

University of KwaZulu-Natal.

Statistical Modelling and Spatial Mapping of Crime in South Africa.



UNIVERSITY OF
KWAZULU-NATAL

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

By

BELISHA NAIDOO

Submitted in fulfilment of the academic requirements for the degree of Master of Philosophy in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban

December 06, 2016

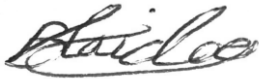
Thesis advisor: **Professor D. North**
Professor T. Zewotir

Candidate's signature:


A handwritten signature in black ink, appearing to read 'Belisha Naidoo', written over a horizontal line.

Disclaimer

This document describes work undertaken as part of a master's programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Candidate's signature: 

As the candidate's supervisor I have /have not approved this thesis/dissertation for submission.

Signed: 
Name: Prof. D. NORTH
Date: 7/12/2016

Abstract

This research investigates factors related to crime rates for the 2013/2014 South African Crime Survey. The survey provides personal information and crime related experiences for all members of the 25 605 households that was part of the study. Using the generalized linear model analysis we show that the crime outcomes significantly differed between provinces. A further data set, containing aggregated crime statistics from 1 140 police stations, had the GPS co-ordinates included which allowed for spatial mapping of crime incidence. Results may be used to predict crime hot spots in the country, thereby having the potential to inform crime reduction initiatives, which could be deployed strategically in order to minimize overall crime by focusing on the potential crime hot spots. In a country where resources are limited and that careful planning is essential, this study potentially has a lot to offer.

Table of Contents

Abstract	3
Acknowledgements	5
1. Introduction	6
2. Literature Review	9
3. Crime Data	12
4. Descriptive Statistics	14
5. Logistic Regression	26
6. Generalized Estimating Equations	45
7. Spatial Analysis	52
8. Conclusion	66
Bibliography	69
List of Figures	71
List of Tables	72
List of Equations	73

Acknowledgements

I would like to thank the South African Statistical Association (SASA) for the granting of a National Research Foundation- crisis in academic statistics bursary; their contribution is a vital part of this thesis.

I would like to thank my supervisor Professor Delia North and my co-supervisor Professor Temesgen Zewotir for all their help, support and guidance.

I appreciate the training in mapping tools from STATS SA, Dr R Naidoo and his team who were influential in the visual analysis of this research.

Final thanks be to God and my family.

1. Introduction

Crime is regarded as an act of breaking the law and is punishable by the state. In a broad sense we have two types of crime; serious crime such as murder, rape, and robbery and statutory crime, such as fraud, drug and alcohol abuse violations and vandalism (Wikipedia, 2015).

It is well documented that crime poses a problem in our country. According to a recent article, crime in South Africa has increased in some areas, though some crime rates have decreased over the past decade (Shaw & Kriegler, 2016). Shaw explains how the murder crime rate (being highest in Cape Town) has been increasing for the last three years after having decreased between 1994 and 2012, while aggravated robbery has also increased in the last decade.

This increase of crime in certain areas raises concerns as emphasized by the National Police Commissioner Riah Phiyega (South Africa's crime stats, 2014). It is evident from this report that the incidence of serious crime has not stabilized in the country. Reports of murder, attempted murder and sexual offences decreased between 2004 and 2013, but serious offences increased significantly between 2013 and 2014.

This recent increase of serious crimes in the country poses a problem for South Africans. Statutory crime has also increased in the 2013/2014 financial year, in particular, property related crime and drug related crime, have occurred with higher incidence.

All South Africans are affected by crime in one way or the other, either by being a victim of crime, or by living in fear of being a victim of crime. Most South African emigrants explained that the high crime in the country was a major factor influencing their decision to emigrate, thus causing a loss in man power for the country (Macdonald, 2008).

The decision to leave South Africa, is thus often due to the high crime rate of South Africa, in comparison to other countries. In particular, as evidence of the high rate of serious crimes in the country, a report on crime in South Africa (Nation Master, 2014), stated that the total number of recorded crimes committed in 2002 was around 2.6 million, i.e. the fifth highest crime rate amongst all countries at that time!

South Africa ranked ninth on the United Nation's top 10 list of world murder rates in 2012, with a murder rate of 31, calculated as the number of murders committed in one year per 100 000 people (Roane, 2014). South Africa has recorded the highest rape rate in the world since 2004, was ranked third for murders amongst

Christian countries and the country had the highest assault rate in 2011, amongst all emerging market countries!

In order to attempt to establish reasons for the high crime rate in South Africa, one needs to take cognisance of the historical background of the country, and in particular, the influence of the Apartheid laws that were implemented from 1948 to 1994, leaving a devastating legacy of unequal access to quality of life, with dire consequences for a large proportion of the citizens of the country, even today.

Williamson (1957) mentions that the Apartheid policy contributed to the increase of crime in the country over the last few decades. He explains how these laws, that forced segregation amongst different races, caused many South Africans to resort to committing crime.

Williamson argues that discrimination in South Africa led to the Blacks or “Bantus” being poorly educated and prepared for a life as servants and labourers. He further mentions that the failure to retain high levels of education amongst the Bantu society, ignited delinquent behaviour among the Blacks in the country. Discrimination leads to poverty, according to the author, with the Blacks historically only earning a fraction of the wages of the non-Blacks. Poverty is a result of unemployment and migration; which in turn leads to increased potential to commit crime, hence not surprising, this is very prevalent among the Black section of the population of the country.

David Bruce, a representative of the Centre for the Study of Violence and Reconciliation (CSV), highlighted the causes of crime in an article. He notes that the economic structure of South Africa consists of high levels of poverty and unemployment, thus causing ideal conditions for crime to be committed. The Safety and Security Minister, Charles Nqakula added in the same article that there was an increase of crime committed by children, with 3000 South African children being detained in 2008. The reason for the delinquent behaviour from children was blamed on the lack of parenting skills, supported by the CSV’s preliminary report (IOL news, 2008).

Gould (2014), blames the lack of respect that South Africans have for the law to be the reason for the high crime rate in the country. The writer believes that when South Africa entered democracy in 1994, immunity was still not gained by those who were victims of the Apartheid laws, which resulted in their disregard of the law.

A National Development Plan is currently being implemented in South Africa, which amongst other aims, hope to contribute to increased safety of all citizens by co-ordinating the work of the South African Police service, which manages 1140 police stations across the country. Currently there is only one police officer for every 346 South Africans (South Africa's crime stats, 2014).

In this study, our main objective will be to look for patterns and predictors of crime, in an attempt to add value to the process of minimizing the crime rate of the country, by better understanding the situation, so that results obtained can inform prevention strategies. We will use a statistical approach, using the crime data to develop a statistical model, which we can then use to make inferences regarding crime in South Africa.

We will further investigate South Africans' perceptions of crime occurring in their neighbourhood and to match that up with the police reported incidents of crime in their area for the 2013/2014 period. This relationship, along with other factors related to crime, will be graphically represented with the aid of graphs.

Statistical modelling and spatial mapping are the methods which will be employed to investigate the nature of crime committed and to identify the factors affecting the different types of crime.

To conclude, we will attempt to locate potential crime hotspots and thereby inform a more optimal use of crime reduction resources, which are always under constraint in a developing country, where there are so many competing urgencies.

2. Literature Review

In this chapter, we aim to discuss the problem of crime in more detail, by considering the research on this topic from authors around the world. We next shift our focus to Crime in South Africa, and apply the methodology of statistical modelling and spatial mapping to unpack the incidence and perceptions of crime in the country.

A few studies are cited below;

The Canadian Crime Statistics report (Brennan, Shannon; Dauvergne, Mia, 2010) used descriptive statistics to present data collected from the annual Uniform Crime Reporting (UCR) survey. The authors computed two measurements; crime rate i.e. the total number of crimes committed divided by the population and the crime severity index i.e. the total weighted crime divided by the population, where more serious crimes were assigned higher weights. Different categories of crime were investigated by the authors, who then depicted the results in graphs for the different provinces of Canada.

The authors found that the Northern part of the country had the highest crime rate and further had a high index of violent crime severity, while the Northwest Territories and Nunavut province, had the highest police-reported crime rate in the categories of homicide, breaking in and entry, motor vehicle theft and drug related crime. Their study found that crime was mainly committed by youth and young adults, as the crime rate was the highest among accused at the age of 18 years.

Their final conclusion was that crime rate in Canada decreased by 5% from the previous year and the crime severity index decreased by 6% in 2010. We will perform a similar descriptive analysis for the different provinces in South Africa.

Frank *et al.*, (2012) conducted a longitudinal study, focusing only on burglaries in Vancouver. Their data came from the Police Information Recording System (PIRS), which recorded 23 659 burglaries in the Metro area. Single-family dwellings were investigated by the authors over a 5-year period and the frequency of crime for each specific dwelling was recorded and consequently analysed. The main findings were that the more frequently a house was broken into, the lower the probability of it being reported to the police.

Their study was aimed at revealing the under-reporting of crime to the police. Prior to the study, only 19.8% of the burglaries for a home being broken into more than once, was reported to the police, but after this study, 47.1% of the burglaries reported to the police, were repeat burglaries. We use the reported crimes in the latter part of our research to study the reported crime per police station around the country in this thesis.

The Canadian Crime Statistics report of 1997 (Kong, 1997), has associations of the characteristics of the victims linked to the accused. The author found that in Canada, males between the ages of 26-32 were most commonly the victims of serious crime, i.e. murder, attempted murder and assault. On the other hand in the case of sexual offences, the victims were most commonly females between the ages of 12 and 17, while reported abductions were most common amongst children around the age of 7, with harassments and hostage victims being most commonly reported in the case of females between the ages of 25-31.

Considering the perpetrators of crime, it was found that for all categories of crime, aside from prostitution, crimes were more commonly committed by men, while abduction crime reported a high percentage of perpetrators, with 42% being females accused of this crime. The median ages for the offenders was between the ages of 23 and 35. We, having no information on the perpetrators for our study, will extensively investigate the characteristics of the victims of crime for the South African data of this thesis.

In South Africa, a Victims of Crime Survey (VOCS) was conducted by Statistics South Africa (Stats SA), from April 2013 to March 2014 (Victims of Crime Survey, 2014). This survey provided information on all types of crimes in South Africa. The main findings are that it is perceived (by 70% of those surveyed) that corruption increased during the period 2010-2013, and a high percentage of households surveyed (76.9%), felt that the reason for this, is that those accused wanted to get rich quickly. For vehicle theft, it was reported that 72% of the households had their vehicles stolen from their own property.

Data from the survey on assaults and sexual offences, revealed that a significant number of the victims, were victimised by their own relatives. Demographic information further revealed that residents from the province of Limpopo, felt the most safe when walking in their neighbourhood at night, while residents from the Free State felt the least safe. It is interesting to note that this study focused on the views of the study group about crime (their perceptions), as well as actual crime incidents experienced by them, as opposed to studies that use only reported crime incidents, making this a very interesting data set to explore.

We use the data provided in this survey for the first part of our research (descriptive analysis and Logistic regression), while we take an alternate approach in investigating the perception against actual incidents, where we drill deeper into prediction analysis, using this data.

Chainey et al., conducted a crime study in 2008, that used spatial analysis to explore incidence of crime. The authors found that hotspot mapping techniques best predicted the location of the occurrence of “street crime”. Spatial patterns were relatively successfully predicted only when sufficient amount of input data was used, along with the correct parameter selection. Consequently, spatial analysis through hotspot mapping was the optimal predictive crime mapping technique. This study will aim to take some of those ideas further in Chapter 7, for the South African crime data, based on the location of each police station where the crime was reported.

Spatial intensity of crime and the indicators of crime levels in the neighbourhood of Omaha, Nebraska, was investigated by Zhang et. al (2007). The authors found that the crime density indicator was more appropriate than the location quotient indicator, as it locates crime incidents, as opposed to locating where the victims of crime are. They studied four types of crime, i.e. assault, robbery, auto-theft and burglary. They applied Ordinary Least Square (OLS) method (SPSS) and revealed that high correlations existed between the demographic and household characteristics variables of crimes, in particular, they found that the greater the percentage of the minority population (i.e. the more severe the poverty, the higher the unemployment rate) the more likely the occurrence of the four types of crime (assault, robbery, auto-theft and burglary). On the other hand, the lower the median household age, the greater the probability of the occurrence of assault, robbery, auto-theft and burglary. This was due to the absence of home ownership and the lack of residence stability. The group further found that assault was associated with poverty, robbery was generally associated with the percentage of the minority population and property crime was associated with the type of property (commercial or multi-family dwellings). Low adjusted R-squared values for several models supported the authors findings that the crime density indicator is a suitable one. In conclusion, it was found by the authors that poverty and racial barriers were the greatest contributors to the occurrence of crime. We investigate the relationship of demographic factors on five categories of crime for the South African data in this study.

This chapter gave an overview of similar studies to the different aspects to be undertaken in this study. We will follow closely, in this studies, the techniques used by the mentioned authors.

3. Crime Data

The crime data used for this study was obtained from the Victims of Crime Survey (VOCS), conducted by Statistics South Africa (Stats SA), the National Statistics Office. The VOCS was designed to study the perceived views of citizens on crime in the country, as well as providing a data source for monitoring of crime rates in the country. The VOCS data is thus a valuable source, providing quantitative and qualitative information on crime levels and perceived crime levels in the country (Victims of Crime Survey, 2014).

Stats SA has conducted this survey annually since 2011, initially questioning households on crime occurring from January to December of the previous year. From 2013 onwards however, the reference frame changed and data collection methods became continuous, i.e. all year around, with surveyed candidates reflecting on the period ending a month before the interview. It is for that reason that the reference period for the VOCS 2013/2014 survey extends from April 2013 to March 2014 (Nesstar metadata, 2014).

The data set for this study comprised of 25 605 households. The sample was selected by first stratifying the Master sample collected during the 2001 census, at provincial level, by metropolitan geographic area type, then secondly, stratifying by the variables of household, i.e. size, education, occupancy status, gender, industry and income. A Probability Proportional to Size (PPS) sampling scheme was used to systematically draw a sample from each stratum.

The questionnaire for the VOCS 2013/2014 (Nesstar questionnaire, 2014) was conducted according to international standards. The survey was aimed to collect information from private households in South Africa, where a household was considered as one sample unit. The questionnaire was divided into 29 sections, i.e. where sections 1-9 relate to households perception of crime, sections 10-20 relate to actual incidents of crime, sections 21-28 relate to individual crimes and section 29 was directed to the interviewer to answer.

It is important to note that in this survey, certain categories of crime were under-reported, such as sexual offences and murder respectively. Consequently, these crimes should not be analysed without taking this into account as it would provide unreliable, biased results according to the authors.

Victimisation surveys do have advantages over police reported crime, in the sense that such surveys include incidents that may not be considered a criminal offence to police, for example, the VOCS includes feelings (perceptions) towards crime. It is important to note that even if you may not have personally experienced a particular type of crime, you may be intensely aware of the potential thereof and could consequently be out of sync with reality!

Consequently, the crime incidents actually experienced, together with opinions about crime and opinions on how to minimise crime, is very valuable information. In addition, it is estimated that the victim surveys uncover between 60% and 70% of crime (South Africa World crime Capital, 2001).

We assume that by using a sample (surveyed data), we could accurately determine traits that would be true for the population in general. To illustrate how population estimates could be misleading we refer to a seminar in 2013 when the South African Police Service (SAPS) released the countries' crime trends report for 2012/2013, which was statistically incorrect. The report was based on population totals *estimated* from the 2001 Census to calculate crime ratios for 2011/2012, as opposed to using the actual population total for the 2011/2012 year. Their estimation was out by 1.7 million people, making the crime ratios totally incorrect. The SAPS however still believe that their estimates are correct, based on their own interpretations, but it has been widely agreed that these results are incorrect and has had a detrimental effect on policy making and identifying focus areas for strategic planning of crime prevention and reduction (Getting the most out of South Africa's crime statistics, 2013).

In this study we accordingly will not use the results from the SAPS crime reports. Instead, we obtained aggregated crime data from the SAPS Crime Research and Statistics Unit (SAPS, 2015) which recorded crimes for different categories of crime, namely contact crime, property related crime, crime as a result of police action and aggravated robbery, for each of the 1140 police stations around South Africa. This data included the GPS co-ordinates of the police stations, which will be used in our spatial analysis in the Chapters to follow.

We will use the two data sets for our research, while in chapters 4 through to chapter 6, we will use the first data set (VOCS) and in chapter 7 we use our second data set (Spatial data).

4. Descriptive Statistics

In this chapter we will analyse data from the victims of crime survey (VOCS) descriptively. We will illustrate trends and associations between different categories of crime, for different locations and by demographic information. Victims' perception and reactions are shown, as well as their suggestions on how to combat crime.

Crime categories:

We first categorise the different types of crime. Figure 4.1 gives a representation of the percentages of people surveyed who hold particular perceptions with regards to types of crimes in the country. It is evident that household crimes such as burglary and robbery were perceived to be most frequently committed crime, followed by street crime, such as pick-pocketing and bag or purse-snatching.

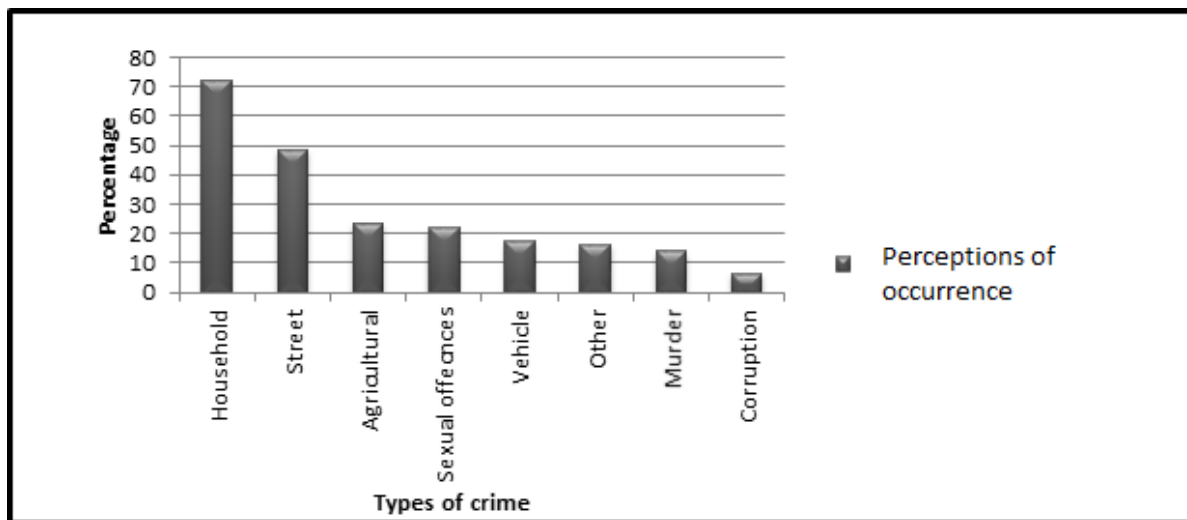


Figure 4.1: Perception of crime in SA (% who believe this crime has occurred)

From Figure 4.1 we have the perceived crimes suggested by the sample group, we next investigate actual incidents of crime experienced by households in the same survey group, over the previous five years, to see the reality of the perceptions held.

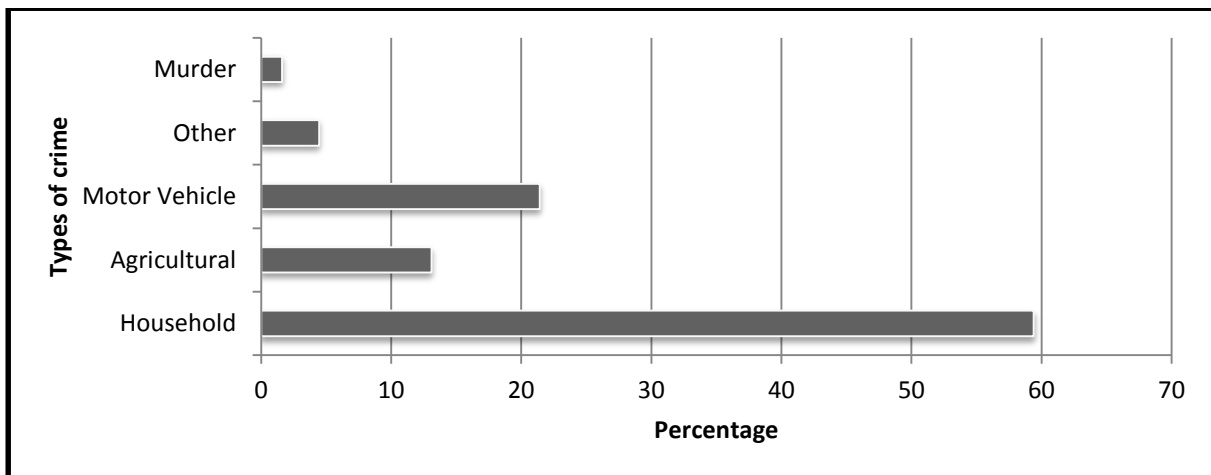


Figure 4.2: Experiences of crime over the last 5 years

Figure 4.2 reveals that household crime was the most frequently experienced crime by South Africans surveyed, i.e. burglary and house break ins. Motor vehicle theft ranked second highest, with 21.4% of those interviewed having experienced motor vehicle theft over the previous 5 years! Agricultural crime (e.g. livestock or crop theft) is also quite frequently experienced by households over the previous 5 years. It further is important to note that the average crime rate for successfully committed crimes is 93%, i.e. 93% of the crimes committed, were in fact successful, which is quite high and dwarfs in comparison with crimes that are attempted, but not carried out.

Crime locations:

We attempt to locate crime at a province level, based on our survey results. Figure 4.3 shows that crime seems to be fairly evenly spread across provinces in the country, with minor peaks in the provinces of Mpumalanga and Western Cape and troughs in Limpopo and the Free State. A more detailed analysis of the distribution of crime will be presented in Chapter 7, using spatial maps.

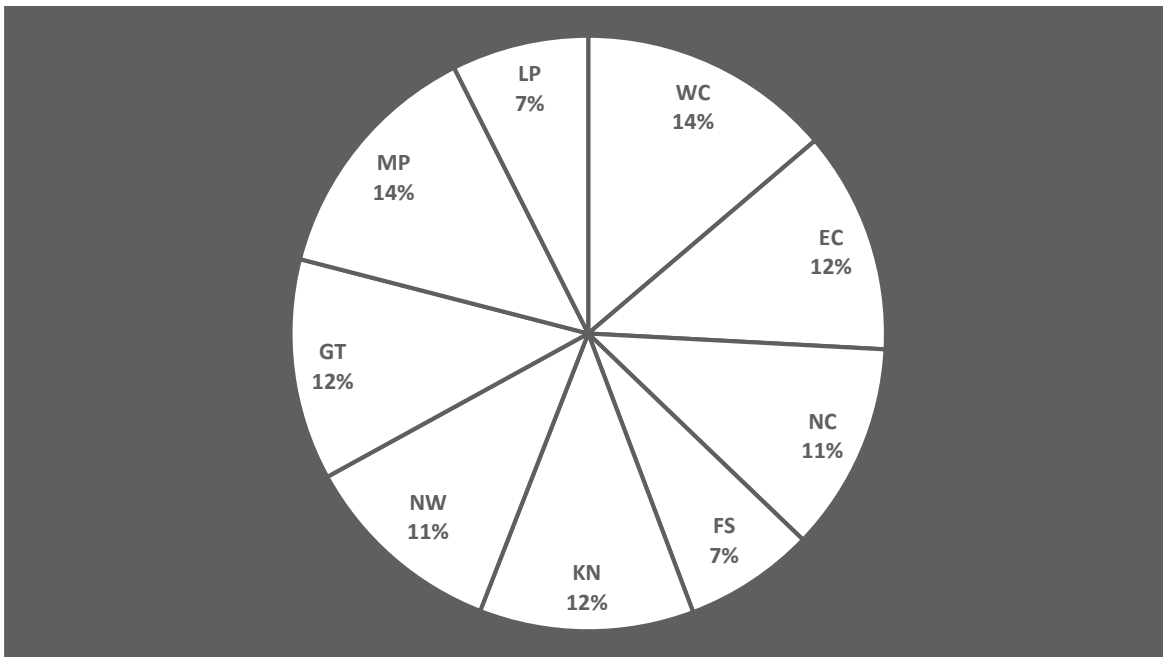


Figure 4.3: Distribution of all crimes in South Africa, by province (%)

We next consider the areas that are believed to have high crime rates. Fear of crime prevents South African residents from doing certain everyday activities, as can be seen from Figure 4.4. In particular, the survey revealed that amongst those surveyed it is clear that going to parks or being in open spaces has the highest perceived risk of crime. Everyday activities which involve children further ranked quite high in the perceived potential for crime, such as children walking to school or playing outdoors was definitely a fear. We further note that the risk of having their house burgled places much fear on South Africans wanting to purchase a house. There is a common trend in perceptions however that crime is most likely to occur in vast, open, desolated spaces in South Africa.

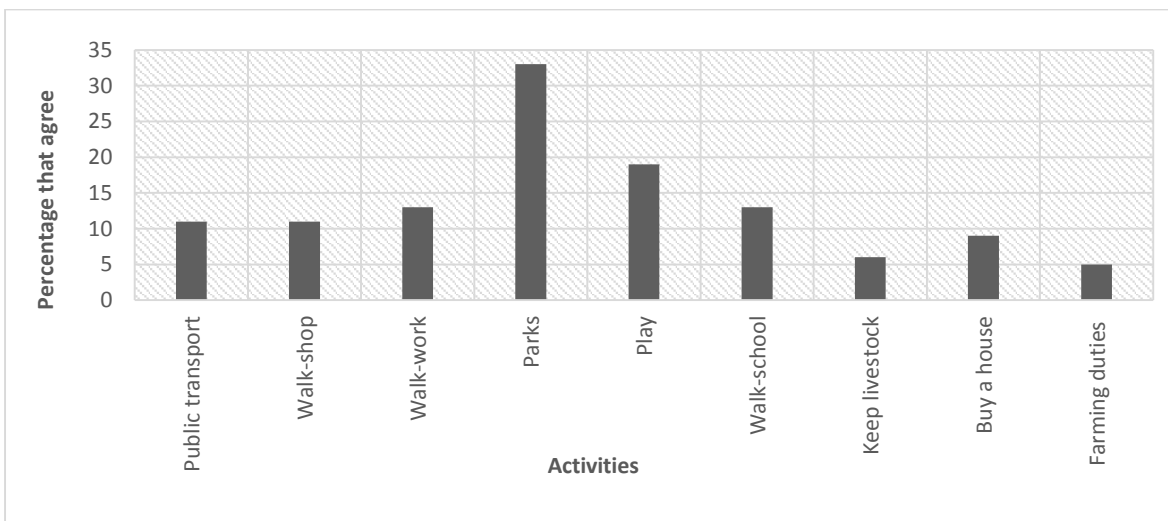


Figure 4.4: Preventions due to crime in SA

Demographic information:

We show victims of crime for different demographic information, first broken down by age group as a percentage, as well as non-victims of crime by age group (categorical) as represented in Figure 4.5. One immediately notices that citizens, in the age group 50 years and older, are the most vulnerable to crime in South Africa, with 32.08% of victims belonging to this age group.

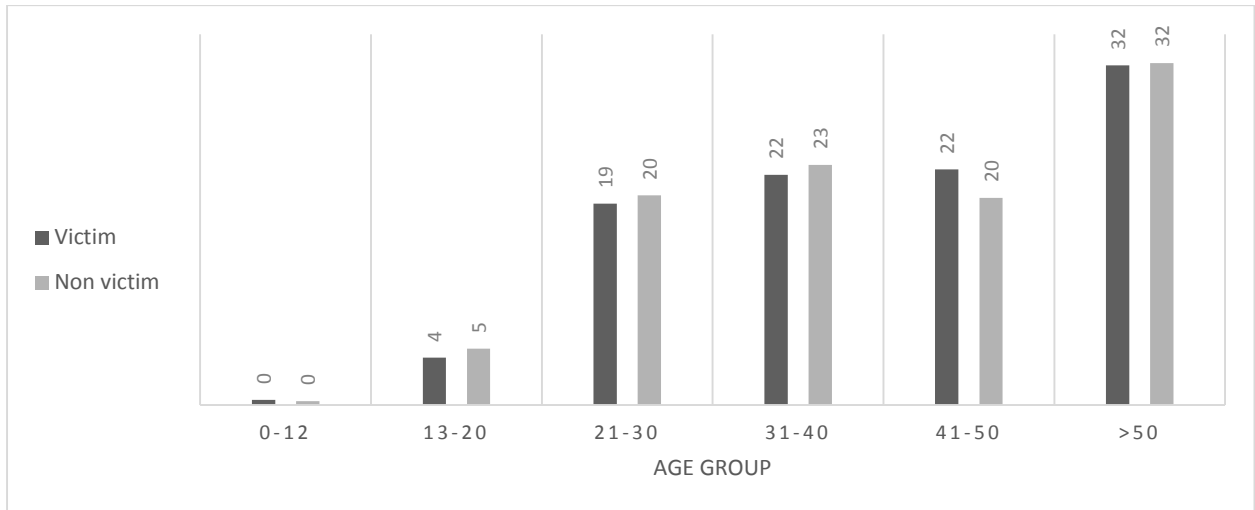


Figure 4.5: Victims and non-victims of crime by age group

Further demographic information of victims (all types of crime) within the last 5 years is broken down by income-type, race and gender, this is depicted in Figure 4.6.

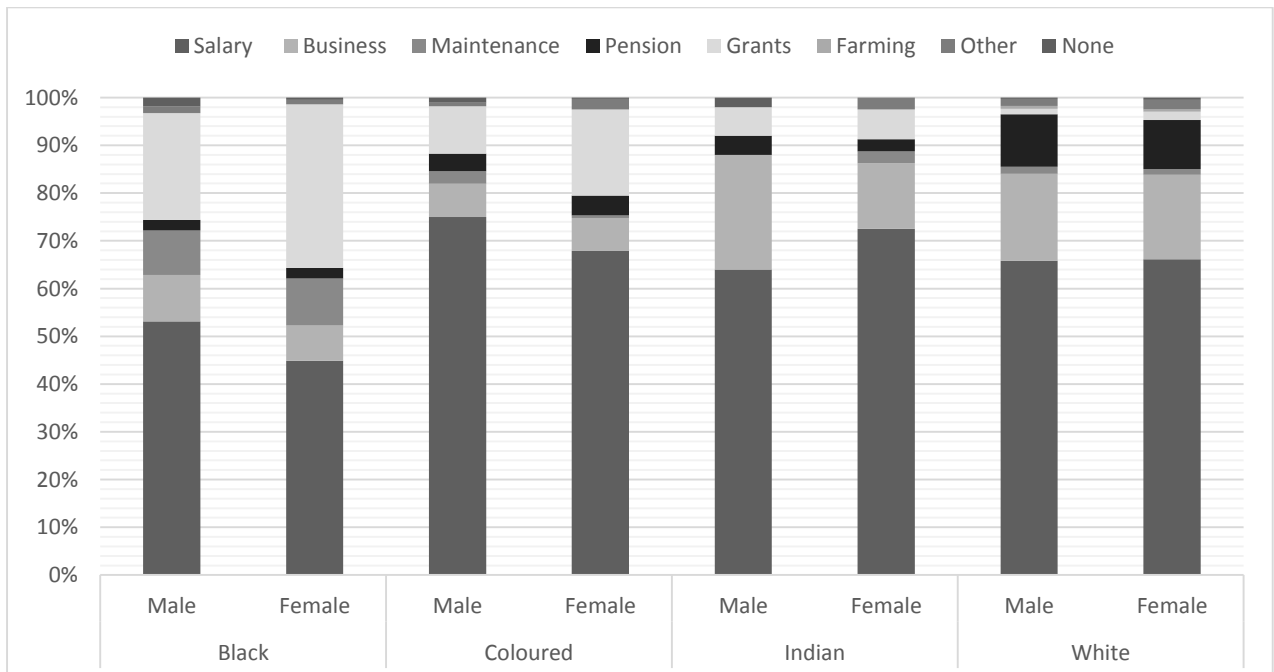


Figure 4.6: Victims of crime (last 5 years), broken down by Income type, race and gender (%)

We notice that across all races and genders, victims of crime are more likely to earn a salary as their main source of income, i.e. crime victims are predominantly salary earners for all races and genders. Receiving grants and maintenance as a source of income is the second most dominant income type amongst victims of crime from the Black race, as is the case, but to a lesser degree, amongst crime victims from the White race (having the lowest proportions of crimes amongst different race groups).

Income from business on the other hand is most common amongst White and Indian victims of crime. We further note that pension pay-outs as a source of income is quite high amongst White crime victims whilst maintenance contributes relatively more to the incomes of Black victims of crime compared to any other race. Sources of income such as pensions, grants and maintenance would describe the income of victims of crime from older age groups and females.

Victim’s reasons for crime:

To understand the reason behind crime being committed, we consider how the survey group felt about corruption in South Africa. They thought the most common reason for corruption was that the perpetrator intended to get rich quickly.

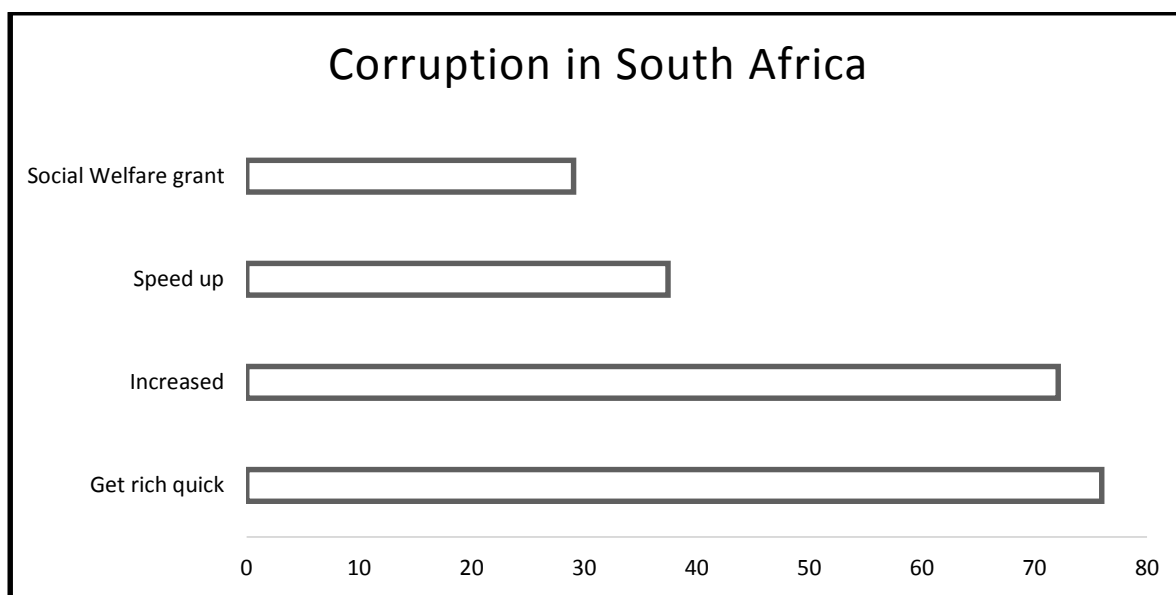


Figure 4.7: Corruption in South Africa – perceptions of the reason for corruption (%)

We note, in Figure 4.7, that around 76% of all those interviewed believe that the reason why people engage in corruption is to get rich quickly (this result was expected, having been mentioned in the literature review), with 72% of households feeling that the level of corruption increased over the last three years, 37% felt that people pay bribes to speed up procedures and 29% reflected that social welfare grant officials were the most corrupt among all the government services!

Victim's suggestions to reduce crime:

We consider the manner in which South African households felt that the government should use resources to reduce crime. This is depicted in Figure 4.8.

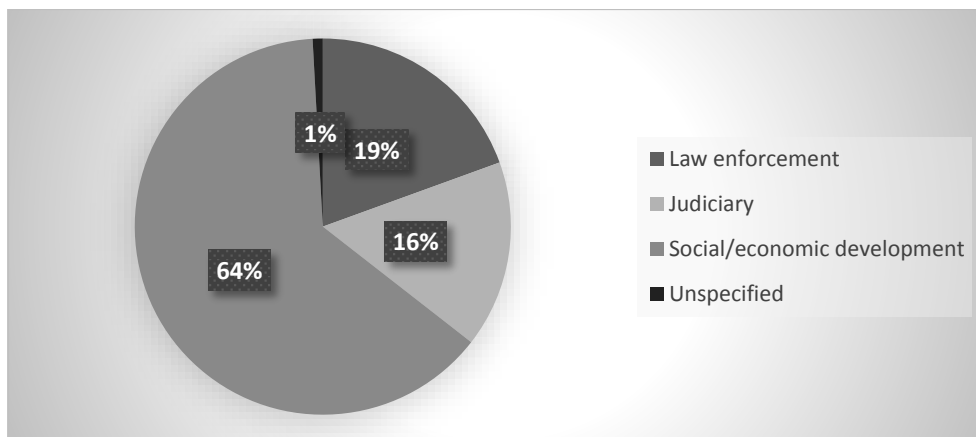


Figure 4.8: Crime reduction methods as suggested by study group

We note that those surveyed largely felt that social and economic developments by government was the best strategy to combat crime, including the undertaking of job creation initiatives.

An analysis was performed on the forms of assistance that the survey group populated. Non-Governmental Organisations (NGO's) and other organisations within the community provide services to victims of crime which include access to medical services, counselling services, or the offer of a place of shelter and safety in the area.

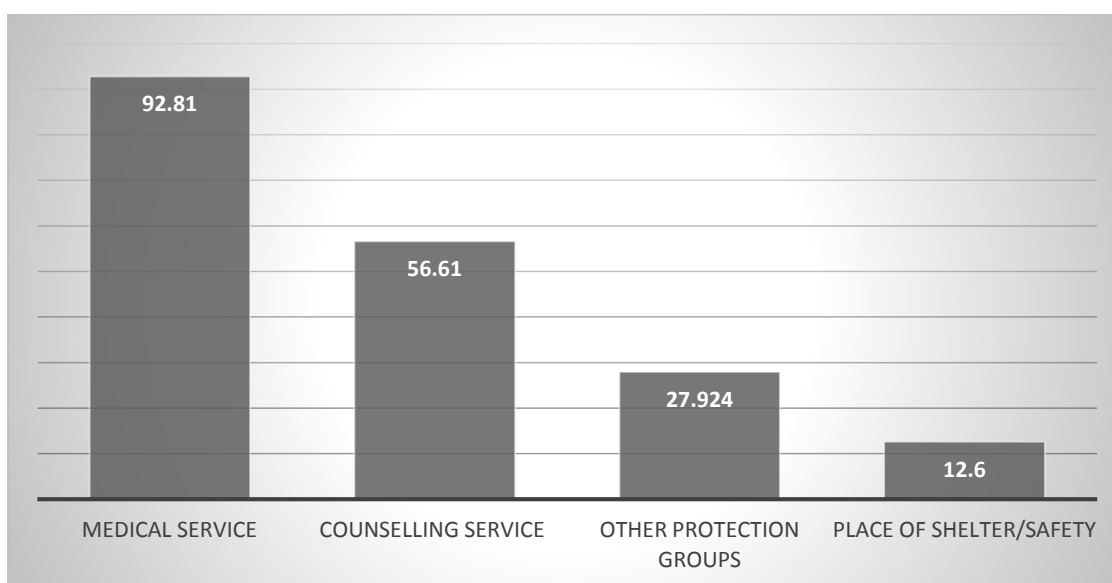


Figure 4.9: Victim support structures (%)

Figure 4.9 indicates that access to counselling services is generally (56.6%) more difficult for households to secure than medical services (92.8%). The lack of a place of shelter and safety in the community is the greatest problem for victims of crime with only 12.6% of South African residents reportedly having access to such a facility. The survey group felt that methods can be put in place to improve these areas.

Victim’s reaction to crime:

We investigated the reaction of the study group during an incident of crime (the first port of call when faced with crime incidents). Results from the survey for this question are summarized in Figure 4.10 and reveal who is the person or organisation individuals feel they will first contact when faced with crime.

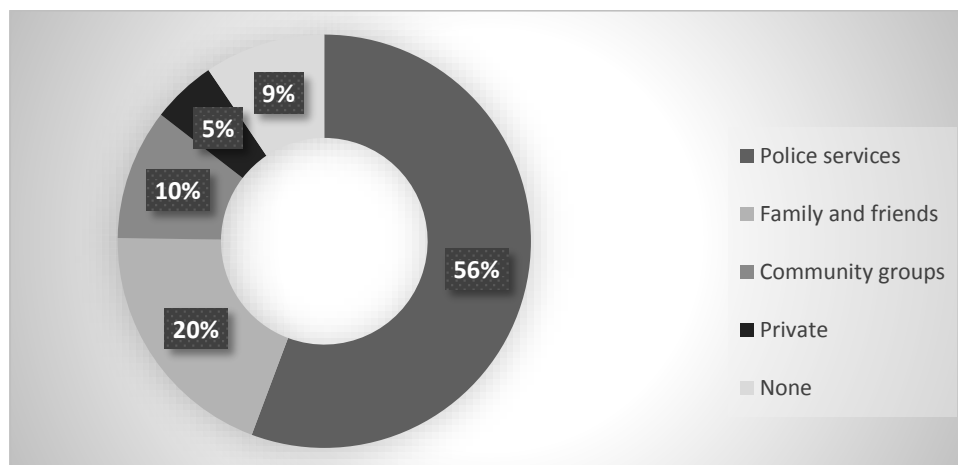


Figure 4.10: Responses to crime incidents (%)

It comes without a surprise that the most common response to crime (56%) is to contact the South African Police Service (SAPS). However, it is immediately clear that a significant proportion of citizens do not report crime to the police. Instead, a high percentage of South African victims of crime seek refuge from relatives or friends, followed by community groups, i.e. traditional authorities. Educational programmes should thus be aimed at community members. Further, the partnership between the criminal justice system and the traditional authorities needs to be strengthened, so that the police and community can more effectively work collaborating to combat and deal with crime.

Victim’s thoughts on Policing:

The opinions of the study group on the response times to crimes of the South African Police Service (SAPS) is given in Figure 4.11.

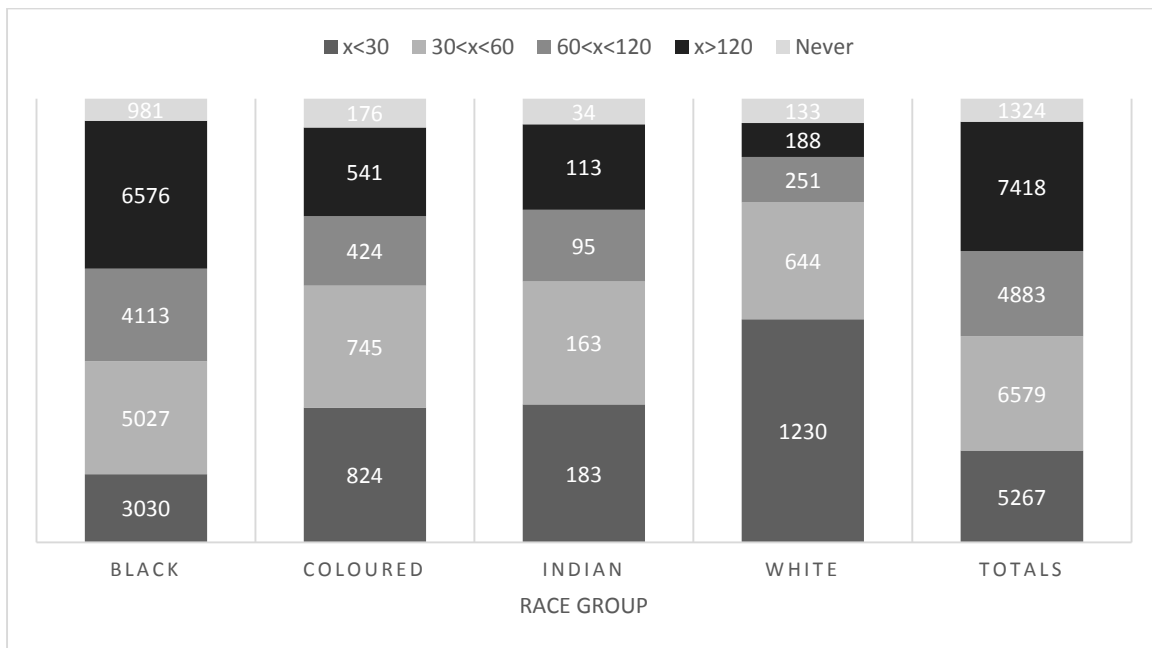


Figure 4.11 Response times (minutes) of police officers, as reflected by different race groups (%)

We note that amongst citizens surveyed, 29% felt that the SAPS took more than 2 hours ($x > 120$ minutes) to respond to a call of emergency. It is worth noting that the proportion of Black residents that felt that it takes more than 2 hours for the SAPS to react to a crime, outweighs the similar proportion for any other race type. The other race groups predominantly felt that the police took less than 30 minutes to arrive to an incident scene, indicating that they had an experience of shorter response times to crimes than was the case for Black citizens. A very low percentage of people (5%) felt that the SAPS never arrived at a crime incident scene!

Next we investigate satisfaction levels of the survey group with the police in general, or with the way in which punishment is handed down by the court.

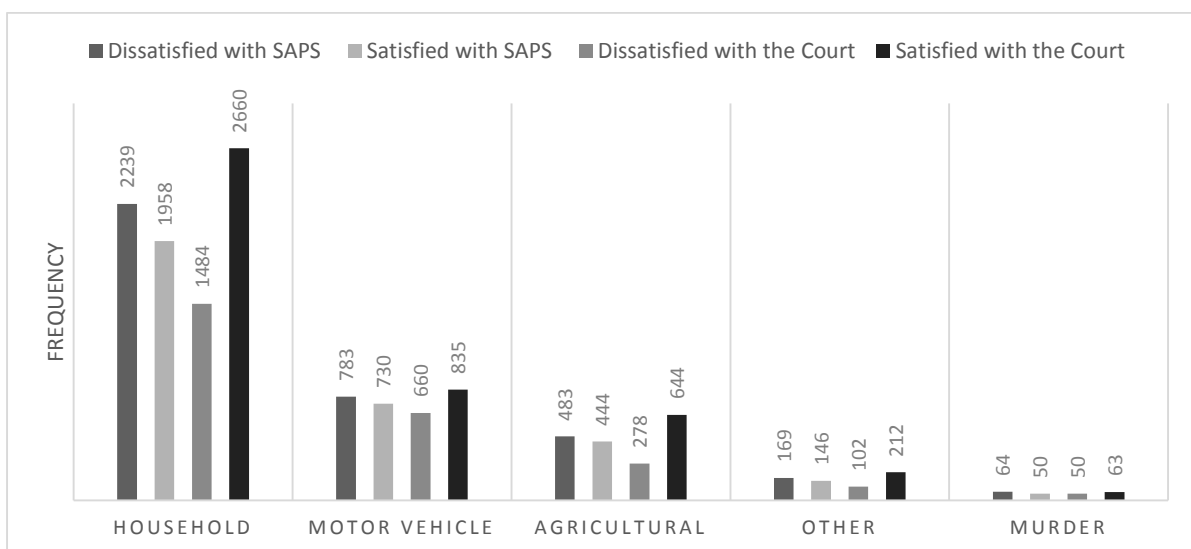


Figure 4.12: Victims satisfaction levels broken down by types of crime

Based on Figure 4.12, we conclude that across all types of crime, the survey group expresses satisfaction towards the decisions made by the court. In contrast to this satisfaction of the victims, we find that they are equally dissatisfied with the South African Police Service (SAPS). The job description of a Police Official is to be involved in preventing, combating or investigating crime (SAPS careers, 2014), however it is felt by this study group that these tasks are not being executed to their satisfaction, in comparison to the way that tasks are performed by the courts.

Highlighting only murder, we notice that there is no considerable difference between satisfaction levels of the study group, evidence by the significant height differences of the bars for the other types of crime in Figure 4.12.

This can be explained by a comment made by Sibusiso Masuku, who points out that only half of murder cases were sent to court, while only a fraction of those resulted in a guilty verdict (Masuku, 2003). We can consequently assume that an equal dissatisfaction is experienced by the study group towards the police, as well as the courts for Murder, in the 2013/2014 study.

We next focus on regularity of visible policing, i.e. how often those surveyed felt that they could see police officers patrolling in their province. Overall 35% of the study group reported that they see a police officer patrolling at least once a day, while 16% of households reported that they never see a police officer patrolling in their province.

We note from Figure 4.13 that in the Eastern Cape, 34% responded that they never see a police officer, which is surprising as the Eastern Cape has 195 police stations, the highest number of police stations per province in the country (total of 1140 police stations in the country).

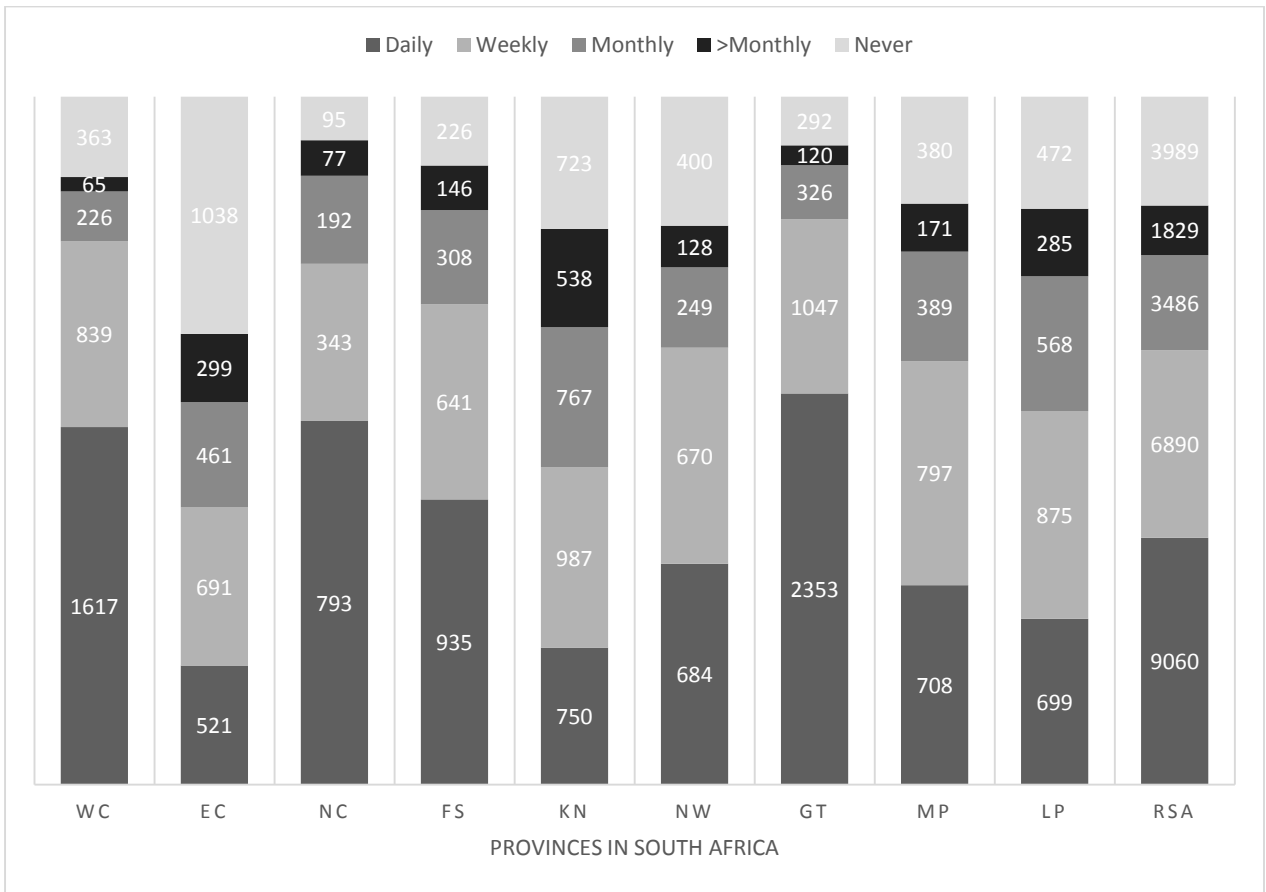


Figure 4.13: Visible policing by province

Perception vs Reality:

Figure 4.14 illustrates the perception of crime, along with the crime statistics (reality), so that the link between reality and perception is depicted graphically.

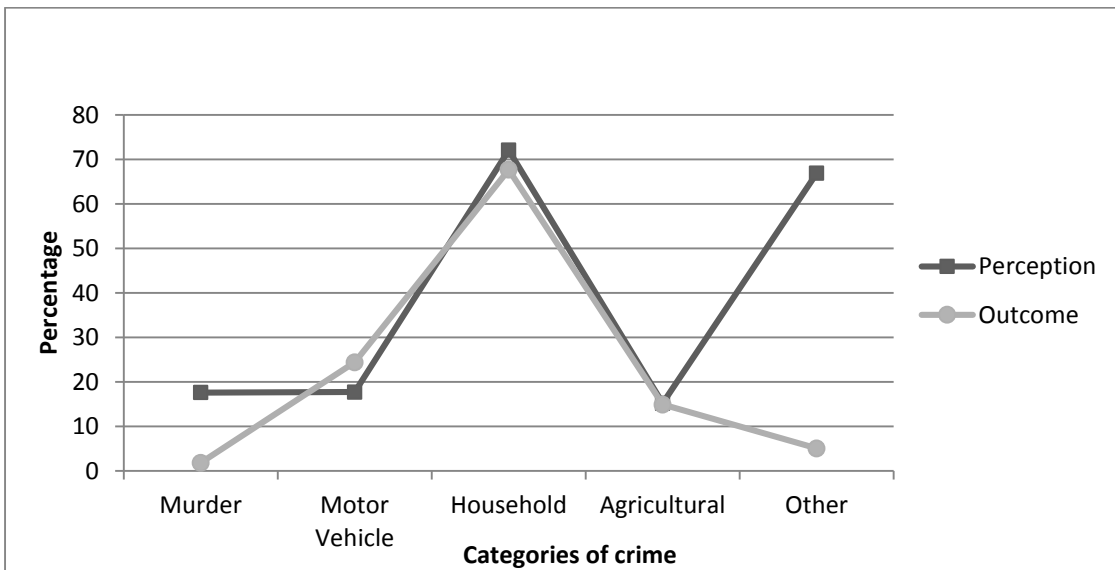


Figure 4.14 Perception verses outcome of crime

Perception of being a victim of crime, is equivalent to an individual having a fear of that crime, while a victim of crime refers to those who have experienced an incident of crime. We note that the perception (fear) of crime matches the crime outcome, for most categories of crime, except for murder. We note that murder, given its severity, is over-perceived, possibly as it is so highly feared. We note that rape was the most feared crime, second only to murder (Masuku, 2002). The “other” crime category is the outlier, where the actual crime far outweighs the perception (fear) of that crime category, but given its vague interpretation (it has various sub categories), it is possibly not surprising that the study group did not accurately or realistically perceive this type of crime.

A cross tabulation of the perceptions (fears) of crime, against crime that has actually occurred, is given in Table 4.1. Comparing by row, we note that 23.28% of households who feared being a victim of crime, had actually experienced crime, while 0.96% of households who do not fear crime, have been victims of crime in South Africa. Citizens who fear crime were thus almost four times more likely to become victims of crime. This confirms that their fear of crime is justified opposed to the small subset who are unaware of crime.

We denote that the probability of being a victim of crime, when one perceives that one could be a victim of crime, exceeds the probability of being a victim of crime when one does not perceive being a victim of crime. This is contrary to the hypothesis of independence between perception and occurrence of crime.

Table 4.1 Perception verses Outcome contingency table

	Victim of crime	Not a victim of crime
Perceived being a victim of crime	5961 (23.28)	16473 (64.33)
Did not perceive being a victim of crime	245 (0.96)	2926 (11.43)
Total	6206 (24.24)	19399 (75.76)

Table 4.2 Odds Ratio output

Type of Study	Value	95% Confidence Limits	
Odds Ratio	4.3217	3.7809	4.9398

The interpretation is done using Table 4.2, as the odds ratio reflected in this table provides an estimate of the relative risk when an event is rare.

The odds ratio for victims of crime, with regards to fear, is calculated by performing a cross multiplication using Table 4.1, i.e. $(5961 \times 2926) / (16473 \times 245) = 4.3217$.

This indicates that the probability (odds) of becoming a victim of crime among those who fear crime is 4.32 times higher than those who do not fear crime. The narrow confidence interval [3.7809; 4.9398] further indicates that this estimate has high precision.

Many of the relationships found in this chapter by examining features of the data descriptively, will be analysed in more depths in chapters to follow.

This chapter was useful in illustrating the relationship between the different attributes of South Africans with regards to crime. The chapters to follow will take on a statistical approach, analysing using statistical tools, predictive models and hotspot maps.

5. Logistic Regression

In this chapter we introduce our first statistical model, where we attempt to predict patterns of crime using characteristic traits of the study group and the occurrences of crime as the response.

Logistic regression models the relationship between a binary response variable (Y) and one or more explanatory variables (vector \mathbf{X}_i) (Wang, 2011). In this study, the binary response variable (Y) was whether the individual had been a victim of crime (Yes; No). We are interested in modeling different types of crime using various exploratory variables (\mathbf{X}_i) (where \mathbf{X}_i represents attributes such as Gender, Age, Province, Income type, Race, etc.).

Generalized linear models, opposed to linear regression models, equates the linear component to a *logit* transformation (natural logarithm) of the probability of a given outcome on the dependent variable (Czepiel). We then set up a model as follows:

$$\text{Logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{i=1}^n \beta_i X_i \quad (1)$$

Where α is the intercept parameter, β_i denotes the coefficients of \mathbf{X}_i , representing the parameter estimates and \mathbf{X}_i are the explanatory variables mentioned for $i = 1, 2, \dots, n$. The vector β_i is calculated using the maximum likelihood estimation method, computed using a statistical software package, SAS.

After some algebra in solving for π_i from Equation 1, the probability of success i.e. $Y=1$, is given by

$$\pi_i \stackrel{\text{def}}{=} \left(\frac{1}{1 + e^{-(\alpha + \sum_{i=1}^n \beta_i X_i)}} \right) \quad (2)$$

The *logit* is an expression for the ‘log odds’ of the outcome Y , under a specific set of \mathbf{X}_i , so that the odds from Equation 2 is given by

$$OR = \frac{\pi_i}{1-\pi_i} = e^{(\alpha + \sum_{i=1}^n \beta_i X_i)} \quad (3)$$

The odds ratio is the ratio of the probability that event Y will occur, divided by the probability that event Y will not occur (Kleinbaum & Klein, 2002). We will use the odds ratio to interpret the effect of factors further in this analysis.

To decide upon model adequacy in this study we will use the Hosmer and Lemeshow test with the Pearson statistic,

$$\sum_{i=1}^t \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - \sum_j \hat{\pi}_{ij}/n_i]} \quad (4)$$

Let y_{ij} denote the binary outcome for observation j in group i of the partition, where n_i denotes the number of observations and $\hat{\pi}_{ij}$ denotes the corresponding fitted probability to the model, $i=1, \dots, t$ and $j=1, 2, 3, \dots, n_i$. The Hosmer and Lemeshow statistic will indicate whether the fit is decent or not, but will not detect any types of lack of fit (Agresti, 2002).

To support this test we will also use the Receiver Operating Characteristic (ROC) curve, which was derived from signal detection theory, used during World War II for the analysis of radar images. The area under the ROC curve measures accuracy, i.e. the ability of the test to correctly classify the outcome success of the study. The ROC curve is fitted using the maximum likelihood estimator method through statistical software and the area under the curve represents the percentage of randomly drawn pairs for which the test correctly classifies the group of the individual (The Area Under an ROC curve, 2015).

Several models will be constructed based on the explained approach. First we will consider an individual's perception of crime in South Africa; where the response variable will be a success if they did perceive that they will be affected by crime and a failure if they perceived to not be affected by crime, this will be split into different models for the different categories of crime, i.e. Murder (and attempted murder), Motor Vehicle related crime (theft, damage), Household crime (burglary), Agricultural crime (theft of crops and livestock) and Other types of crime. Similarly we will consider an individual's victim status, where a success denotes whether an individual has been a victim of crime or a failure if they have not.

The model introduces explanatory variables as factors: Age as a continuous variable and Gender, Province, Income type, and Race of households in South Africa as categorical variables. The response variables for this model is binary (yes or no to the question "have you experienced crime?").

The categorical variables have respective reference groups Male, KwaZulu-Natal Province, No Income and the White Race group.

Considering Table 5.1, we find that at a 5% level of significance, the factors Gender, Age, Race, Province and Income are all significant in the model as well as the interactions Province*Race, Age*Province and Gender*Province are significant.

Table 5.1: Summary of significant factors

Effect	DF	Wald Chi-Square	Pr > ChiSq
Province*Race	24	90.4236	<.0001
Gender*Province	8	24.7277	0.0017
Age*Province	8	19.1765	0.0139
Province	8	60.0907	<.0001
Gender	1	10.4885	0.0012
Age	1	10.4933	0.0012
Income	5	74.5455	<.0001
Race	3	27.6495	<.0001

A significant Province*Race interaction for example, means that a household's victim status will be influenced by the household's race but this status will vary from one province to another. Similar interpretations exists for the Age*Province and Gender*Province interactions. We note that all these interactions include Province which supports our aim of locating crime.

The interpretation of the parameter estimates is for example, if Income (Business) = 0.4337, then compared to a household belonging to Income (None), the log odds of Victim status is 0.4337, however the parameter estimates outputs are omitted from this chapter.

The Hosmer and Lemeshow Goodness-of-fit test is used for model adequacy, the test statistic calculation is shown in Equation 4. We find that the HL test statistic is $\chi_{HL}^2 = 5.3692$ with 8 degrees of freedom and p-value=0.7175, which exceeds 0.05 (we do not reject model adequacy with 95% certainty), thus indicating that this measure supports the models adequacy for this data.

We further find that the area under the ROC curve is 0.595, with 58.8% of the observed pairs being concordant. We then conclude that the model is adequate.

The odds ratios are given in Table 5.2, this helps us to understand the outcome of crime in relation to income better. We established that the risk of being a victim of crime for those who receive income from farming is 1.727 times the risk for those who receive a fixed salary. The 95% confidence interval for the odds ratio, has a narrow width and includes the value of 1.0, so it is plausible that the true odds of being a victim of crime are equal for farmers and those who receive a salary.

Table 5.2: Odds Ratio Estimates for victims of crime

Effect	Point Estimate	95% Wald Confidence Limits	
		Lower	Upper
Income Business vs Salary	1.500	1.345	1.673
Income Farm vs Salary	1.727	0.796	3.747
Income None vs Salary	0.972	0.798	1.185
Income Other vs Salary	0.924	0.830	1.028
Income Pension vs Salary	0.910	0.844	0.981

The effect plots presented in Figure 5.1 through Figure 5.3, profiles the predicted probability of crime for different attributes of an individual based on our study.

The association of the gender and province in Figure 5.1, reveals that the probability of becoming a victim of crime is higher for females as compared to males, in the provinces of Free State, KwaZulu-Natal, Gauteng and Mpumalanga.

This result follows from controlling for Age at 43 years, income type as “none” and race group “White”.

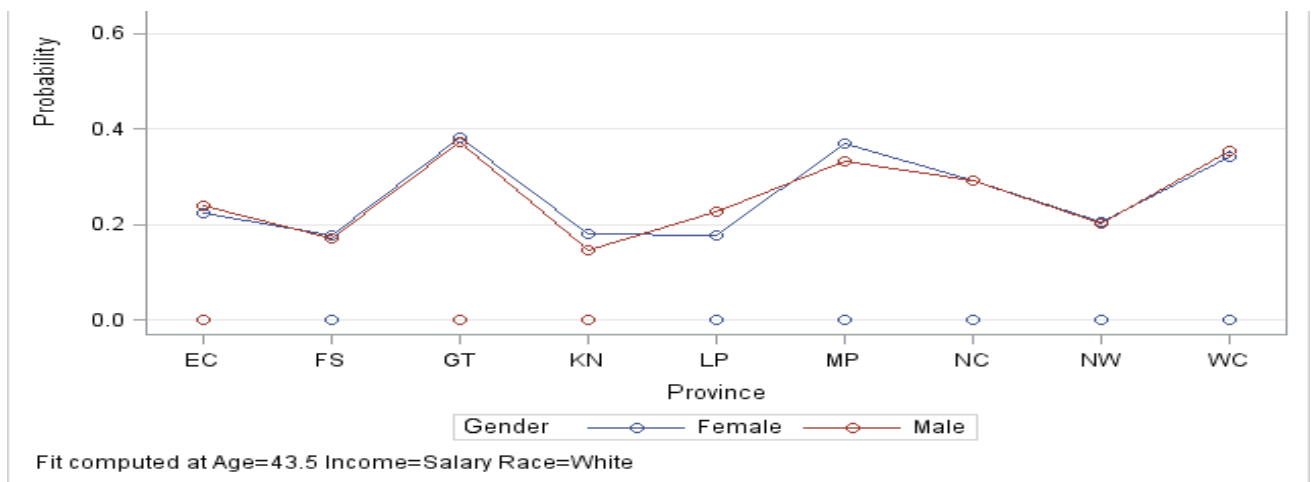


Figure 5.1: Probability diagram for interaction Gender*Province

Based on Figure 5.2, the predicted probability of being a victim is the highest for the White race group across provinces Western Cape, Northern Cape, Free State, Gauteng, and Mpumalanga. The Black race group has a higher probability of being a victim of crime in the provinces of Eastern Cape and KwaZulu-Natal, with the Indian race group having greatest probability of being a victim of crime in the North West province and the Coloured race group has the highest risk of being crime victims in the province of Limpopo. These results are based on taking Gender fixed as Females, income type as “none” and an average Age is set.

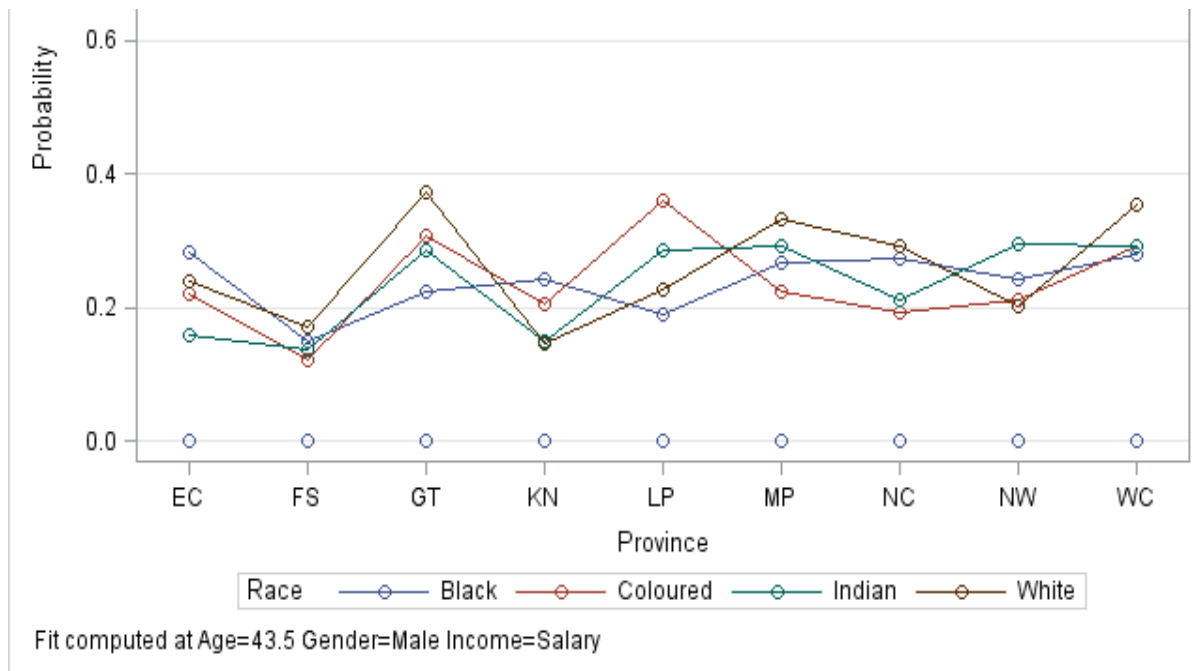


Figure 5.2: Probability diagram for interaction Race*Province

Figure 5.3, shows that the probability of being a victim of crime for those under the age of 20 years, is the highest in the province of the Western Cape and Gauteng also substantially high. The probability of being a victim for those over the age of 20 years, is generally higher, whilst being in the province of Gauteng, this probability constantly increases with age.

All provinces reveal a linear correlation that is either positive or negative as age increases. It is important to note that as age increases, so does the probability of being a victim of crime in KwaZulu-Natal (this line has the steepest gradient). The White race group, female gender and no income were set constant in Figure 5.3.

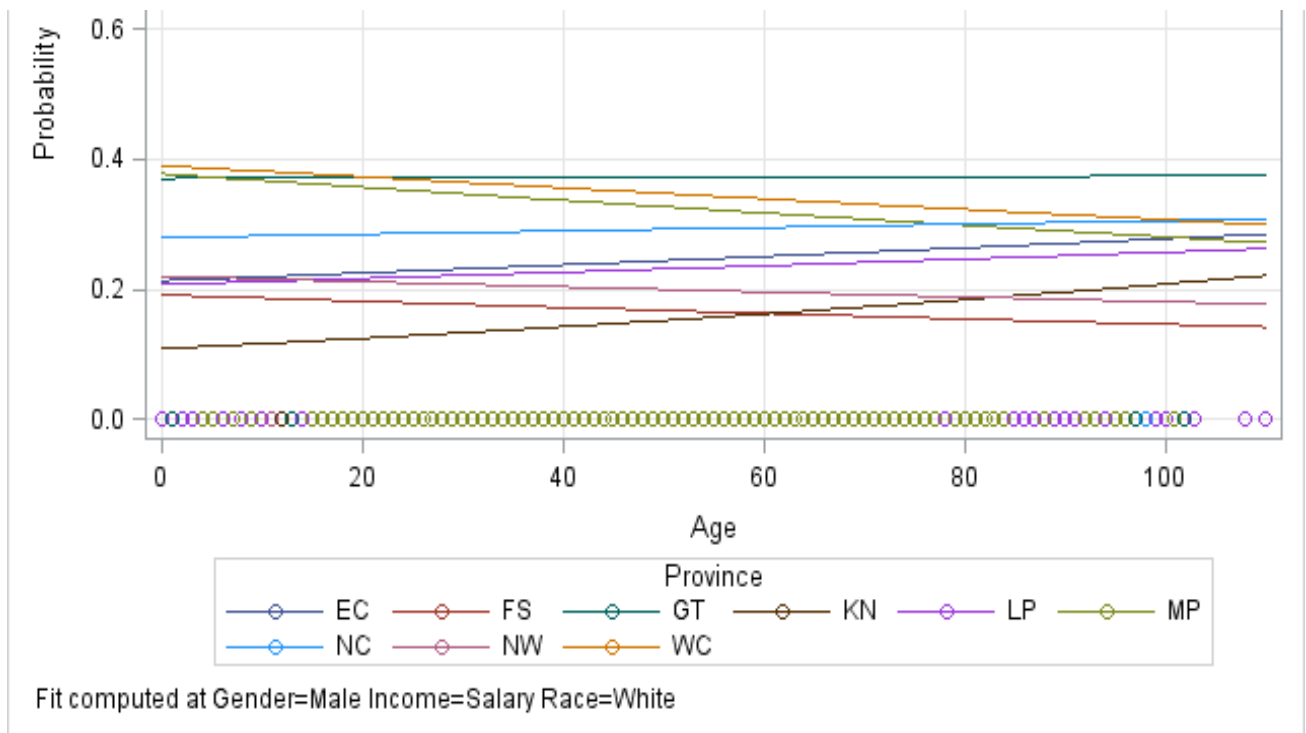


Figure 5.3: Probability diagram for interaction Age*Province

In the same way we can model the response variable of being a victim of crime or not, for separate categories of crime, for example, the probability modeled is Murder experienced, Motor Vehicle theft experienced, and so on. The individual results are tabulated in Table 5.3 for simplicity. The values arise from conducting a statistical analysis using SAS. All models follow the same methodology of a Logistic regression.

Table 5.3: Summary of SAS output for victim models

Types of crime	Variable	Estimate	P-value	Odds Ratio example	HL p-value	ROC
MURDER	Province	29.6417	0.0002	Race Black vs White: 2.624	0.7342	0.685
VEHICLE	Province	30.3644	0.0002	Province WC vs KN: 3.153	0.0294	0.732
	Income	124.8734	<.0001			
	Race	83.8733	<.0001			
	Age*Race	15.7853	0.0013			
	Age*Province	15.2094	0.0552			
AGRICULTURE	Age	7.4081	0.0065	Income Farm vs Salary : 14.591	0.0331	0.790
	Province	48.4248	<.0001			
	Income	174.2171	<.0001			
	Race	41.5574	<.0001			
HOUSEHOLD	Province	13.9634	0.0827	Gender Female vs Male: 1.032	0.4565	0.594
	Income	71.9098	<.0001			
	Province*Race	54.9356	0.0003			
OTHER	Province	58.2288	<.0001	Province MP vs KN: 1.693	0.5345	0.578
	Income	14.1256	0.0148			
	Race	7.1174	0.0682			

Analyzing Table 5.3, we view only significant variables, either at a 5% or 10%, which we derive because the p-values are below 0.05 or 0.1 respectively. For example, Province is significant at a 10% level for the Household crime model (we can only be 90% sure), while the other factors were found to be significant for Household crime at a 5% level of significance. A significant main effect in a model suggests that that attribute (Age, Race, etc.) of the individual in the study group significantly affects the odds of being a victim of that crime category or not.

Table 5.3 further shows which models have a significant interaction term (association of attributes) we note that the victims of Murder, Agriculture and Other types of crime models have no interaction terms that explain the model. An interaction term, for example, Race*Province in the victims of Household crime model, suggests that the Province of the individual surveyed depends on the Race of the individual because in interaction these variables predicted whether or not they were a victim of burglary or not, Race*Province (Black*GT) predicts that burglary occurs more in Black South Africans from the Province of Gauteng compared to the reference groups being White individuals from KwaZulu-Natal, keeping the factors of Gender, Age and Income type constant.

Model adequacy, which tells us how well the data fits or explains the model, will be measured by the Hosmer and Lemeshow p-value as well as the ROC curve areas. The Hosmer and Lemeshow performs a test and for example in the victims of Murder model a p-value of 0.7342 (Table 5.3) is given which implies that we do not reject model adequacy at the 0.05 level, thus this measure supports the models adequacy for the data. Using this test we see that model adequacy is questionable for the victims of Motor Vehicle crime and Agriculture crime models. We will thus use the Hosmer and Lemeshow test as an alternative to the ROC diagnostics test. However, the ROC curve areas provide more influential results. This diagnostic test can be classified into different levels of accuracy:

- .90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail, (The Area Under an ROC curve, 2015).

Regarding our models, the victims of Agricultural crime model has the highest ROC value of 79% (in other words having an area of 0.79 under the ROC curve), which means that this model is the most accurate (although only fairly accurate as area lies in the interval 0.7-0.8) and especially this accuracy tells us how well the test separates the two groups being modelled, namely those being a victim of Agricultural crime and those who are not.

We can thus conclude that victims of crime for Murder and Motor Vehicle crime models are poor and fairly accurate respectively, but the victims of crime models for House hold and Other crimes fail the accuracy test. However, these two mentioned models that fail the accuracy test, do pass the adequacy test using the Hosmer and Lemeshow p-value (these statistical measurements can be found in Table 5.3).

Lastly, a few significant Odds Ratio Estimates are listed. Again we can measure the strength of association using the knowledge that:

- $OR > 3$ suggests a strong association
- $1.6 \leq OR \leq 3$ suggests a moderate association
- $1.1 \leq OR \leq 1.5$ suggests a weak association, (Wang, 2011)

From our results in Table 5.3, the factor of Province is strongly associated with the victims of Motor Vehicle related crime and Income type is strongly associated with the victims of Agricultural crime. To explain what an Odds ratio estimate means, we use the model that predicts the victim of Agricultural crime, given in Table 5.3 is Income Farm vs Salary = 14.591. The risk (odds) of being a victim of Agricultural crime among South Africans who receive income from farming is about 14.591 times that of those earning a salary. This result comes as no surprise! Consider one further odds ratio estimate, in the victims of Motor Vehicle crime model, where the odds of an individual being a victim of Motor Vehicle related crime in the Western Cape Province is 3.153 times what it is for an individual from the Province of KwaZulu-Natal. This accounts for a detailed analysis of the SAS output for the five different Victims of crime models.

We consider the perception that individuals have of crime, regardless of whether they have been a victim of crime, i.e. they still perceive that they might become victims. Their perception differs for different types of crime (there are five categories of crime). These models will predict the probability that an individual perceives to be a victim of the different types of crime.

Table 5.4 summarizes the SAS output for the five different Logistic Regression models. The probability modelled is for example whether the individual perceived to experience household crime or any of the different types of crime.

Table 5.4: Summary of SAS output for perception models

Types of crime	Variable	Estimate	P-value	Unadjusted Odds Ratio example	HL p-value	ROC
MURDER	Gender	2.8782	0.0898	Income Farm vs Salary = 1.685	0.8976	0.634
	Age	3.1336	0.0767			
	Province	16.0108	0.0422			
	Income	9.7845	0.0816			
	Race	27.0256	<.0001			
	Gender*Province	14.9154	0.0608			
	Gender*Race	9.9205	0.0193			
	Age*Province	14.5847	0.0677			
VEHICLE	Province	49.4702	<.0001	Gender Female vs Male = 0.966	0.8947	0.718
	Income	10.1713	0.0705			
	Race	105.4023	<.0001			
	Age*Province	19.4863	0.0125			
	Age*Race	8.1821	0.0424			
	Province*Income	61.2416	0.0169			
	Province*Race	115.0462	<.0001			
	Income*Race	23.1978	0.0571			
AGRICUL	Age	8.0051	0.0047	N/A	0.9410	0.800
	Province	52.7453	<.0001			
	Income	16.3245	0.0060			
	Race	39.9981	<.0001			
	Age*Income	16.9076	0.0047			
	Province*Income	113.5686	<.0001			
	Province*Race	109.9319	<.0001			
	Income*Race	44.9083	<.0001			
HOUSE	Province	71.0864	<.0001	N/A	0.0447	0.581
	Income	10.5720	0.0606			
	Gender*Province	19.1151	0.0143			
	Province*Income	56.5205	0.0434			
	Province*Race	115.2161	<.0001			
	Income*Race	25.4623	0.0303			
OTHER	Gender	3.1031	0.0781	N/A	0.7332	0.631
	Province	34.7240	<.0001			
	Income	27.5201	<.0001			
	Race	67.1975	<.0001			
	Gender*Province	19.2437	0.0136			
	Gender*Income	10.1267	0.0717			
	Age*Province	21.0572	0.0070			
	Province*Income	134.4919	<.0001			
	Province*Race	111.6293	<.0001			

Now using Table 5.4, we find that for example the model for murder, has all main effects significant in the model and further that the interactions Gender*Race, Province*Race, Gender*Province and Age*Province are all significant.

To express what these mean we will use Table 5.5.

Table 5.5: Parameter estimates output (subset)

Gender(Female)*Race(Indian)	1.0495
AGE	0.00510

This is the ordered log-odds estimate for a one unit increase in age on the perception of Murder, given the other variables are held constant in the model. If a subject were to increase in age by one year, you'd expect the log-odds of their perception of Murder crime to increase by a factor of 0.0051 given in the table above, on the ordered log-odds scale, while the other variables in the model are held constant.

Similarly, a Gender*Race interaction for example, also in Table 5.5, infers that a household's perception of Murder will be influenced by the household's gender, but this perception will vary between races.

Individuals who are Indian Female, compared to those that are White Male, have an increased log odds of their perception of Murder (by 1.0495- from Table 5.5). By this logic, we see that the perception for Agricultural crime model is the model that is most explained, where four main effects and six interaction effects make up this model. Among the five types of crime, Agricultural crime is thus the best described and predicted amongst all crime types mentioned.

The Hosmer and Lemeshow Goodness-of-fit test for model adequacy, displays p-values all greater than 0.05, aside from the perception for Household crime model. This means that all the models are adequate, aside from the perception for burglary model. Even though this model fails the adequacy test, it needs to be checked for accuracy. We further find that the area under the ROC curve is high overall, but accuracy fails for the perception of Murder model. The concerns for this model failing are not severe as an individual's view towards burglary (Household crime) could be exaggerated. The most accurate model is shown to be the perception of Agricultural crime model (accuracy level = good), with an area of 0.8 under the ROC curve.

The odds ratio estimates are not calculated for the perception of Agricultural, Household and Other crime models (represented as N/A). The possible reason for this could be that there are no events, or all events are observed in both groups, either a denominator of zero arises or the standard errors cannot be calculated (Measures of relative effect: the risk ratio and odds ratio, 2015).

The odds ratios that are present in the table, help us to understand the perception of Murder and Motor Vehicle crime better. For example, given Gender Female vs Male=0.966, we conclude that the odds for perceiving to be a victim of Motor Vehicle theft, is 0.966 times higher for females than what it is for males, keeping the other factors of Age, Race, Income type and Province fixed.

We now use the survey logistic procedure (a tool for logistic regression when using survey data). We understand that the sample mean, \bar{y} in Equation 5, is an unbiased estimator of the population mean \bar{Y} .

$$\bar{y} = \frac{\sum_i^n y_i}{n} \tag{5}$$

$$var(\bar{y}) = \frac{\sum_i^n (y_i - \bar{y})^2}{n(n - 1)} * \left[1 - \frac{n}{N}\right] \tag{6}$$

Where $i=1, \dots, n$ and n is the total number of observations in the sample. The corresponding sample variance without replacement is given in Equation 6, with N as the population size, the last term is known as the finite population correction.

Beyond the theoretical components, we note that the survey data has to include certain features. Those are clustering (the data should be partitioned), stratification (a mutually exclusive group variable) and unequal weighting (number of population units a sample unit represents).

The survey design of our research is to use the primary sampling unit (PSU) number as the cluster variable (there were 3080 in the Master Sample), the binary household type (Metropolitan or not) was the strata variable and the weights were calculated as the inverse of the sampling rate (for example, in the Western Cape province those sampled were 94.9% responsive, the inverse of this would allocate a weight of 1.05).

The Response Profile reveals that 6558 (weighted) of the respondents were victims of crime.

Table 5.6: Survey Logistic Profile

Response Profile			
Ordered Value	Victim	Total Frequency	Total Weight
1	0	19399	20413.1
2	1	6206	6557.86

The Type 3 analysis in Table 5.6, similar to our previous Logistic regression results (Table 5.1), shows that all main effects and three selected interaction terms were found significant.

Table 5.7: Parameter estimates output (survey)

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	8.8604	0.0029
Gender	1	8.3447	0.0039

Province	8	47.5903	<.0001
Income	5	74.2602	<.0001
Race	3	19.0239	0.0003
Age*Province	8	18.3379	0.0188
Gender*Province	8	19.6148	0.0119
Province*Race	24	69.5938	<.0001

Finally, we can conclude that using this Survey Logistic method correctly design-adjusts all the estimates in Table 5.7. The fitted model is now more precise.

We next introduce, Ordinal logistic regression. This is a model where the response variable has multiple outcomes that are ordered, for example from weak to strong, or mild to severe (Gould W. , 2000). In our study this would be to model the response ‘Occurrence’ which is the frequency that crime is experienced by one household (for example, one household could be a victim of Motor Vehicle and Household crime which amounts to two types of crimes experienced).

The Occurrence values for our study, is ordered as 0,1,2,3,4,5 with 0 being no crime experienced over the study period and 5 being quite severely affected by all types of crime.

The difference in ordinal regression compared to ordinary logistic regression is the consideration of the probability of an event and all events that are before that event in the ordered listing, instead of the probability of a single event in logistic regression (Norusis, 2015). Ordinal logistic regression are solved using logit models (natural logarithms of odds explained in Equation 5). We could use a multinomial logistic model, but that would not take into account the ordering of the target variable (Benoit, 2012).

The simplest application of Ordinal logit model is the Cumulative Logit Models which can be grouped as the Proportional Odds Model, Non-Proportional Odds Model or the Partial Proportional Odds Model (Ari & Yildiz, 2014).

In the Cumulative Logit model, the event of interest is observing a particular score or less. For the Occurrence of crime type experienced, you model the following odds:

$$\theta_0 = P(0 \text{ types}) / P(\text{greater than } 0 \text{ types})$$

$$\theta_1 = P(0 \text{ or } 1 \text{ types}) / P(\text{greater than } 1 \text{ types})$$

$$\theta_2 = P(0, 1, \text{ or } 2 \text{ types}) / P(\text{greater than } 2 \text{ types})$$

$$\theta_3 = P(0, 1, 2, \text{ or } 3 \text{ types}) / P(\text{greater than } 3 \text{ types})$$

$$\theta_4 = P(0, 1, 2, 3, \text{ or } 4 \text{ types}) / P(\text{greater than } 4 \text{ types})$$

The last category doesn’t have an odds associated with it since the probability of experiencing greater than 5 types of crime would be zero (Norusis, 2015).

Let $\theta_i = P(\text{event} \leq i) / P(\text{event} > i)$, and Y is the response variable with i ($i = 0, 1, 2, 3, 4, 5$) being each ordered category, where X_1 to X_k are the k explanatory variables. This would exist for observations $j = 1, 2, \dots, n$. The following holds for each event for each observation and category:

$$\log(\theta_i) = \log(P(\text{event} \leq i) / P(\text{event} > i)) = \alpha_i - (\beta_1 X_{1j} + \dots + \beta_k X_{kj}), \quad (7)$$

The intercept term α_i denotes the threshold values (intercept values for each logit). Due to the subtraction of the terms $\beta_1 X_{1j} + \dots + \beta_k X_{kj}$, a large coefficient (large magnitude of the parameter estimate) indicates an association with more types of crime experienced.

The Cumulative Logit Model (Proportional Odds Model) works under the assumption of cumulative logit parallelity. This assumption states that the categories of the Target (dependent) variable should be parallel, in other words $P(\text{event} \leq 1) \parallel P(\text{event} \leq 2) \parallel P(\text{event} \leq 3) \parallel P(\text{event} \leq i)$. This assumption implies that the correlation between the independent variables and the dependent variable remains constant for each level of the dependent variable. The Likelihood Ratio Test or the Wald Chi-Square test, tests the null hypothesis below of equality of β_k coefficients of the independent variable for every level of the dependent variable (Ari & Yildiz, 2014).

$$H_o : \beta_{1i} = \beta_{2i} = \dots = \beta_{(k-1)i} = \beta \quad (8)$$

The sign of β is also of importance when we see a positive β value for the categorical variables (Gender, Race, Province, and Income Type), it implies that more occurrences of crime types are more likely compared to the reference category, while a negative β value tells us that fewer crime types are likely to be experienced. For the continuous variable Age a positive β value tells us that as Age increases in years, the likelihood of experiencing more types of crimes increases. An association with more occurrences of crime types means smaller cumulative probabilities for lower occurrences of crime types, since they are less likely to occur.

According to our SAS analysis an Ordinal logistic regression was run having 6 levels in the response. PROC LOGISTIC was used to fit the cumulative logit model. The probabilities modeled are summed over the responses having the lower Ordered Values. The technique used to calculate the parameter estimates is Fisher's scoring.

The score chi-square for testing the proportional odds assumption is 24867.6576, which is not significant with respect to a chi-square distribution with 72 degrees of freedom ($p < 0.0001$). This indicates that the proportional odds model does not adequately fit the data.

This is confirmed by the low ROC value of 57.4% ($c = 0.574$). This is almost a fair accuracy level, so we will continue to use this model.

The Type 3 analysis in Table 5.5, indicates that only 3 factors are of significance in the model, those are Province, Income Type, and Race.

Table 5.8: Ordinal regression type 3 analysis

Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	0.9445	0.3311
Age	1	0.2828	0.5948
Province	8	245.9826	<.0001
Income	5	72.6896	<.0001
Race	3	51.4906	<.0001

Considering these variables in more detail, we refer to Table 5.6. The significant factors for the different types of crime experienced is Province (FS, LP, MP, and WC), Income (Business) and Race (Black, Coloured, and Indian).

Table 5.9: Parallel line assumption test

Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
20.0939	92	1.0000

Using the Wald Chi-Square test for parallelity, we find the p-value in Table 5.7 to be greater than 0.05, so we do not reject the null hypothesis of equality of the coefficients of the independent variables for every category of the dependent variable.

Table 5.10: Ordinal regression parameter estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0	1	0.6622	0.1207	30.0959	<.0001
Intercept	1	1	2.7828	0.1238	505.4363	<.0001

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	2	1	5.3260	0.1616	1086.4634	<.0001
Intercept	3	1	7.5751	0.3542	457.3530	<.0001
Intercept	4	1	8.6740	0.5897	216.3803	<.0001
Gender	Female	1	-0.0285	0.0293	0.9445	0.3311
Age		1	-0.00051	0.000967	0.2828	0.5948
Province	EC	1	-0.0572	0.0553	1.0687	0.3012
Province	FS	1	0.6566	0.0683	92.5080	<.0001
Province	GT	1	0.0124	0.0513	0.0582	0.8094
Province	LP	1	0.4669	0.0609	58.7925	<.0001
Province	MP	1	-0.1809	0.0572	10.0136	0.0016
Province	NC	1	0.0384	0.0725	0.2813	0.5959
Province	NW	1	0.0951	0.0621	2.3449	0.1257
Province	WC	1	-0.1699	0.0617	7.5820	0.0059
Income	Business	1	-0.2839	0.1061	7.1638	0.0074
Income	Farm	1	-0.3590	0.3968	0.8183	0.3657
Income	Other	1	0.1841	0.1055	3.0417	0.0811
Income	Pension	1	0.1891	0.0982	3.7097	0.0541
Income	Salary	1	0.0987	0.0956	1.0678	0.3014
Race	Black	1	0.2941	0.0491	35.8572	<.0001
Race	Coloured	1	0.3760	0.0645	33.9883	<.0001
Race	Indian	1	0.4903	0.1084	20.4695	<.0001

The value given to Indian is 0.4903 (Table 5.8), is positive, this implies that there is a tendency towards more types of crime experienced for that race group compared to the White race group, in other words the White race group experiences fewer types of crime than the Indian race group. The magnitudes of these estimates are Black=0.2941, Coloured=0.3760 and Indian=0.4903, this confirms that the Indian race group experiences the most types of crime, followed by the Coloured race, while the Black race experiences the least number of types of crime, all compared to the White race.

A similar interpretation exists for the other main effects given in Table 5.8. To summarize, Mpumalanga has the lowest number of types of crime occurring and the Free State Province has the highest number of types of crime occurring, in comparison to the Province of KwaZulu-Natal. With Income, Business = -0.2839 (in Table 5.8) stipulates that a lower count of types of crime (or no crime types) are more likely to occur for those earning income from Business compared to those who earn no income, keeping all the other factors constant.

Table 5.11: Odds Ratio Estimates for Ordinal Regression

Effect	Point Estimate	95% Wald	
		Confidence Limits	
Gender Female vs Male	0.972	0.918	1.029
Age	0.999	0.998	1.001
Province EC vs KN	0.944	0.847	1.053
Province FS vs KN	1.928	1.687	2.204
Province GT vs KN	1.012	0.916	1.120
Province LP vs KN	1.595	1.416	1.797
Province MP vs KN	0.834	0.746	0.933
Province NC vs KN	1.039	0.902	1.198
Province NW vs KN	1.100	0.974	1.242
Province WC vs KN	0.844	0.748	0.952
Income Business vs None	0.753	0.612	0.927
Income Farm vs None	0.698	0.321	1.520
Income Other vs None	1.202	0.977	1.478
Income Pension vs None	1.208	0.997	1.465
Income Salary vs None	1.104	0.915	1.331
Race Black vs White	1.342	1.219	1.478
Race Coloured vs White	1.456	1.284	1.653
Race Indian vs White	1.633	1.320	2.019

Considering the values in Table 5.9, the odds ratio estimates help us to understand the association within a factor, say Race for example.

Given in the table are the odds ratio estimates; Race Black vs White = 1.342, from this we can say that controlling for the factors of Age, Gender, Province and Income type, individuals who are of the Black Race group have 34.2% higher odds than individuals who are White of having a response that indicates that they would experience more types of crimes. In other words, the odds of high types of crime versus the combined effect of lower types of crime is 1.342 times higher for Blacks than Whites given all the other variables are held constant.

Similarly for the factors of Gender, Province, and Income.

To summarize the only continuous variable, Age (Age= 0.999). Controlling for the other explanatory variables, 1 additional year in Age is associated with a 99.9% increase in odds of facing higher types of crime relative to lower types of crime.

We use graphical interpretation to explain the interaction effects. The key interaction terms were Province*Gender, Province*Race and Province*Age, all of which included Province (the locating factor).

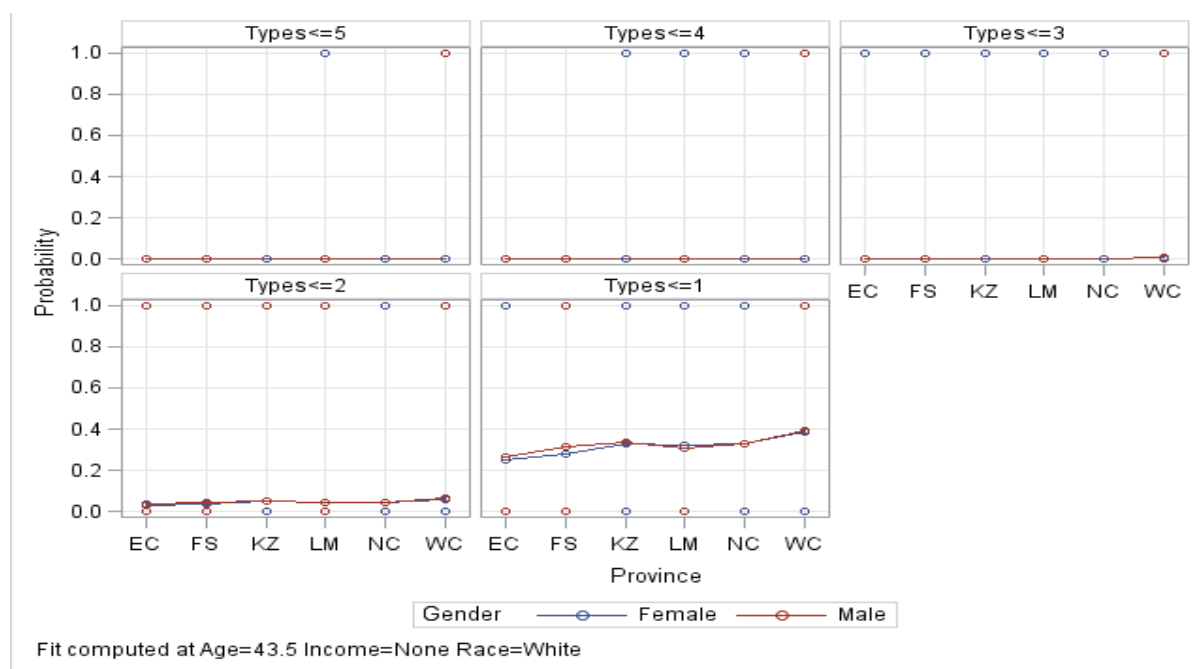


Figure 5.4: Ordinal Province vs Gender interaction

This first interaction shows that Females seem to be experiencing few types of crime in the Eastern Cape (Types <=2) at a higher probability and Males experience few types in Free State (Types <=1) with higher probability than Females. Elsewhere the probabilities are even for both genders and the cumulative event of experiencing 3, 4, or 5 types of crimes has a low and almost zero probability.

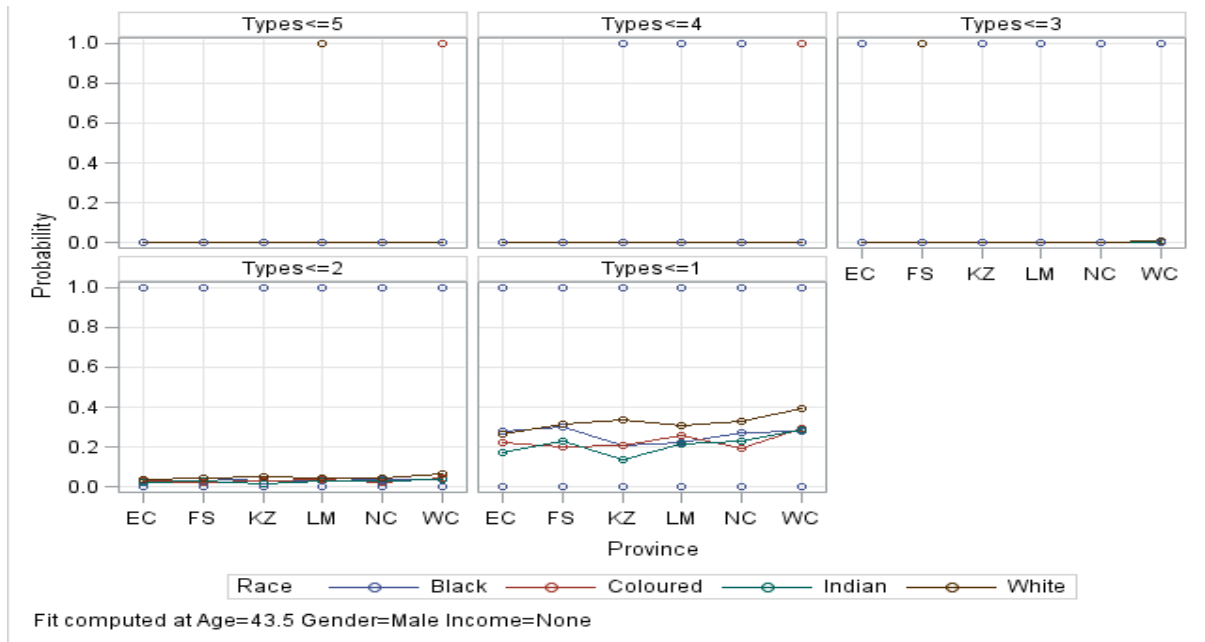


Figure 5.5: Ordinal Province vs Race interaction

In Figure 5.5 we see a clear split where the White race group has highest probability of experiencing less than or equal to one type of crime across all significant Provinces found. Again, we see that more types of crime are less likely to be experienced.

Figure 5.6 focuses on the Age and Province interaction, we notice that on average, 55 year old individuals are more likely to experience one or less types of crime in the Free State province. Younger individuals (on average 27) in the Northern Cape were more likely to experience one or less types of crime. The rare events (zero to none probability) is an individual experiencing three or more types of crime.

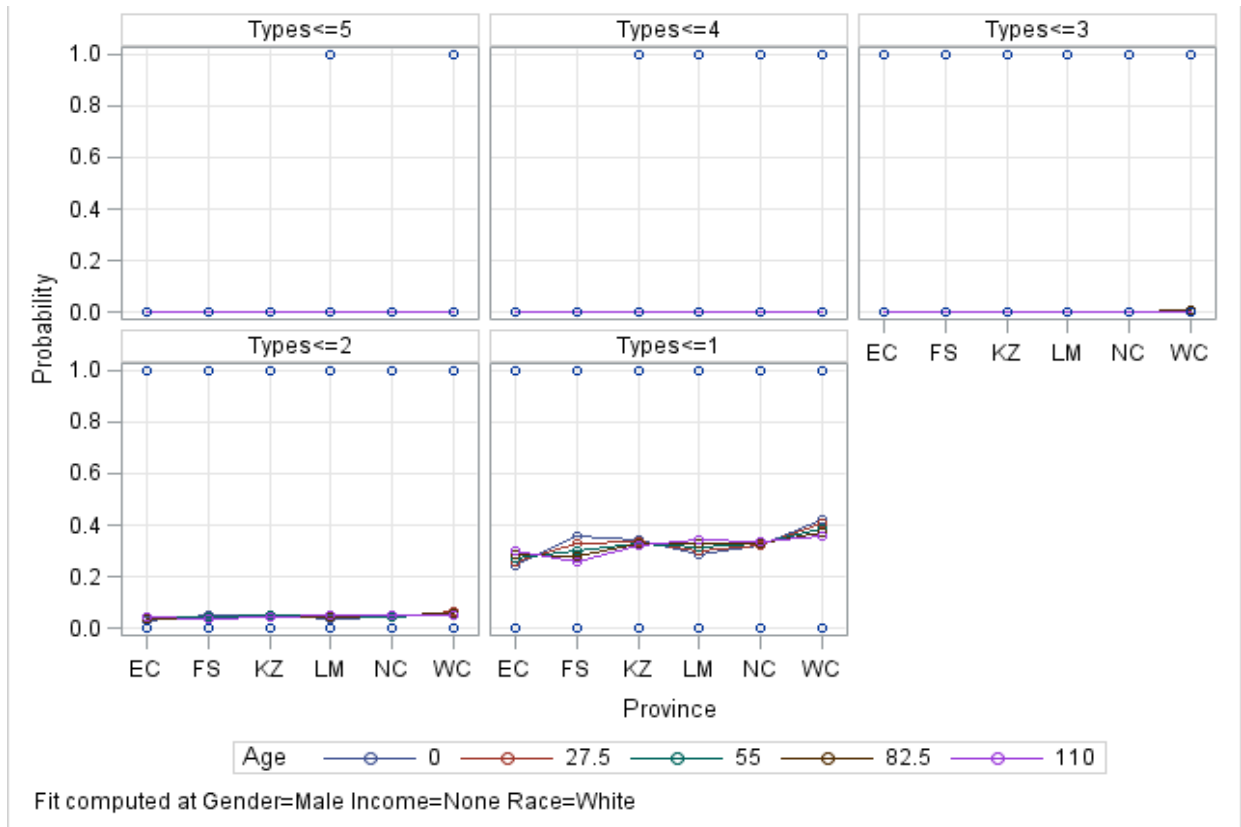


Figure 5.6: Ordinal Province vs Age interaction

In this chapter, we conducted analysis on three different model structures, a logistic model with actual crime incidents as the dependent variable, a logistic model with perceived crime incidents as the dependent variable and an ordinal logistic model with the number of different types of crimes experienced by an individual as the dependent variable.

We extend our statistical models in the next chapters to predict different outcomes. The methodology and theory changes and will be explained accordingly.

6. Generalized Estimating Equations

In this chapter we aim to investigate the effect of our categorical variables on the combined effect of the responses, specifically we aim to investigate whether a correlation exists between the binary (or discrete) responses recorded. The previous chapter considered how these categorical variables effects each response individually, with no assumption of correlation between responses.

We will use Generalized Estimating Equations (GEE's), which falls under the umbrella of the generalized linear model (GLM) analysis. GEE differs from a standard GLM in the sense that the distribution of the response is not fully explained. In essence, GEE is a method of analyzing correlated data, where aside from this existence of a correlation between the responses, a standard GLM approach could be used (Bandyopadhyay, 2011).

Given that the GEE method stems from a GLM, we first focus on a GLM. A generalized linear model (GLM) is broken down into two components, a systematic component and a random component (Johnston, 1996). The systematic component relates to the model form that is, joining the means of the responses to the linear predictors by a link function. The random component relates to the model distribution or the probability distribution from an exponential family. The common distributions from the exponential family are either the Binomial, Poisson, Normal, Gamma, or Inverse Gaussian distributions.

The equation for a GLM that summarizes what these components represent, is given in (9),

$$g(E(\mathbf{Y}_i)) = g(\boldsymbol{\mu}_i) = \mathbf{X}'_i \boldsymbol{\beta} \quad (9)$$

In this equation, g represents the link function, for example, a 'logit' link function, that joins the response variable to the explanatory variables, so that \mathbf{Y}_i represents the dependent measures for $i=1, \dots, n$. It is further known that $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ (individual means) with, \mathbf{X}'_i a vector of explanatory variables (these are the attributing factors, independent variables) for subject i , and $\boldsymbol{\beta}$ a vector of regression parameters to be estimated.

Generalized estimating equations (GEE's) on the other hand, also consist of these two components. In a similar manner let Y_{ij} represent the j^{th} measurement on the i^{th} subject, for $j = 1, \dots, m$ and $i = 1, \dots, n$, with the link function g chosen as a log link (or logit). This can accordingly be represented as

$$g(\mu_{ij}) = \log(\mu_{ij}/(1 - \mu_{ij})) \quad (10)$$

$$v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij}) \quad (11)$$

In addition, equation 10 represents the variance function which characterizes the distribution of the exponential family, as described in the random component.

The final aim of the GEE method is to estimate $\boldsymbol{\beta}$, which is obtained by solving (12)

$$\sum_{i=1}^n \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0 \quad (12)$$

where $\mathbf{Y}_i = [Y_{i1}, \dots, Y_{ij}]'$, $\boldsymbol{\mu}_i = [\mu_{i1}, \dots, \mu_{ij}]'$ and \mathbf{V}_i represents the covariance matrix of \mathbf{Y}_i . This covariance matrix is calculated as follows

$$\mathbf{V}_i = \phi \mathbf{A}^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}^{\frac{1}{2}} \quad (13)$$

where \mathbf{A} is a square matrix, with $v(\mu_{ij})$ in the main diagonal. This brings us to the working correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$, which is estimated using values of the vector $\boldsymbol{\beta}$. There are different structural choices for the working correlation matrix, namely exchangeable, AR(1), stationary m-dependent (or Toeplitz), and unspecified (or unstructured), which is the most efficient (GEE for Longitudinal Data, 2015).

GEE's have consistent and asymptotically normal solutions, even with miss-specification of the correlation structure. Other advantages of using a GEE analysis includes the fact that the GEE

- yields both robust and model-based standard errors for the parameter estimates,
- computes solutions for all kinds of outcomes, for non-normal outcomes,
- provides population-averaged (or marginal) estimates of $\boldsymbol{\beta}$.

It is worth noting however that the GEE assumptions are more stringent if there is regarding missing data (Johnston, 1996).

In this study, General Estimating Equation (GEE) analysis is used to model the correlation between the multiple response variables that is the types of crime, where all the responses are binary in this study.

The predicted model in this study takes the form given in (14).

$$\begin{aligned} & \text{MURDER | VEHICLE | HOUSEHOLD | AGRICULTURE | OTHER} \\ & = \beta_0 + \beta_1 * \text{Gender} + \beta_2 * \text{Age} + \beta_3 * \text{Province} + \beta_4 * \text{Income} + \beta_5 * \text{Race} + \\ & \beta_6 * \text{Gender} * \text{Province} + \beta_7 * \text{Age} * \text{Province} + \beta_8 * \text{Province} * \text{Race} \end{aligned} \quad (14)$$

An analysis of this study as per the output by SAS, under the general modelling (GENMOD) procedure, was followed. The probability of each category of crime is modeled as a logistic regression model, using explanatory variables Age (continuous), Gender, Province, Income and Race (as categorical), along with the interaction terms Gender*Province, Age*Province and Province*Race. The crime data entries for this study included 25605×5 (128025) data lines, due to the “repeated households”. SAS clustered the results by a specific household and finally 25605 data lines resulted. We thus modeled the 5 categories of crime as the response variables and investigated the existence of a correlation between them. The Unstructured Correlation structure was used. Recall from (13) that there are several choices for the working correlation matrix the choice in this study was the generic, i.e. when the exact correlation cannot be determined.

According to (14), we identify parameter estimates by each β for the respective parameters (variables), these values are calculated using software (SAS). Using these values, we complete our model defined in (14), noting that some parameter estimates are vectors and some scalars, depending on how many categories the parameter has. When we consider a parameter estimate, we first investigate the significance thereof for this we use a p-value approach and only once the parameter estimate is found to be significant, do we investigate the magnitude and sign of the estimate and how these effect the interpretation of our study model.

The p-value approach is used to test the significance of a parameter, while we use hypothesis testing to explain this concept. Let the null hypothesis be $H_0: \beta = \mathbf{0}$, which is tested against the alternative $H_a: \beta \neq \mathbf{0}$. If the p-value is less than 0.05 (we use 5% as the norm), we reject H_0 and conclude with 95% certainty that the parameter is significant, while the converse is also true.

The parameter estimates computed by the GEE procedure can be positive or negative valued. The interpretation of parameter estimates are always done as a comparison, where we make sense of the values by comparing them to a reference parameter (here the parameter estimate is zero), as illustrated in Table 6.1. The negative value then implies a tendency towards the reference category and a positive value implies a tendency toward that parameter itself, while the magnitude of the parameter estimate shows the strength of that tendency. We use examples from Table 6.1 to explain this relationship.

Using the p-value approach explained above, we illustrate only the significant parameter estimates and their odds ratio as given in Table 6.1 (note that the insignificant variables were omitted).

Table 6.1: Significant Parameter Estimates

Parameter			Estimate	Odds Ratio	P-value
Intercept			-3.1871	0.0412914	<.0001
Gender	Female		0	1	
Gender	Male		-0.1675	0.8457766	0.0075
Age	All		0.0065	1.0065212	0.0008
Province	KZN		0	1	
Province	EC		0.5804	1.786753	0.0252
Province	GT		1.084	2.9564819	<.0001
Province	MP		1.2364	3.4431956	<.0001
Province	NC		0.6737	1.9614814	0.0184
Province	NW		0.7775	2.1760254	0.0137
Province	WC		1.0186	2.769315	<.0001
Income	No Income		0	1	
Income	Business		0.229	1.257342	0.0105
Income	Pension		-0.165	0.8478937	0.0491
Race	White		0	1	
Race	Black		0.4066	1.5017033	0.0091
Gender*Province	Female	KZN	0	1	
Gender*Province	Male	LP	0.4032	1.4966062	0.0001
Age*Province	All	KZN	0	1	
Age*Province	All	FS	-0.0112	0.9888625	0.0031
Age*Province	All	GT	-0.0061	0.9939186	0.0232
Age*Province	All	MP	-0.0092	0.9908422	0.0013
Age*Province	All	NW	-0.0098	0.9902479	0.0022
Age*Province	All	WC	-0.009	0.9910404	0.0018
Province*Race	KZN	White	0	1	
Province*Race	FS	Black	-0.6848	0.5041911	0.0033
Province*Race	GT	Black	-1.0899	0.3362501	<.0001
Province*Race	LP	Black	-0.8126	0.4437029	0.0065
Province*Race	MP	Black	-0.7653	0.4651944	0.0002
Province*Race	NC	Black	-0.5259	0.5910232	0.021
Province*Race	WC	Black	-0.6939	0.4996237	0.0002

Table 6.1 uses GEE methodology to model the predicted probability of being a victim of crime (any of the five categories of crime). Consider the parameter Gender (Male) with the estimated value of -0.1675, which indicates that the predicted probability of experiencing an incident of crime is higher for females (the reference

group) than males, since the negative value explains the tendency towards the reference category. The odds for this event is 0.8458, which means that the odds of a female being a victim is 84.58% higher than for males.

The similar conclusion applies to the Province parameter. All significant variables have positive parameter estimates, indicating that the predicted probability of experiencing an incident of crime is higher in these Provinces than the Province of KwaZulu-Natal.

The entries in Table 6.1 includes positive and negative estimates for income, which can be interpreted as meaning that the predicted probability of experiencing an incident of crime is higher for those with income from business as opposed to those that have no income, and their predicted probability of experiencing an incident of crime is higher for those with no income, as opposed to that that receive a pension. Similarly for ethnic group, the predicted probability of experiencing an incident of crime is higher for those from the Black race group than for Whites.

When considering an interaction of parameters, we note for example that in the case of Gender*Province; illustrated in Figure 6.1, the predicted probability of experiencing an incident of crime is higher for Male subjects from Limpopo than females (parameter estimate of 0.4032 confirms this). Females have a higher probability of being a victim of crime than males, in the Province of Mpumalanga.

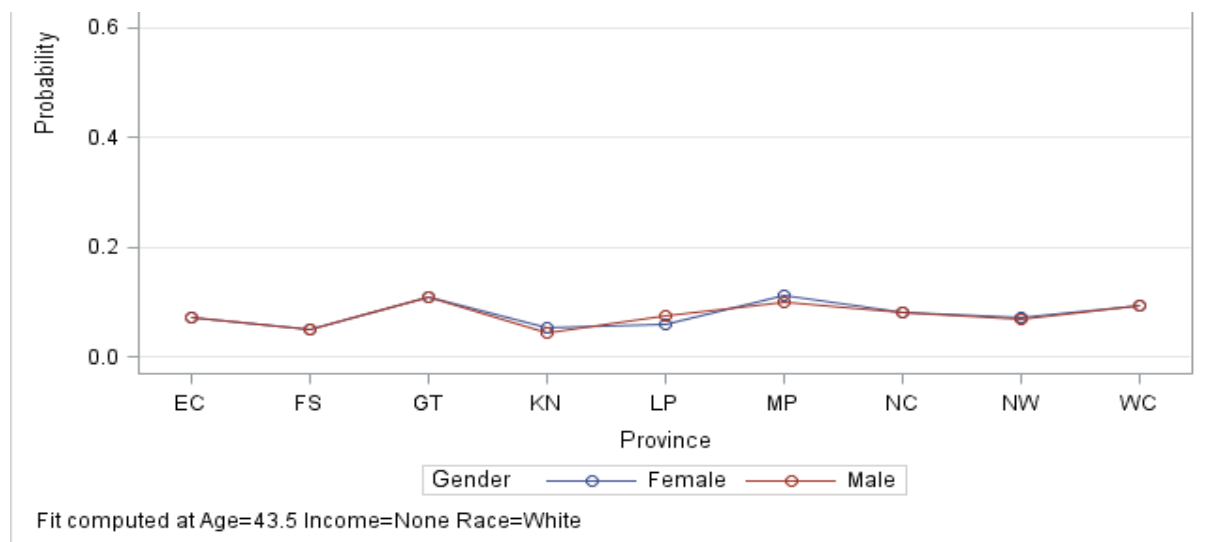


Figure 6.1: Province vs Gender interaction

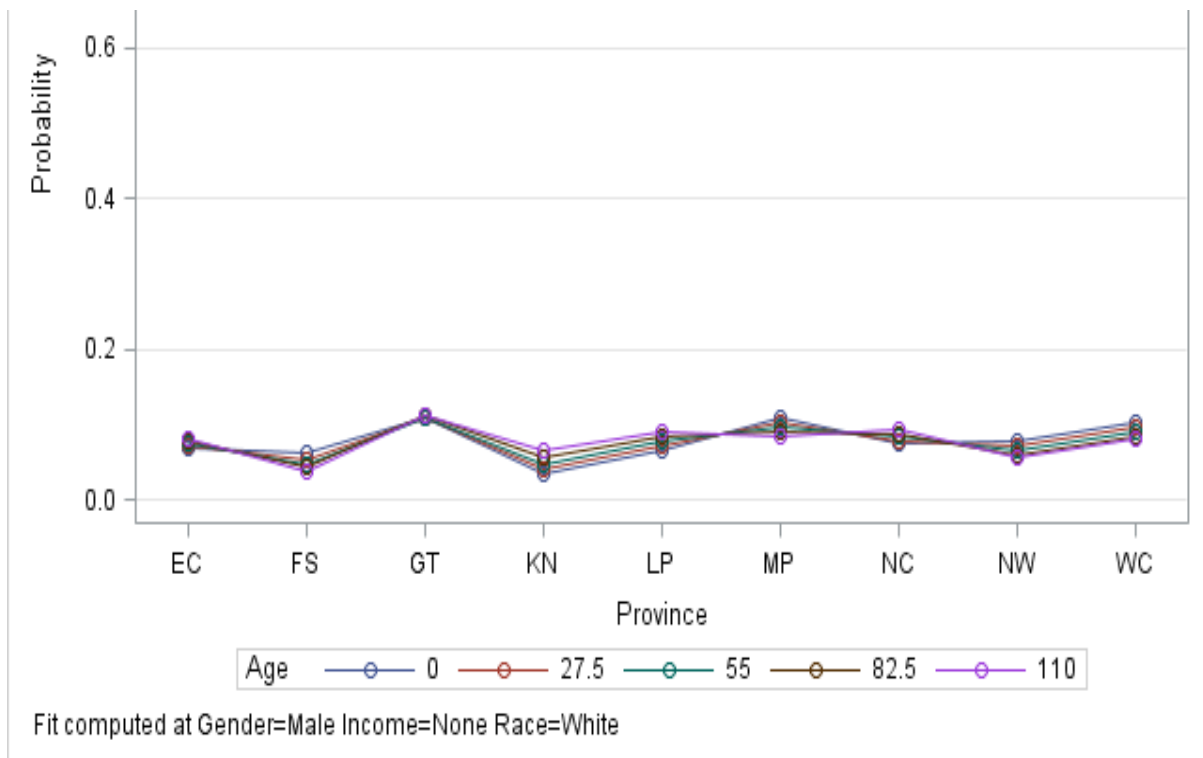


Figure 6.2: Province vs Age interaction

For the Age*Province interaction, we first see that the predicted probability of experiencing an incident of crime is the same for all households irrespective of their age in Gauteng. We notice that the younger aged individuals are more likely to be affected by crime in Mpumalanga and Eastern Cape, whereas the older victims are predicted more highly to be victims of crime in the remaining provinces.

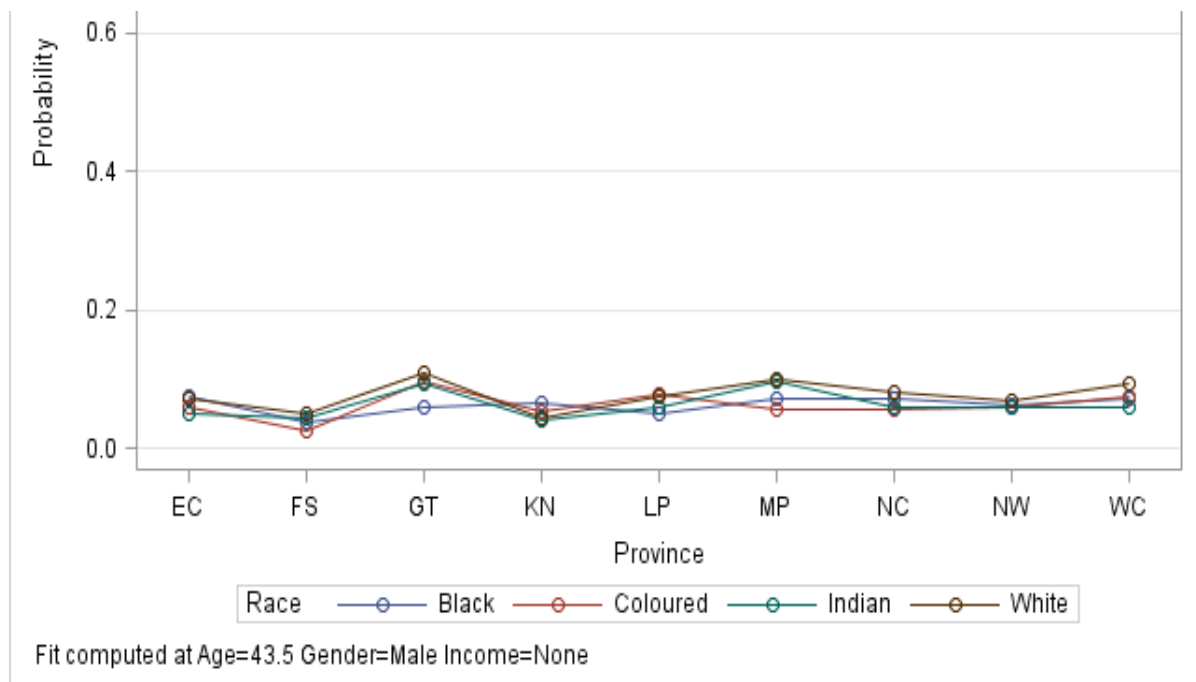


Figure 6.3: Province vs Race interaction

From focusing on the Province*Race interaction, we note that the Black race group have the highest probability of being a victim of crime in KwaZulu-Natal, with the Indian group having the highest chance of being victims of crime in the Eastern Cape and Mpumalanga, with the White race group recording the highest chance of being a victim to crime as compared to other races in the remaining provinces.

Table 6.2: Correlation matrix for categories of crime

	Murder	Vehicle	House	Agric	Other
Murder	1.0000	0.0041	-0.0986	0.0280	0.0245
Vehicle	0.0041	1.0000	0.1623	0.0060	0.0472
House	-0.0986	0.1623	1.0000	-0.0404	0.0026
Agric	0.0280	0.0060	-0.0404	1.0000	0.0187
Other	0.0245	0.0472	0.0026	0.0187	1.0000

The working correlation matrix, discussed earlier in this chapter, accounts for the relationship between the response variables (binary). The symmetric matrix quantifies the strength of the correlation between the response variables.

Comparing the magnitudes of the entries of the matrix in Table 6.2, we see that there exists a correlation between Vehicle theft and Household crime (0.1623) though this value is small, it is the highest off diagonal element, so that we can deduce that the strongest correlation existing between any two response variables, is between Vehicle theft and Household crime. These two types of crime are accordingly more strongly associated with each other than is the case for occurrence of any other two crime types listed.

In this chapter we used a Generalized Estimating Equation approach to investigate the different effects that the attributes within each variable have on the combined outcome of crime. In the same manner we investigated the interaction variables and interpreted the parameter estimates. We further investigated each interaction by including the Province variable, which aided in locating crime. We also considered the relationship between the different categories of crime.

We next extend our study to visually locate crime, by using spatial mapping methodology.

7. Spatial Analysis

In this study we have analyzed crime using statistical models and representing relationships, using graphs. We next attempt to visually locate crime around the country.

We focus on how a certain location may be a factor of crime, so that we may illustrate the relationship between criminal behavior and those who reside in that area (who are at risk of experiencing that crime).

Hot spot mapping (finding crime hot spots) is the visual methodology that we intend to use, where in addition to visual representations, we provide some statistical estimates in explaining the occurrence of crime. We bear in mind that graphical output alone will not ensure that proper interpretation can be obtained (Anselin, Luc; Cohen, Jacqueline; Cook, David; Gorr, Wilpen; Tita, George;, 2000).

The GLM (generalized linear model) was fully explained in the previous chapter, we extended this methodology by introducing a generalized linear mixed model (GLMM). As explained, the GLM assumes the vector of observations \mathbf{y} to be uncorrelated, while GLMM on the other hand, assumes the observations \mathbf{y} to be Normally distributed and having a spatial correlation structure.

GLMM is a class of models which combines GLM and mixed models, this allows us to cater for scenarios where observations are repeated, based on some group. For example we investigate the scenario where the response is the frequency of crime (count), where there are 9 categories of crime (repeated measure) in each police station (Bolker, 2013).

The SAS statistical procedure that we use in this chapter, is GLIMMIX. The GLIMMIX procedure fits statistical models (GLMM) to data with correlations or non-constant variability and where the response is not necessarily normally distributed (SAS/STAT(R) 9.2 User's Guide, Second Edition, 2010).

We use the same source to explain the basic model structure that links GLIMMIX and GLMM. We let \mathbf{Y} represent the vector of observations and $\mathbf{X}\boldsymbol{\beta}$ the fixed effects, with \mathbf{X} the design matrix and $\boldsymbol{\beta}$ remaining to be the matrix of independent variables, $\mathbf{Z}\boldsymbol{\alpha}$ is the random effect, \mathbf{Z} is the design matrix and $\boldsymbol{\alpha}$ defines the distribution of s (s is computed for any location s_i), so that this random effect incorporates the spatial aspect into the GLMM model (Ayele, Zewotir, & Mwambi, 2013).

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} \quad (15)$$

The basic model structure for the GLMM is given in (15), $\mathbf{X}\boldsymbol{\beta}$ being the fixed effects and $\mathbf{Z}\boldsymbol{\alpha}$ being the random effects.

Estimation methods for the fitted model include maximum likelihood, generalized estimating equations and the penalized quasi-likelihood method to name a few (McCulloch, 1997).

We expand on each of these briefly. The maximum likelihood (ML) equation uses the integral (with respect to the dimension of \mathbf{Z}) to find the maximum value of the likelihood, the likelihood equation involves Equation 15, in an attempt to solve for $\boldsymbol{\beta}$. ML estimation and likelihood ratio tests can however be quite complex to compute for many GLMMs.

The generalized estimating equations method is fully explained in Chapter 6, the fitted model uses Equation 11 to estimate the parameters $\boldsymbol{\beta}$. We used this method in our research.

The penalized quasi-likelihood method is similar to results from the Laplace approximation. Consider Equation 15, let $\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\varepsilon}$, we have the below transformation:

$$g(\mathbf{y}) \approx g(\boldsymbol{\mu}) + (\mathbf{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \mathbf{z} \quad (16)$$

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}g'(\boldsymbol{\mu}) \quad (17)$$

The aim is to both find the estimate of $\boldsymbol{\beta}$ as well as the best linear unbiased predictor of $\boldsymbol{\varepsilon}$, the error term. This approach is computationally easier however does not work well with non-normal data (binary data), (McCulloch, 1997).

The distribution of $\boldsymbol{\alpha}$, which includes the spatial effect, is a Gaussian distribution, i.e. $\boldsymbol{\alpha} \sim \text{Gau}(0, \Sigma_{\boldsymbol{\alpha}}(\boldsymbol{\theta}))$ and the spatial correlation is parameterized by $\boldsymbol{\theta}$ in $\Sigma_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$. The error term of the model in Equation 15, $\boldsymbol{\varepsilon} \sim \text{Gau}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I})$ accommodates for over-dispersion. We assume that $y(s_i|\boldsymbol{\alpha})$ is conditionally independent for any location s_i , with conditional mean $E[y(s_i)|\boldsymbol{\alpha}] = \mu(s_i)$.

Geostatistics uses three core functions to describe spatial correlation, these are the correlogram, covariance and the semi-variogram. The semi-variogram has a nugget, sill (or scale) and range represented as in Figure 7.1. We can define our variogram as in (18).

$$\text{Semi-variogram} = \gamma(s_1 - s_2) = \frac{1}{2} \text{var}[Z(s_1) - Z(s_2)] \quad (18)$$

We assume $\mu(s)$ to be constant, and the semi-variogram to be a function $\gamma(\cdot)$. We define the semivariance $\gamma_o(t)$, lag distance class t , nugget variance $c_o \geq 0$, structural variance $c_1 \geq c_o$ and the range parameter R . The spatial covariance structures are spherical, exponential, power and gaussian (to name a few). We consider the exponential form

$$\gamma_o(t) = \begin{cases} 0 & \text{if } t = 0 \\ c_o + c_1 \left(1 - e^{-t/R}\right) & \text{if } t > 0 \end{cases} \quad (19)$$

Semivariogram

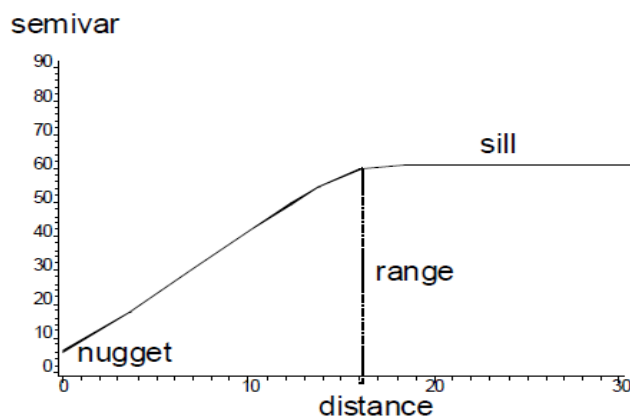


Figure 7.1: Semi-variogram

The model we select for the semi-variogram, explained above, is used in kriging (spatial prediction). Kriging, which we will use the SAS software to compute, aids in predicting our outcome (frequency of crime) at locations that are not in our sample, using our sampled locations.

We introduce the Least Squares means algorithm, where LS-means are predicted population margins. The result would be an estimation of the summation of the means, where the means are based on the model (linear) used. Each LS-mean is represented by $\mathbf{L}\hat{\boldsymbol{\beta}}$, where \mathbf{L} is the design matrix and $\hat{\boldsymbol{\beta}}$ is the estimated values of the independent variable parameters (fixed effects). The variance is computed as $\mathbf{L}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\mathbf{L}'$, where the variance matrix depends on the estimation method (SAS Knowledge Base, 2010).

To represent differences in the LS means visually we use the diffogram, which is a plot of lines at 45° angles (rotated anti-clockwise) where the co-ordinates of the midpoint of these lines represent the respective LS-means for the two independent variables being compared. The center of the line co-ordinates (x, y) is computed as $x = \min\{\hat{\eta}_i, \hat{\eta}_j\}$ and $y = \max\{\hat{\eta}_i, \hat{\eta}_j\}$, where $\hat{\eta}_i$ and $\hat{\eta}_j$ denote the i^{th} and j^{th} LS-mean for the two independent variables being compared (High, 2011).

After explaining the theory, we consider our data for this chapter, as explained in Chapter 3 of this study, will contain the GPS co-ordinates for the 1140 police stations across the country, each reporting of 9 different types of crime (detailed reports of crime were categorized into simply 9 types as per data cleaning).

Viewing our results in geographic context adds a new level of understanding to the results. We first illustrate our findings using maps and plots before we use our statistical approach.

Using the iNZight software, developed at the Auckland University (Wild, 2016), we visually compare the different police station according to the incidents of crime experienced. Figure 7.2 shows that the trend is positively skewed (the tail is to the right), implying that most of the police stations experience low incidents of crime. The purple dots present throughout the figure reveal that Robbery (of possessions) occur most frequently across the country. In an attempt to locate the highest occurrences of reported crime, we find that Cape Town Central Station records among the highest, followed by Mitchells Plain, Johannesburg Central and Durban Central areas. The Provinces that we locate as the higher crime areas are Western Cape, Gauteng and Kwa Zulu-Natal.

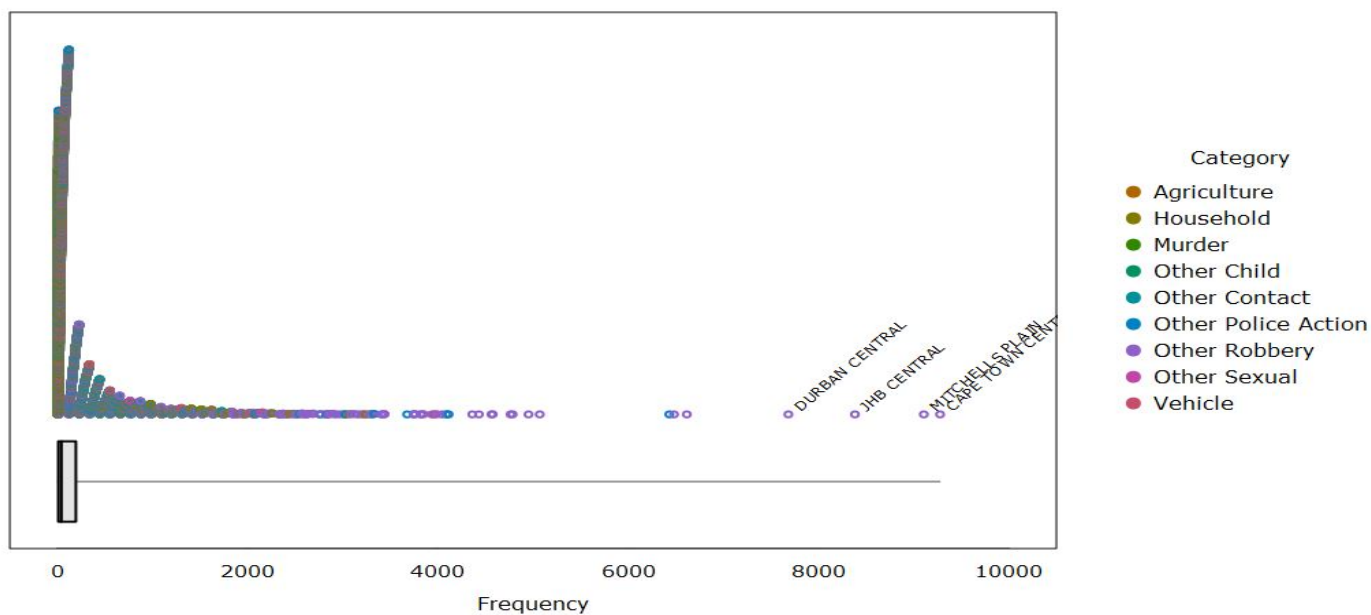


Figure 7.2: iNZight plot of police stations

We explore this data further at a province level. We illustrate the occurrences of crime for each category, within each province. We use QGIS as demonstrated by Stats SA representatives during a dissemination training on mapping tools, as visual aid in this regard.

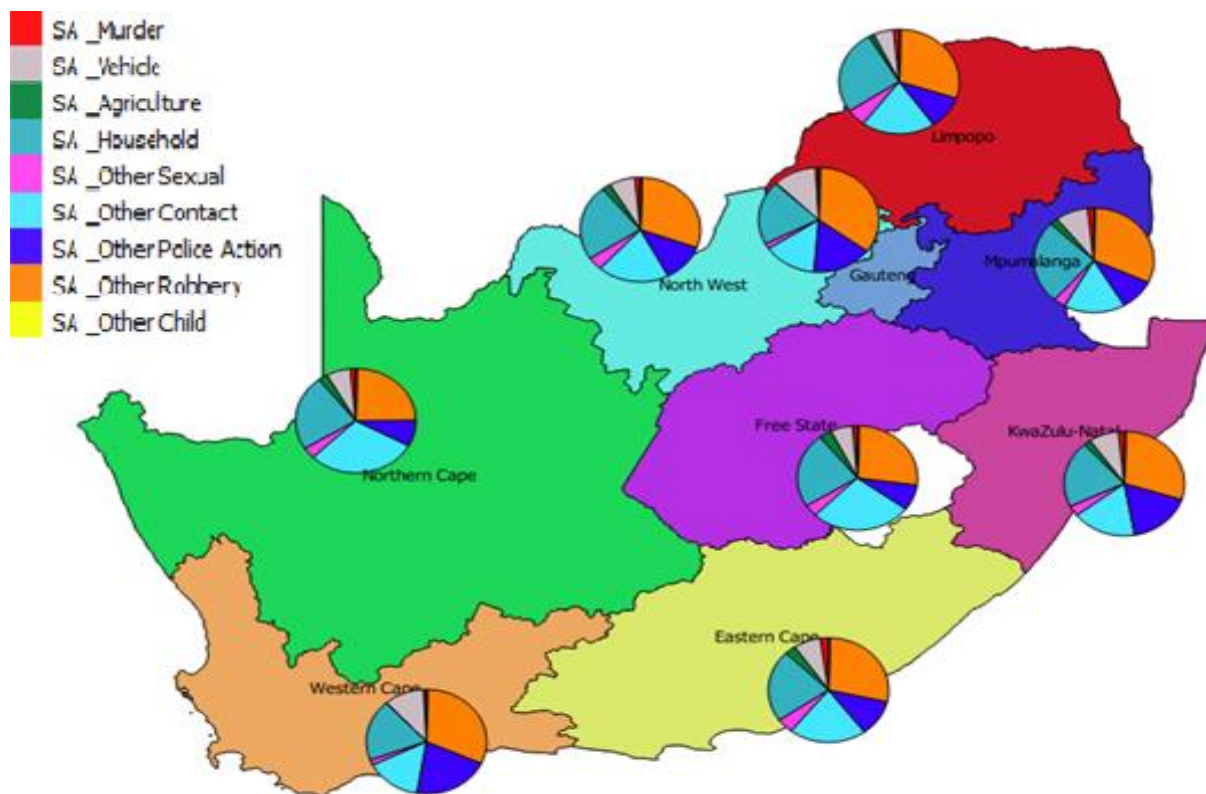


Figure 7.3: Spread of crime per province

It is evident that robbery is most frequently occurring crime type across the country, which includes the categories of common or aggravated robbery, commercial crime, non-residential robbery and all theft. Other contact crime types are also present, for example public violence, assault and even attempted murder. Household crime, which is burglary of homes, is among the mentioned crimes that occur often in South Africa, across all provinces.

We also determine the types of crime that occur more in some particular provinces than the others. Figure 7.3 bears evidence that crime detected as a result of police action (drug related, driving under the influence or unlawful possession) occurs highly in the provinces of the Western Cape, KwaZulu-Natal and Gauteng. Agricultural crime (theft of crops and livestock) are almost not present in the provinces of the Western Cape, Limpopo and Gauteng, while it is the highest in the Free State.

We however need be mindful to present our results of occurrences and should rather draw conclusions based on population totals for those areas, as this would reveal a more accurate result. We understand that our data provides only crime reported to the police station, by considering how these reports compare to the population of those areas would be more ideal and yield calculated crime rates per Province, as given below.

Table 7.1: Crime rates of province by crime category

Prov- ince	Mur- der	Vehicle	Agricul- ture	House- hold	Other Sexual	Other Con- tact	Other Police Action	Other Robbery	Other Child	Over- all
KZN	0.054%	0.272%	0.065%	0.691%	0.109%	0.582%	0.578%	0.995%	0.010%	3.355%
LIM	0.032%	0.118%	0.030%	0.553%	0.112%	0.421%	0.216%	0.633%	0.007%	2.122%
MP	0.047%	0.250%	0.060%	0.728%	0.092%	0.465%	0.268%	0.888%	0.008%	2.806%
NW	0.046%	0.226%	0.073%	0.747%	0.131%	0.613%	0.371%	0.961%	0.014%	3.181%
NC	0.062%	0.281%	0.107%	0.996%	0.148%	1.263%	0.345%	1.046%	0.010%	4.258%
WC	0.063%	0.857%	0.014%	1.505%	0.130%	1.200%	1.643%	2.505%	0.017%	7.935%
EC	0.068%	0.232%	0.094%	0.698%	0.143%	0.656%	0.351%	0.890%	0.009%	3.141%
FS	0.061%	0.286%	0.158%	1.095%	0.171%	1.283%	0.371%	1.260%	0.022%	4.706%
GT	0.044%	0.582%	0.007%	0.975%	0.083%	0.712%	0.792%	1.718%	0.015%	4.928%

These statistics show that overall, Western Cape experienced the highest potential probability of crime, that is 7.9% of the total population on average reported an incident of crime to the police. The respective crime rates are shown for other categories of crime, with the trend leaning towards Western Cape having the highest crime rates across most of the categories. Reasonably low overall crime rates can be seen for the Mpumalanga and Limpopo provinces with 2 in every 100 people on average, to have reported that they have experienced some type of crime.

Based on the raw data, we have provided some descriptive results that will further be explored in this chapter. We will now use the statistical approach that we have explained at the beginning of this chapter.

We modelled the data, that includes repeated Police Station names, for the nine different categories of crime, according to the GLIMMIX principles. The data has names of the different types of crimes with latitude and longitude attached, which we use as a measurement for mapping. The response variable is Frequency of each crime category reported per police station, so that the distribution that the model uses is Spatial Exponential.

Using the p-value approach we obtain the output as in Table 7.2, similar to other chapters, where now we test the null hypothesis that there is no difference between the categories of crime. The reference crime category here is Vehicle theft (default). We notice that all the p-values are less than 0.0001, this falls within the rejection region of the null hypothesis, thus supporting the conjecture that there is a significant difference between all the other categories of crime compared to the Vehicle theft Category.

Table 7.2: GLIMMIX model output

Effect	Category	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		187.96	13.1807	1139	14.26	<.0001
Category	Agriculture	-163.09	15.2368	9112	-10.70	<.0001

Effect	Category	Estimate	Standard Error	DF	t Value	Pr > t
Category	Household	232.48	15.2368	9112	15.26	<.0001
Category	Murder	-162.94	15.2368	9112	-10.69	<.0001
Category	Other Child	-181.98	15.2368	9112	-11.94	<.0001
Category	Other Contact	157.49	15.2368	9112	10.34	<.0001
Category	Other Police Action	112.67	15.2368	9112	7.39	<.0001
Category	Other Robbery	434.29	15.2368	9112	28.50	<.0001
Category	Other Sexual	-132.97	15.2368	9112	-8.73	<.0001
Category	Vehicle	0

We use the same GLIMMIX procedure for the data but explore a different technique, the Least Squares Means (LSM) technique. The derivation of the LSM was explained at the beginning of this chapter. The output in Table 7.3 shows that Murder, Agricultural crime and other child related crimes are not significant in the model (the p-values are larger than 0.05). The estimates in the table are the calculated values for the least-square mean of that corresponding category of crime.

Table 7.3: LSM model estimates output

Category	Estimate	Standard Error	DF	t Value	Pr > t
Murder	25.0132	13.1807	10251	1.90	0.0578
Vehicle	187.96	13.1807	10251	14.26	<.0001
Agriculture	24.8623	13.1807	10251	1.89	0.0593
Household	420.44	13.1807	10251	31.90	<.0001
Other Sexual	54.9842	13.1807	10251	4.17	<.0001
Other Contact	345.45	13.1807	10251	26.21	<.0001
Other Police Action	300.63	13.1807	10251	22.81	<.0001
Other Robbery	622.25	13.1807	10251	47.21	<.0001
Other Child	5.9798	13.1807	10251	0.45	0.6501

Table 7.3 corresponds to Figure 7.4. The coordinates (x, y) of the midpoint for each line corresponds to the two least-square means being compared. For example, for the comparison of Categories “Other Police action” and “Other Contact” crime, the respective estimates of their LS means are 300.63 and 345.45 respectively (this was taken from Table 7.3). The center of the line segment (midpoint) for Police

action and Contact crime is placed at (300.63; 345.45). Lines associated with significant comparisons do not touch or cross the reference line (broken line). The blue line present for Contact crime and crime as a result of police action, represents a significant comparison between these categories, but shows that they are unrelated.

On the contrary, according to the output, Murder and crime related to Children (i.e. kidnapping), as well as Murder and Sexual related crimes result in insignificant comparisons (you cannot compare one with the other - indistinguishable), so that these two types of crime are significantly related.

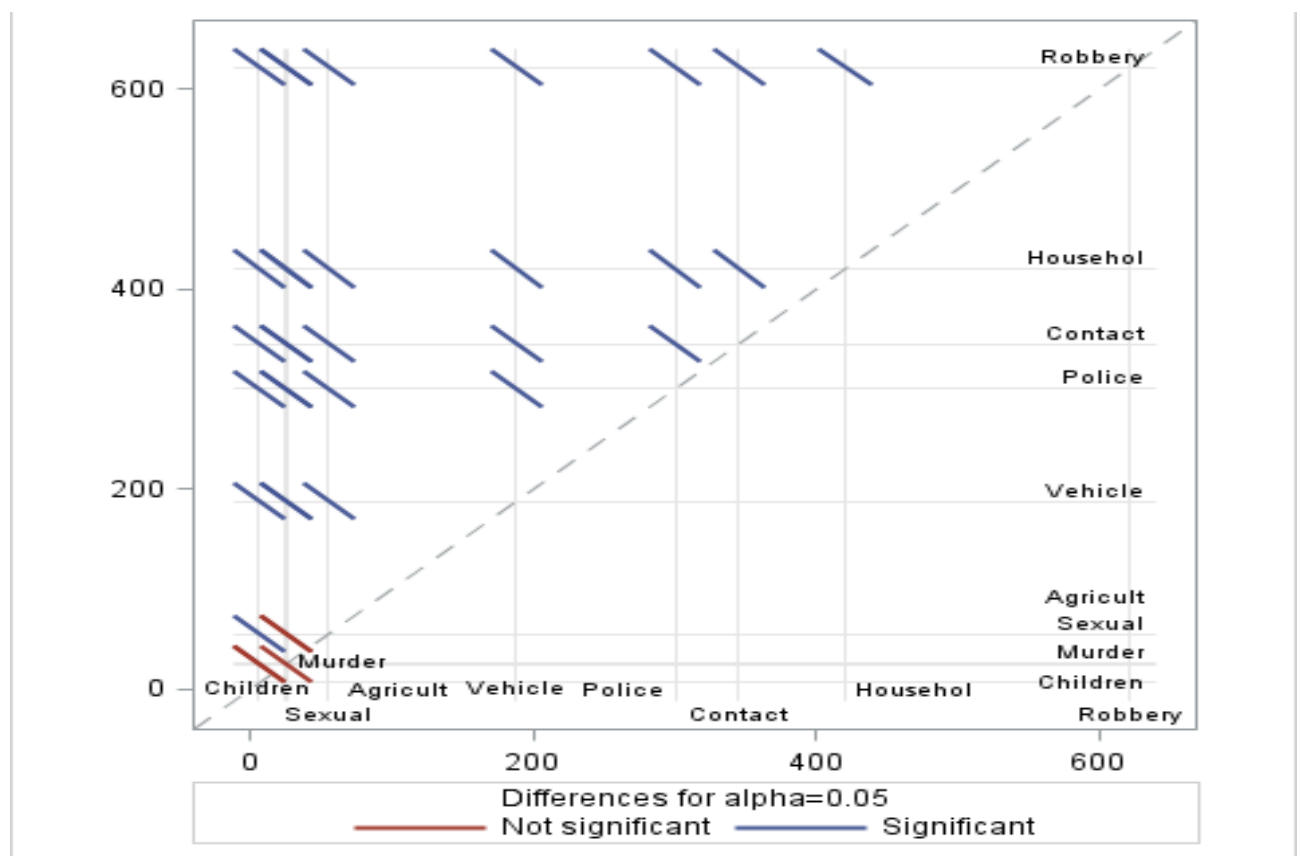


Figure 7.4: LSM comparison of categories of crime

We next explore this comparison further by considering Murder to be a fixed category. Table 7.4 presents the Differences of Category the LS Means output. Using the p-value approach once again, we find that those categories of crime that have p-values that exceed 0.05, support the null hypothesis that there is no difference between those categories (LS mean for Murder = LS mean for another crime category).

It is accordingly confirmed that Murder and Agriculture, Murder and Sexual crime, and Murder and Child related crime, are all related at a 5% level of significance.

In other words, we could say for example that when a child is kidnapped, statistics show that this would also quite possibly end in murder.

Table 7.4: Differences of Category least squares means (Dunnett comparison)

Category	Category	Estimate	Standard Error	DF	t Value	Pr > t	Adj P
Vehicle	Murder	162.94	18.6404	10251	8.74	<.0001	<.0001
Agriculture	Murder	-0.1509	18.6404	10251	-0.01	0.9935	1.0000
Household	Murder	395.43	18.6404	10251	21.21	<.0001	<.0001
Other Sexual	Murder	29.9711	18.6404	10251	1.61	0.1079	0.4609
Other Contact	Murder	320.44	18.6404	10251	17.19	<.0001	<.0001
Other Police Action	Murder	275.61	18.6404	10251	14.79	<.0001	<.0001
Other Robbery	Murder	597.24	18.6404	10251	32.04	<.0001	<.0001
Other Child	Murder	-19.0333	18.6404	10251	-1.02	0.3072	0.8755

Table 7.4 can further be illustrated as given in Figure 7.5, where the vertical lines that lie within the shaded band (between the Lower decision line and upper decision line), illustrate those categories that are related to Murder (control). The results given here further confirm all that was provided above (the association of murder with the other crime types).

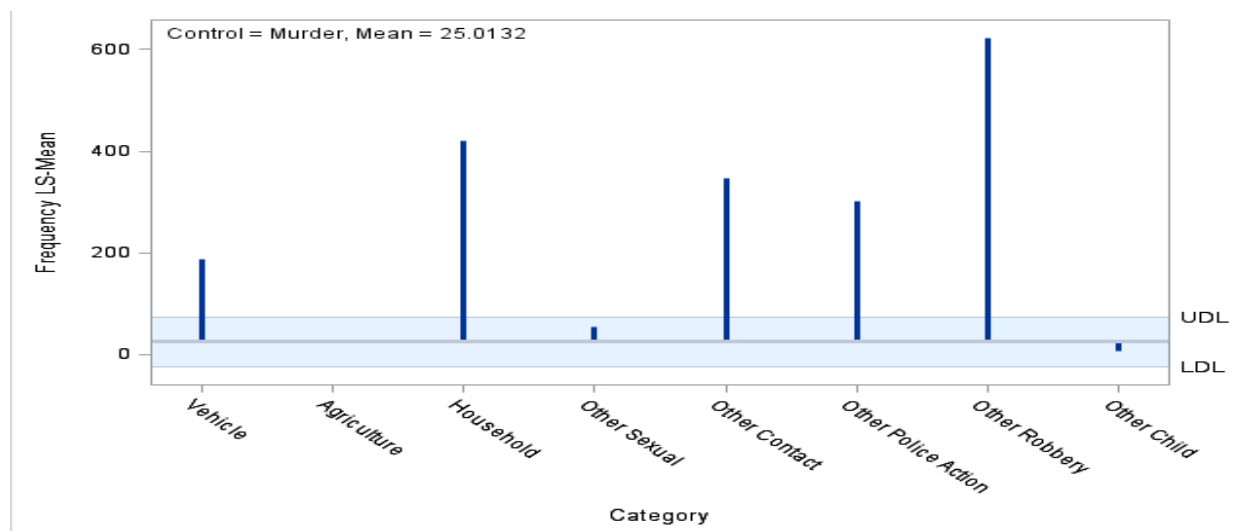


Figure 7.5: LSM graph comparing murder to all categories of crime

We next consider using the mixed procedure (proc mixed), to perform a lattice analysis. The intention of a lattice analysis is to reduce experimental error and to increase precision. We alter the data structure for this analysis to include groups, blocks and treatments.

The data used in this study represents the group variable as the 9 provinces, where within each province there are blocks taken as the police station (the block size for each province is not equal), and within each block (police station) there are 9 categories of crime recorded (the treatments).

The linear model for a lattice design has the format expressed in (20).

$$Y_{ijl} = \mu + \tau_i + \gamma_j + \rho_{l(j)} + \varepsilon_{ijl} \quad (20)$$

We let Y represent the frequency of crime that has occurred, with μ the mean frequency of crime, τ the treatment effect (1 to i) in our study ranging from 1 to 9, γ_j the group effect (our Provinces) extending from 1 to j (provinces 1 to 9), the block effects are represented by $\rho_{l(j)}$, where these police stations range from 1 to l and for each province 1 to j , we further note that l changes for each province, while ε_{ijl} denotes the random error of the linear model for the lattice design.

The mixed procedure was conducted using SAS, to analyze the lattice design. The results obtained were first the parameter estimates for the prediction model expressed in (20).

We consider the block variables (Police Stations) that were significant in the model (p value <0.05), of those Police Stations that weighed a significance in the model. We note that Mitchells plain in the Western Cape had the greatest positive impact on increasing the occurrence of crime (Estimate = 1643) and the Kameeldrift Station in Gauteng had the greatest negative impact on occurrence of crime (the least records of crime predicted) having an estimate of -514.

Further, the treatment effect represent the crime types, and have their least square means are displayed in Table 7.5. We notice that Other Robbery is most occurring and Other Child related crime are least frequent.

Table 7.5: LSM for Lattice design

Crime Type	Estimate	Standard Error	DF	t Value	Pr > t
Agriculture	11.6455	48.3983	9112	0.24	0.8099
Household	407.22	48.3983	9112	8.41	<.0001
Murder	11.7964	48.3983	9112	0.24	0.8074
Other Child	-7.2369	48.3983	9112	-0.15	0.8811
Other Contact	332.23	48.3983	9112	6.86	<.0001
Other Police Action	287.41	48.3983	9112	5.94	<.0001
Other Robbery	609.03	48.3983	9112	12.58	<.0001
Other Sexual	41.7675	48.3983	9112	0.86	0.3882
Vehicle	174.74	48.3983	9112	3.61	0.0003

The corresponding spatial map can be found in Figure 7.6. This was coded using SAS and the legend key provided shows that the frequency of crime was lower than 2000 for the entire data set (there were no records higher than 8000, hence there are no red dots). The spatial distribution shows that there high frequencies of crime in the province of Gauteng, KwaZulu-Natal, Eastern Cape and Western Cape.

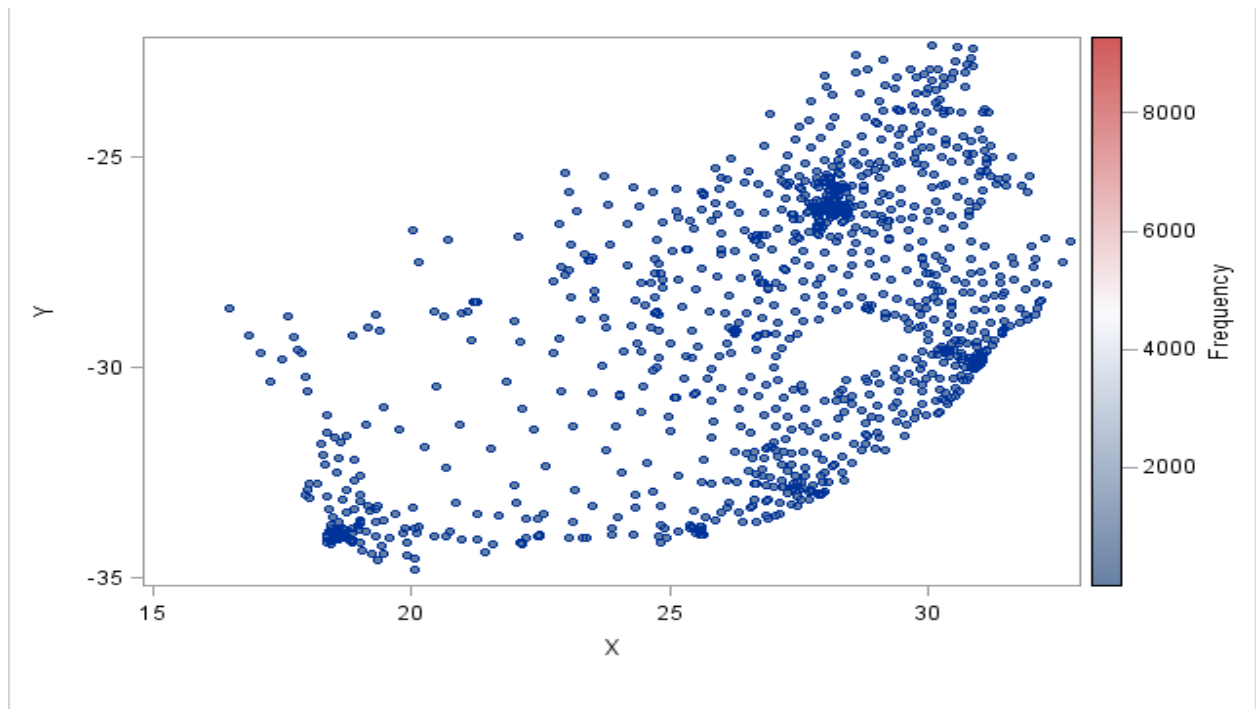


Figure 7.6: Spatial Distribution of Frequency Observations

This figure further shows large areas with no or low crime and also possibly due to lacking police stations, quite probably as there is very low or no inhabitants of some of the areas where there are no dots.

Figure 7.7 maps the spatial distribution of crime in a 3D plot. This shows the small-scale variation typical of spatial data, but there does not appear to be any surface trend (random peaks appearing).

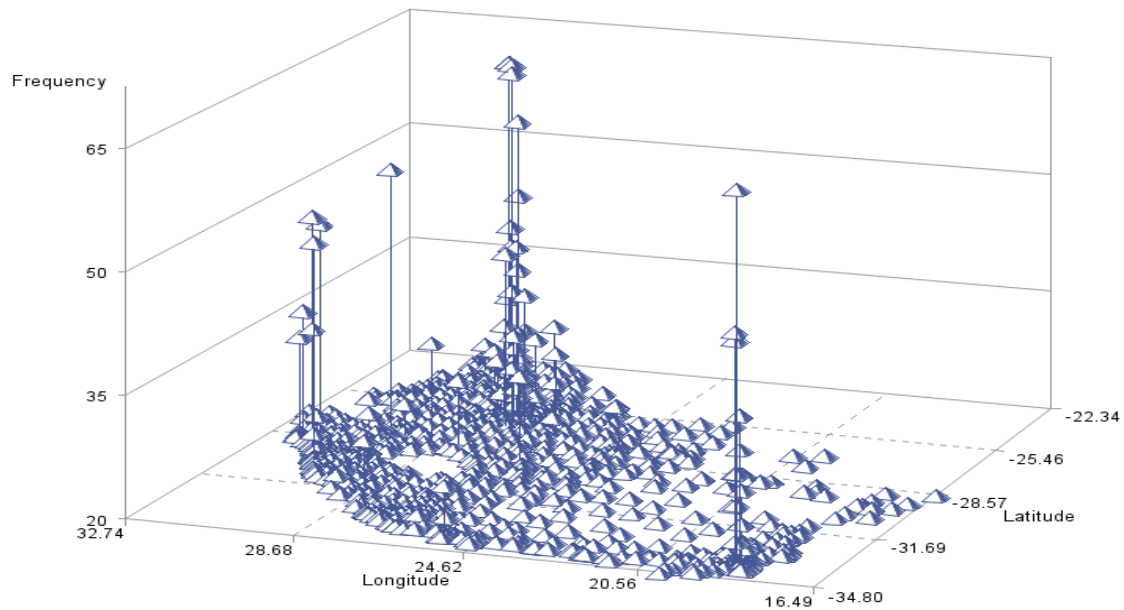


Figure 7.7: Surface Plot of Crime in South Africa

This illustrates further that there are high frequencies of crime occurring in the province of Gauteng, Western Cape, Eastern Cape and KwaZulu-Natal (as agreed by the spatial distribution of crime).

We now conduct our spatial correlation analysis by considering the semi-variance. Figure 7.8 shows the semi-variogram for the log transformation of our data (log output of the frequency of crime). The behavior of the semi-variance is spherical, this can be seen by the linear trend at the origin of the empirical model (this represents properties of higher levels of crime with short-range variability).

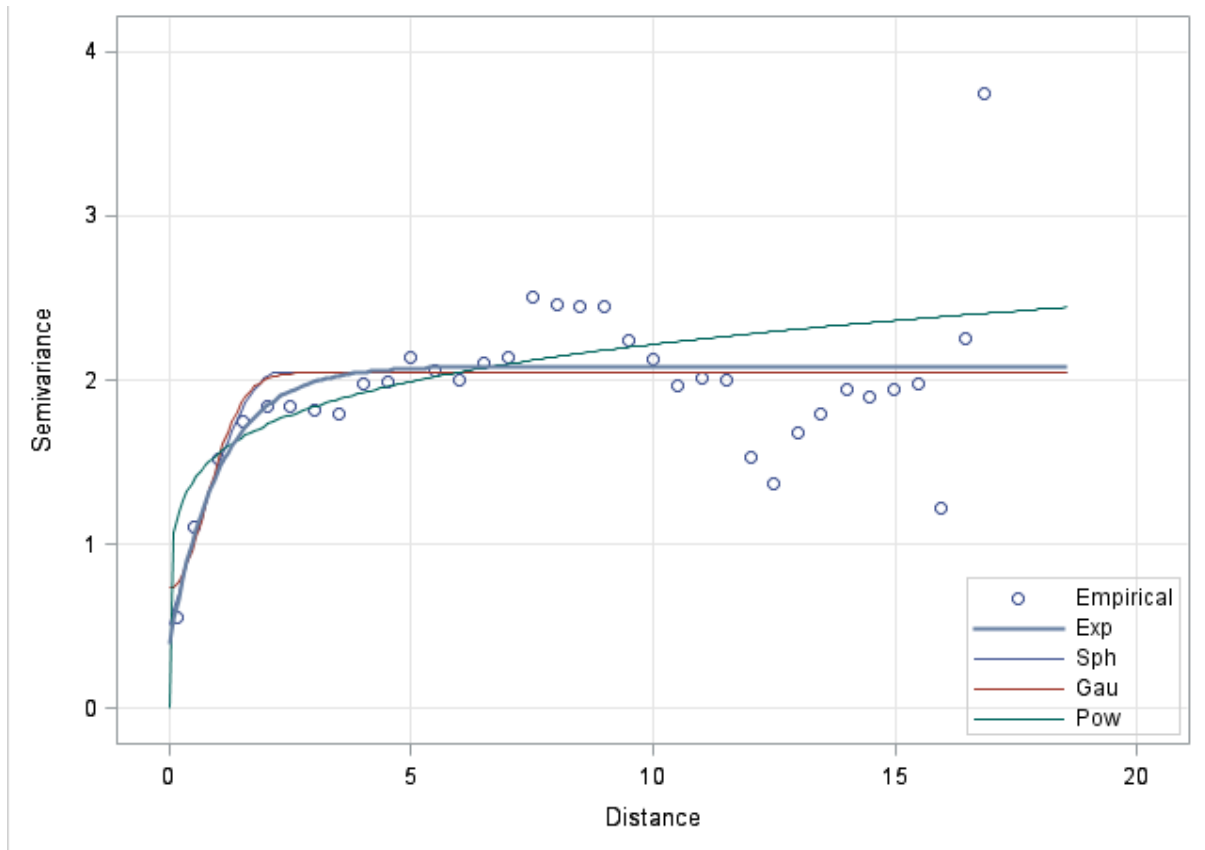


Figure 7.8: Semi-variogram for our data

We next test fitting all models to find the overall best model. The results also bears evidence that some models are indistinguishable, but we found that between the models that are distinguishable, with the best fitted model overall chosen by the AIC classification criteria (the preferred model is the one with the minimum AIC value). The different models are shown on Figure 7.8, the best fitting selection model is confirmed to be the Exponential model (having the lowest AIC value of 171.238).

Based on Table 7.6, the logarithm of frequency of crime (spatial correlation), a small nugget effect of 0.392 rises to a sill (or scale) value of 1.693, a rise of the exponential type. The observed log frequencies go as high as 10.15, which corresponds to the frequency of crime of 25 575 incidents reported.

Table 7.6: Semi-variogram function values

Parameter	Estimate	Approx Std Error	DF	t Value	Approx Pr > t	Gradient
Nugget	0.3917	0.01426	32	27.46	<.0001	-0.00212
Scale	1.6930	0.01351	32	125.34	<.0001	-0.00155
Range	1.0376	0.01732	32	59.89	<.0001	0.000591

Any of models are expected to exhibit similar behavior in terms of spatial correlation. They would result in the same output, however we choose the exponential behavior to use for the spatial prediction.

Figure 7.9 illustrates the kriging aspect, discussed at the beginning of this section, and it shows a surface of the predicted Frequency of crime values. According to the predicted values, the highest frequencies of crime are located as the white regions within the contours (we notice white patches in and around Gauteng), whereas the lowest frequency is observed at the north western parts. This map provides a useful indication of the spatial distribution of Crime in South Africa. The prediction errors appear to be relatively high (550-675) throughout the domain and, as you would expect, they increase as you move further away from the observations (550 increases to 600 then to 650 and so on as you move outwards).

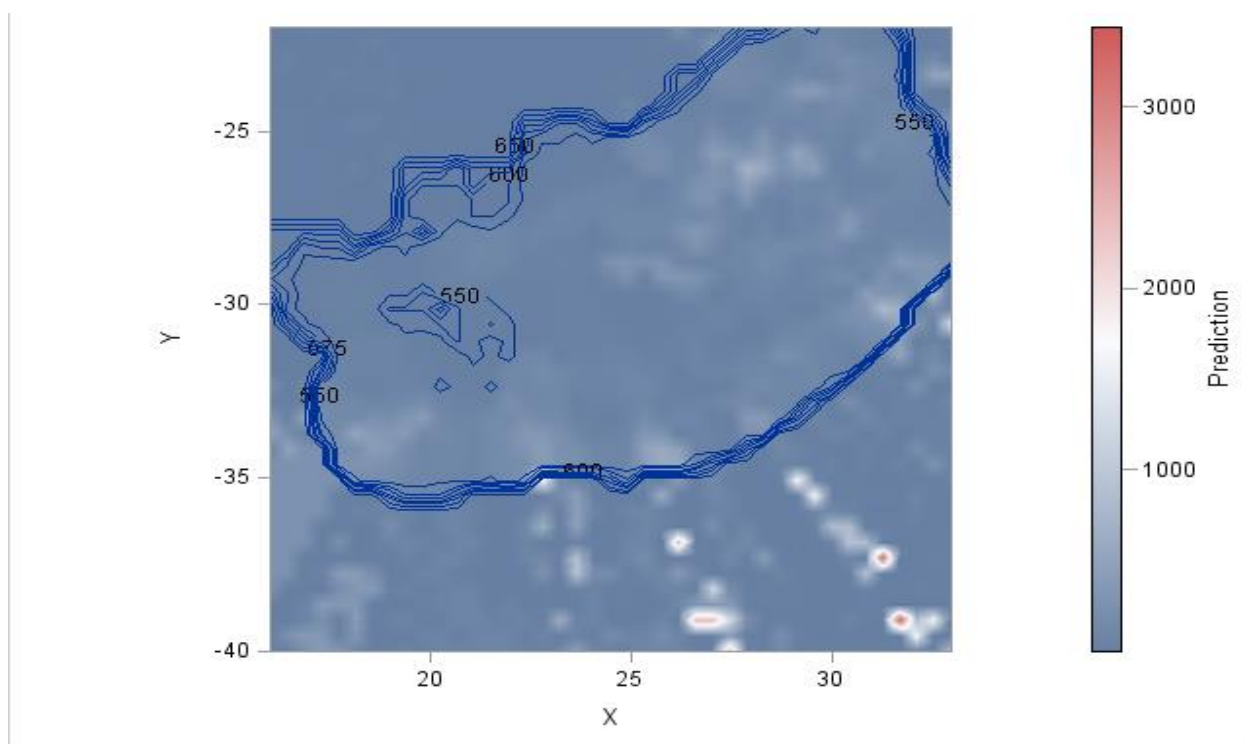


Figure 7.9: Contour map of Frequency of crime prediction

The major part of this chapter was dedicated to visually representing crime with the means of spatial plots, where in particular we illustrated both point in time crime occurrences, as well as predicted occurrences of crime. We analyzed the model estimates that explained the different relations to crime, and the association between different types of crime. This chapter was unique in the sense of using spatial data (co-ordinates), where we were mindful to model location of crime within the country, and then found that the results were in accordance with that found in previous chapters!

8. Conclusion

The closing chapter serves as a summary of the entire work presented, where we consider the aims set out in our opening chapters (including the literature review) and give evidence of this being fulfilled.

In the introduction, we explained the particular crime situation unique to South Africa. Different leaders in the country emphasized this problem within the country and in particular made statements about crime in the county in comparison to the other countries.

We presented the Victims of crime survey (VOCS) in the next chapter, where the findings of this survey were summarized. For the main parts of this study we used this information as the data source, so that we performed descriptive statistics as well as performing statistical modelling to analyse this data.

Our study differed from the VOCS report given in the Literature Review, in the sense that our task was not to prove or disprove any of the findings, yet it was to expand on them. We analysed the perception data in comparison to the actual incidences of crime.

Our comparative analysis showed that the odds of becoming a victim of crime among those who perceived it is 4.32 times higher than those who did not perceive to be a victim of crime. Our other simple statistics showed that Household crime occurred the most in comparison to vehicle theft, murder, agricultural and other crime.

Our fourth chapter generated the attributes of a crime victim, similar to the Canadian Statistics Report in the Review, the authors profiled both the crime victims as well as the perpetrators. We found that they were typically female, of the Black race group, older than 20 years of age and earning a pension income.

With regards to the location aspect of crime, we found that less crime occurred in the Free State, and the most in Mpumalanga (sample crime rates used similar to the Canadian Crime Study), and that the hot spots were in vast open areas like parks and also in bushy areas. The VOCS findings summarized in the Literature review notes that residents in the Free State Province feel least safe at night, which contradicts our findings that the Free State Province has the lowest crime rate. The contradiction can be explained that one instance is a perception and the other instance is an actual outcome.

We began our predictive analysis, where using a logistic model approach, we found the interaction of province with age, gender, and race to be significant. Our aim was to model location of crime and we found that the linking factor of province depicted that location is indeed influential in the occurrence of crime.

We continued trying alternate predictive models in our sixth chapter (using different/multiple target variables). It was interesting to find that there was a significant association between Household and Vehicle crime, where both occurred in concurrence. Like the reviewed Vancouver study, we were able to find high frequencies of burglary recorded as well as association between vehicle theft and burglary.

In our penultimate chapter we studied spatial data, where we first located most of the reported crimes to have come from the Cape Town police station, followed by Mitchells Plane and then Johannesburg central.

Spatially there are high frequencies of crime in Gauteng, Western Cape, KwaZulu-Natal and Eastern Cape. We found evidence that the Western Cape had the highest proportion of reported crime and Limpopo province had the least crime.

Using spatial modelling we found that Murder (the most severe type of crime) can be significantly predicted when agricultural, sexual, or child related crimes were committed. We were able to graphically predict further occurrences of crime as occurring more at inland areas, one of which was the Gauteng province. This prediction methodology (Kriging) can relate to Spatial intensity methods explored in the literature review.

Limitations of this study are similar to any prediction modelling research that is, the lack of accuracy. We discover the crime hot spots around the country based on historical survey data and we assume these patterns to be true for future incidences of crime in the same areas.

We recall further that the crime database has little to no information available for the perpetrators of these crimes or the reasons thereof. The next step analysis of this research would be to delve into characteristics of offenders which gives us an insight as to why crimes are occurring, i.e. the factors of a perpetrator of crime found significant could be used to profile a perpetrator, we could then better understand the qualities of crime offenders or in other words reasons for committing crime.

In this study, our main objective was to look for patterns and predictors of crime, in an attempt to add to the process of minimizing the crime rate of the country. We trust that results obtained can inform prevention strategies, so that we hope to have fulfilled these aims outlined at the outset of the study.

In laymen's terms, this research was done so that we could find out where crime was committed most often. People who were surveyed in South Africa agreed that open fields or parks were the most dangerous. Further, the Province of Mpumalanga had the highest number of criminal activities. The results provided by the South African Police Services showed that the Cape Town Police Station (in the Western Cape Province) had the highest number of crimes recorded. In conclusion, these findings can serve as to warn South African's of high risk areas as well as inform the SAPS of which areas to increase their security.

Bibliography

- Agresti, A. (2002). Logistic Regression. In *Categorical Data Analysis, Second Edition* (pp. 165-210). Gainesville, Florida: A John Wiley & Sons,inc., Publication.
- Anselin, Luc; Cohen, Jacqueline; Cook, David; Gorr, Wilpen; Tita, George;. (2000). Spatial Analyses of Crime. *Criminal Justice*, 214-222.
- Ayele, D. G., Zewotir, T. T., & Mwambi, H. G. (2013). Spatial distribution of malaria problem in three regions of Ethiopia. *Malaria Journal*.
- Bandyopadhyay, D. (2011, Spring). Lecture 27: Introduction to Correlated Binary Data . *BMTRY 711: Analysis of Categorical Data* . Division of Biostatistics and Epidemiology: Medical University of South Carolina.
- Benoit, K. (2012, August 22). Multinomial and Ordinal Logistic Regression. *ME104: Linear Regression Analysis*.
- Bolker. (2013). *GLMM notes*. Retrieved from Mc Master University: http://ms.mcmaster.ca/~bolker/classes/s4c03/notes/GLMM_Bolker_draft5.pdf
- Brennan, Shannon; Dauvergne, Mia;. (2010, July). Police-reported crime statistics in Canada. *Juristat article*.
- Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 4-28.
- Czepiel, S. A. (n.d.). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*. Retrieved from <http://czep.net/contact.html>
- Frank , Richard; Brantingham, Patricia L; Farrell, Graham;. (2012). *Estimating the True Rate of Repeated Victimization from Police Recorded Crime Data: A study of Burglary in Metro Vancouver*. Metro, Vancouver: Institute for Canadian Urban Research Studies, Simon Fraser University.
- GEE for Longitudinal Data- Chapter 8*. (2015). Retrieved from www.uic.edu/classes/bstt/bstt513
- Getting the most out of South Africa's crime statistics*. (2013, November 6). Retrieved March 2015, from ISS Africa: <http://www.issafrica.org/events/getting-the-most-out-of-south-africas-crime-statistics>
- Gould, C. (2014). *Why is crime and violence so high in South Africa?* Institute for Security Studies.
- Gould, W. (2000). Interpreting logistic regression in all its forms. *Stata Technical Bulletin (STB)*, published by Stata Corporation, 24-28.
- High, R. (2011). Interpreting the Differences Among LSMEANS in Generalized Linear Models. Omaha, NE: University of Nebraska Medical Center.
- IOL news. (2008, August 25). *What causes crime?* Retrieved March 2011, 03, from <http://www.iol.co.za/news/south-africa/what-causes-crime-1.413755#.VOHjPjGIXIY>
- Johnston, G. (1996). Repeated Measures Analysis with Discrete Data Using the SAS System. SAS Institute.
- Kleinbaum, D. G., & Klein, M. (2002). Introduction to Logistic Regression. In *Logistic Regression, A Self-Learning Text, Second edition* (pp. 15-19). Atlanta, GA, USA: Springer.
- Kong, R. (1997). Canadian Crime Statistics. *Juristat*, 12-13.
- Macdonald, I. (2008). Guestline. *South Africa: Should I stay or should I go?* SA Good News.
- Masuku, S. (2002, November). *PREVENTION IS BETTER THAN CURE: Addressing violent crime in South Africa*. Retrieved from Institute for Security Studies. Published in SA Crime Quarterly No 2: <https://www.issafrica.org/publications/south-african-crime->

- quarterly/south-african-crime-quarterly-2/prevention-is-better-than-cure-addressing-violent-crime-in-south-africa-sibusiso-masuku
- Masuku, S. (2003). *South African crime trends in 2002*, Institute for Security Studies. Retrieved from FOR BETTER AND FOR WORSE: <https://www.issafrica.org/Pubs/CrimeQ>
- Measures of relative effect: the risk ratio and odds ratio*. (2015). Retrieved from General methods for Cochrane reviews: handbook.cochrane.org/chapter_9/9_2_2_2_measures_of_relative_effect_the_risk_ratio_and_odds.htm
- Nation Master*. (2014). Retrieved February 16, 12:43PM, 2015, from Nation Master: <http://www.nationmaster.com/country-info/profiles/South-Africa/Crime>
- Nesstar metadata*. (2013/2014). Retrieved March 25, 2015, from Stats SA: <http://interactive.statssa.gov.za:8282/webview/>
- Nesstar questionnaire*. (n.d.). Retrieved March 20, 2015, from Stats SA: http://interactive.statssa.gov.za:8282/metadata/surveys/VOCS/2013_2014/VOCS2013_2014.htm
- Norusis, M. (2015). *Ordinal Regression*. Retrieved from Chapter 4: www.norusis.com/pdf/ASPC_v13
- Roane, B. (2014, April 22). *SA in top 10 on world murder list*. Retrieved from IOL beta: beta.iol.co.za/news/crime-courts/sa-in-top-10-on-world-murder-list-1678894
- SAPS. (2015). Retrieved April 28, 2015, from <http://www.saps.gov.za/services/boundary.php>
- SAPS careers*. (2014). Retrieved from South African Police Service, Department of Police: www.saps.gov.za/careers/careers.php
- SAS Knowledge Base*. (2010). Retrieved from SAS/STAT(R) 9.2 User's Guide, Second Edition: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glimmix_sect006.htm
- SAS/STAT(R) 9.2 User's Guide, Second Edition*. (2010). Retrieved from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glimmix_a0000001398.htm
- Shaw, M., & Kriegler, A. (2016, August 30). *Rand Daily Mail*. Retrieved from <http://www.rdm.co.za/politics/2016/08/30/a-guide-on-what-to-look-for-in-sa-s-troubling-crime-statistics>
- South Africa World crime Capital? (2001). *Nedbank ISS Crime Index Volume 1*.
- South Africa's crime stats*. (2014). Retrieved February 2015, from The good, the bad and the ugly.: <http://www.thesouthafrican.com/south-africas-crime-stats-2014-the-good-the-bad-and-the-ugly-infographic/>
- The Area Under an ROC curve*. (2015, 7). Retrieved from <http://gim.unmc.edu/dxtests/roc3.htm>
- Victims of Crime Survey 2013/14*. (2014). Statistics South Africa.
- Wang, F.-L. (2011). *Logistic Regression: Use & Interpretation of Odds Ratio (OR)*. Community Health Sciences, the University of Calgary.
- Wikipedia*. (2015). Retrieved February 05, 2015, from Crime: en.m.wikipedia.org/wiki/Crime
- Wild, C. (2016). *iNZight for Data Analysis*. Retrieved from <https://www.stat.auckland.ac.nz>
- Williamson, R. C. (1957). Crime in South Africa: Some Aspects of Causes and Treatment. *Journal of Criminal Law and Criminology*, 185-192.
- Zhang, Haifeng; Peterson, Michael P.; (2007). A spatial analysis of neighbourhood crime in Omaha, Nebraska using alternative measures of crime rates. *Internet Journal of Criminology*, 1-31.

List of Figures

Figure 4.1: Perception of crime in SA (% who believe this crime has occurred).....	14
Figure 4.2: Experiences of crime over the last 5 years	15
Figure 4.3: Distribution of all crimes in South Africa, by province (%).....	16
Figure 4.4: Preventions due to crime in SA.....	16
Figure 4.5: Victims and non-victims of crime by age group	17
Figure 4.6: Victims of crime (last 5 years), broken down by Income type, race and gender (%).....	17
Figure 4.7: Corruption in South Africa – perceptions of the reason for corruption (%).....	18
Figure 4.8: Crime reduction methods as suggested by study group	19
Figure 4.9: Victim support structures (%).....	19
Figure 4.10: Responses to crime incidents (%).....	20
Figure 4.11: Response times (minutes) of police officers, as reflected by different race groups (%).....	21
Figure 4.12: Victims satisfaction levels broken down by types of crime	21
Figure 4.13: Visible policing by province.....	23
Figure 4.14: Perception verses outcome of crime.....	23
Figure 5.1: Probability diagram for interaction Gender*Province.....	29
Figure 5.2: Probability diagram for interaction Race*Province.....	30
Figure 5.3: Probability diagram for interaction Age*Province.....	31
Figure 5.4: Ordinal Province vs Gender interaction.....	42
Figure 5.5: Ordinal Province vs Race interaction.....	43
Figure 5.6: Ordinal Province vs Age interaction.....	44
Figure 6.1: Province vs Gender interaction.....	49
Figure 6.2: Province vs Age interaction.....	50
Figure 6.3: Province vs Race interaction.....	50
Figure 7.1: Semi-variogram.....	54
Figure 7.2: iNZight plot of police stations.....	55
Figure 7.3: Spread of crime per province.....	56
Figure 7.4: LSM comparison of categories of crime.....	59
Figure 7.5: LSM graph comparing murder to all categories of crime.....	60
Figure 7.6: Spatial Distribution of Frequency Observations.....	62
Figure 7.7: Surface Plot of Crime in South Africa.....	63
Figure 7.8: Semi-variogram for our data.....	64
Figure 7.9: Contour map of Frequency of crime prediction.....	65

List of Tables

Table 4.1: Perception verses Outcome contingency table.....	24
Table 4.2: Odds Ratio output.....	24
Table 5.1: Summary of significant factors.....	28
Table 5.2: Odds Ratio Estimates for victims of crime.....	29
Table 5.3: Summary of SAS output for victim models.....	31
Table 5.4: Summary of SAS output for perception models.....	34
Table 5.5: Parameter estimate output (subset).....	34
Table 5.6: Survey Logistic Profile.....	36
Table 5.7: Parameter estimate output (survey).....	36
Table 5.8: Ordinal regression type 3 analysis.....	39
Table 5.9: Parallel line assumption test.....	39
Table 5.10: Ordinal regression parameter estimates.....	39
Table 5.11: Odds Ratio Estimates for Ordinal Regression.....	41
Table 6.1: Significant Parameter Estimates.....	48
Table 6.2: Correlation matrix of categories of crime.....	51
Table 7.1: Crime rates of province by crime category.....	57
Table 7.2: GLIMMIX model output.....	57
Table 7.3: LSM model estimates output.....	58
Table 7.4: Differences of Category least squares means (Dunnett comparison).....	60
Table 7.5: LSM for Lattice design.....	61
Table 7.6: Semi-variogram function values.....	64

List of Equations

Equation 1: Logit transformation.....	26
Equation 2: Probability derivation.....	26
Equation 3: Odds ratio.....	26
Equation 4: Hosmer Lemeshow test statistic.....	27
Equation 5: Sample mea.....	36
Equation 6: Sample variance.....	36
Equation 7: Ordinal odds model.....	38
Equation 8: Likelihood Ratio Test.....	38
Equation 9: Generalized linear model (GLM).....	45
Equation 10: GEE link function.....	46
Equation 11: GEE variance function.....	46
Equation 12: Generalized estimating equation (GEE).....	46
Equation 13: GEE covariance matrix.....	46
Equation 14: Prediction model using GEE.....	47
Equation 15: Generalized linear mixed model (GLMM).....	53
Equation 16: Transformation PQL.....	53
Equation 17: Final PQL estimating equation.....	53
Equation 18: Semi-variogram calculation.....	53
Equation 19: Exponential spatial covariance structure.....	54
Equation 20: Lattice linear model.....	61