4):375–392.

Baumer, B. 2015. A Data Science Course for Undergraduates: Thinking with Data. The American Statistician. 69(4):334–342.

Box, G.E.P. & Cox, D.R. 1964. An analysis of transformations. Journal of the Royal Statistical Society. 26(2):211–252.

Chen, C., Hardle, W. & Unwin, A. 2008. Handbook of Data Visualization. Berlin: Springer.

Cox, T.F. & Cox, M.A.A. 2001. Multidimensional Scaling, Second Edition. Boca Raton: Chapman & Hall/CRC.

CSVR. 2009. Why does South Africa have such high rates of violent crime? [Online], Available: http://www.csvr.org.za/archive/docs/study/7.unique_about_SA.pdf [2016, March 12].

Explaining the official crime statistics for 2013 /14. 2014. [Online], Available: https://www.issafrica.org/uploads/ISS-crime-statistics-factsheet-2013-2014.pdf [2015, January 11].

Fajnzylber, P., Lederman, D. & Loayza, N. 2002. What causes violent crime? European Economic Review. 46(7):1323–1357.

Friendly, M. & Meyer, D. 2016. Discrete data analysis with R : visualization and modeling techniques for categorical and count data. Boca Raton: CRC Press.

Gower, J.C. & Hand, D.J. 1996. Biplots. London: Chapman & Hall.

Gower, J., Lubbe, S. & Le Roux, N. 2011. Understanding Biplots. Chichester, West Sussex, UK: John Wiley.

Greenacre, M. 2007. Correspondence Analysis in Practice. Boca Raton: Chapman & Hall/CRC.

Greenacre, M. 2009. Power transformations in correspondence analysis. Computational Statistics and Data Analysis. 53(8):3107–3116.

Greenacre, M. 2010. Biplots in Practice. Spain: Fundación BBVA.

Greenacre, M. 2011. Measuring Subcompositional Incoherence. Mathematical Geosciences. 43(6):681–693.

Greenacre, M. & Primicerio, R. 2013. Multivariate Analysis of Ecological Data. Bilbao: Fundación BBVA.

Hartigan, J.A. & Kleiner, B. 1981. Mosaics for Contingency Tables. In Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. New York: Springer. 268–273.

Horning, A.M., Salfati, C.G. & Labuschagne, G.N. 2015. South African Serial Homicide: A Victim-Focused Behavioural Typology. Journal of Investigative Psychology and Offender Profiling. 12(1):44–68.

Is South Africa's murder rate in the top three globally? 2015. [Online], Available: https://africacheck.org/spot_check/is-sas-murder-rate-in-the-top-three-globally/ [2016, March 12].

Kroonenberg, P. 2008. Applied multiway data analysis. Hoboken NJ: Wiley.

Le Roux, B. & Rouanet, H. 2004. Geometric data analysis: from correspondence analysis to structured data analysis. Dordrecht: Kluwer Academic Publishers.

Le Roux, N. & Lubbe, S. 2013. UBbipl: Understanding Biplots: Data Sets and Functions. R package version 3.0.4.

Leman, S., House, L. & Hoegh, A. 2015. Developing a New Interdisciplinary Computational Analytics Undergraduate Program: A Qualitative-Quantitative-Qualitative Approach. The American Statistician. 69(4):397–408.

Mccall, P.L. & Nieuwbeerta, P. 2007. Structural Covariates of Homicide Rates. Homicide Studies. 11(3):167–188.

Messner, S.F. 1989. Economic Discrimination and Societal Homicide Rates: Further Evidence on the Cost of Inequality. American Sociological Review. 54(4):597–611.

Meyer, D., Zeileis, A. & Hornik, K. 2015. VCD: Visualizing Categorical Data. R package version 1.4-1.

Nolan, D. & Temple Lang, D. 2015. Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Analysis. The American Statistician. 69(4):292–299.

Rencher, A.C. 2002. Methods of multivariate analysis. J. Wiley.

Republic of South Africa. 1992. Drugs And Drug Trafficking Act No. 140 of 1992. Government Gazette. 325(14143):1–72.

Republic of South Africa. 2001. Firearms Control Act No. 60 , 2000. Government Gazette. 460(22214):1–58.

Republic of South Africa. 2007. Criminal Law (Sexual Offences and Related Matters) Amendment Act No. 32 of 2007. Government Gazette. 510(30599):1–74.

Salfati, C.G. & Labuschagne, G.N. 2015a. Serial Homicide in South Africa Introduction to the Special Issue. Journal of Investigative Psychology and Offender Profiling. 12(1):1–3.

Salfati, C.G. & Labuschagne, G.N. 2015b. An examination of serial homicide in South Africa: The practice to research link. Journal of Investigative Psychology and Offender Profiling. 12(1):4–17.

Salfati, C.G., Labuschagne, G.N., Horning, A.M., Sorochinski, M. & De Wet, J. 2015. South African Serial Homicide: Offender and Victim Demographics and Crime Scene Actions. Journal of Investigative Psychology and Offender Profiling. 12(1):18–43.

Salfati, C.G., Horning, A.M., Sorochinski, M. & Labuschagne, G.N. 2015. South African Serial Homicide: Consistency in Victim Types and Crime Scene Actions Across Series. Journal of Investigative Psychology and Offender Profiling. 12(1):83–106.

SAPS. 2014. An Analysis of the National Crime Statistics. [Online], Available: www.saps.gov.za [2016, January 11].

Sorochinski, M., Salfati, C.G. & Labuschagne, G.N. 2015. Classification of Planning and Violent Behaviours in Serial Homicide: A Cross-National Comparison Between South Africa and the US. Journal of Investigative Psychology and Offender Profiling. 12(1):69–82.

Thomson, J.D.S. 2004. Coloured homicide trends in South Africa. South African Crime Quarterly. (7):9–14.

UNODC. 2013. UNODC: Global Study on Homicide. [Online], Available: http://www.unodc.org/gsh/ [2015, January 31].

Violent crime damaging SA economy – govt. 2013. [Online], Available: http://www.citypress.co.za/business/violent-crime-damaging-sa-economy-govt/ [2015, February 03].