Spatial patterns and determinants of residential burglary in Gauteng for 2011/2012

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

Sabelo Mguli: 0709828w

2019-07-30

Supervisor: Mr S Salau

Co-supervisor: Dr S Merlo

Declaration

I declare that this report is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

____

Signature of candidate

30th day of July 2019 at Johannesburg

Abstract

South Africa ranks amongst the highest in terms of reported crimes in the world, yet there are not many published research papers that study the spatial dynamics of crime. This report used statistical and spatial methods to explore determinants and patterns of burglary in Gauteng, South Africa. The variables considered for modelling were informed by the Routine Activities (RA) theory and other local studies. The methods of analysis used were ordinary least squares (OLS) regressions and geographically weighted regressions (GWR) - an extension of the OLS regression used when there is spatial heterogeneity in the data. The findings show that the measure of social cohesion and physical neighbourhood characteristics such as roads density, percentage of the population unemployed together with the percentage of the population that moved in the last 5 years were significant predictors of residential burglary. The findings provide stronger support for the social disorganisation (SOD) than for the RA theory in the South African context. The spatial patterns suggested that there were hot spots, colds spots clusters in certain parts of the study region. Evidence of heteroscedasticity of the influential factors of burglary was found, although marginally. Although heteroscedasticity at 5 percent significance level was not found in the model variables, the results of the GWR improved the OLS model at 10 percent significance level. The GWR results implied that different strategies of crime prevention could be implemented in different parts of Gauteng.

Dedications

This report is dedicated to my late parents; Nomalizo O. Mguli and Mbulelo L. Mguli, my siblings, Nombongo Mguli and Ntombifuthi Sonqolo. I dedicate this report to my nephews and nieces, Ezingaka, Kungawo, Mandilake and Yongama. I hope this report will encourage you to excel in your future endeavours. Lastly, this thesis is dedicated to all the Mguli family, my colleagues, and friends.

Acknowledgements

I would like to thank God, the creator, for giving me the strength and courage to finish this report despite many challenges that frustrated me. Without God, I, would not be where I am given my upbringing and challenges I have faced in life.

I would like to wholeheartedly thank my supervisors, Dr. S. Merlo and Mr. S. Salau for their guidance, encouragement, and support. Without them, this would not be possible.

I would like to thank my sponsor, Department of Rural Development and land Reform for providing financial assistance. I also thank my supervisor at work, Mrs. E.T Janse van Rensburg for allowing me to pursue my dreams despite tight deadlines and volumes of work, I unreservedly thank her. I would like to thank Mr. Anton Reynecke for proofreading my work despite his busy schedule at work.

Lastly my nephew Mandilake Mguli for all the support and help at home. Without his consistent help, courage and emotional support I would not have been able to finish this report.

Table of Contents

Table of figures

List of Tables

1 Introduction

South Africa ranks amongst the highest in terms of reported crimes in the world (Bhorat *et al.*, 2017*).* This is despite the efforts by Government departments and other security agencies in combating crime in the republic. Although access to safety and security are constitutional rights of citizens in South Africa (SA), Statistics South Africa (StatsSA) states that the fear of criminal victimisation amongst citizens is high and affects the way citizens conduct their lives (StatsSA, 2014). The fear of crime victimisation and the high levels of crime in SA suggest the need for research and studies to aid crime prevention and accelerate crime reduction efforts.

The South African Police Service (SAPS) categorises crimes into four broad groups, namely, contact crimes, contact-related crimes, property crimes, and other crimes. Property crime refers to crimes whereby tangible assets of an individual or institution are stolen in the absence of the owner or guardian of such asset (SAPS, 2016). SAPS further sub-categorises property crimes into residential burglary, non-residential burglary, theft of the motor vehicle and motorcycle, theft out of or from motor vehicles and stock theft (SAPS, 2016). Burglary is the most feared kind of victimisation amongst South Africans (StatsSA, 2014). Residential burglary is defined as illegal breaking and entry into a residential dwelling with the intention of stealing goods while the owners or guardians are not inside the dwelling (StatsSA, 2014; SAPS, 2016). Non-residential burglary refers to a burglary committed against property that is not used for residential purposes (SAPS, 2016). Residential burglary in the presence of the owners is referred to as robbery (StatsSA, 2014). Robbery was excluded from this study. Residential burglary (crimes against buildings) will likely show evidence of geographic pattern or spatial effects in the form of spatial dependence/spatial autocorrelation (spatial grouping of similar values), and spatial heterogeneity (variation in spatial relationships throughout a study region) (Brantingham and Brantingham, 1975; Chen *et al.*, 2017; Bernasco and Nieuwbeerta, 2005; Malczewski and Poetz, 2005; StatsSA, 2014; Townsley *et al.*, 2015).

There is not a lot of empirical research published in SA in relation to property crime or residential burglary. While the ideal unit of analysis for burglary is at the household level, SAPS only publishes aggregated crime data and researchers normally go through a lengthy process to access incidence level data from SAPS. Thus, analysis of burglary data is commonly conducted using aggregate-level data at sub-provincial unit of analysis and have mainly used exploratory spatial data analysis and regression techniques (Fanaroff, Levine, Liebernman and Glaser, 2004; Breetzke, 2010, 2012; Breetzke and Cohn, 2013; StatsSA , 2014, 2017; Hiropoulos and Porter, 2014).

Despite observed decreases in the counts of residential burglaries committed nationally, the levels of crime in SA remain unacceptably high, (SAPS, 2016). Statistics released by SAPS suggest that Gauteng (the most populous province in SA) accounted for 26% of residential burglaries and contributed around 45% to the category of property crimes recorded nationally for the financial year 2011/2012. These statistics justify the need for research into factors that contribute to burglaries in Gauteng. This report aims at adding value to existing studies by conducting a province-wide analysis of residential burglary using a combination of 2011/2012 crime data for the 143 SAPS precincts (policing units) in Gauteng and data from the 2011 census.

1.1 Aim

The aim of the research was to explore the spatial patterns and the possible determinants that explain the distribution of burglary in Gauteng.

1.2 Objectives

The objectives were as follows;

Objective 1: Examine the Gauteng burglary data for spatial heterogeneity, autocorrelation, and clustering.

Objective 2: Fit spatial regression models to identify significant variables corresponding to the observed burglary patterns.

Objective 3: Test if the results change when the spatial unit of analysis is changed.

1.3 Research questions

The following research questions were posed in the research;

Question 1: Is there spatial auto-correlation present in the residential burglary rates in Gauteng?

Question 2: Are there significant clusters (hot spots/cold spots or outliers) of burglary in Gauteng?

Question 3: Is there spatial non-stationarity of burglary in Gauteng?

Question 4: What variables explain the spatial patterns of burglary in Gauteng?

1.4 Research hypotheses

Based on published literature on crime in SA, the following hypotheses are put forward:

I. Population density and accessibility (measured in terms of roads density) increases the risk of burglary in Gauteng.

II. Lack of social cohesion as measured by the percentage of people moving within a period of 5 years and ethnic heterogeneity index of a neighbourhood increases the chances of burglary in Gauteng.

III. Female-headed households have a higher chance of burglary victimisation in Gauteng.

1.5 Structure of the report

This is laid out in the following manner: Chapter 2 contains the literature review followed by Chapter 3 which explains the methodology adopted for the research. Chapter 4 outlines the results and Chapter 5 contains a discussion of the results. Chapter 6 has conclusions and recommendations, and limitations.

2 Literature review

2.1 Introduction

This report aimed to explore possible determinants (variables) and spatial patterns of burglary in Gauteng. The literature review covers the routine activities (RA) theory and other related environmental criminology theories. It also explored ways in which variables were chosen in previous studies and gives a summary of pertinent crime analysis literature.

Burglary is defined as an “illegal and intentional entry in part or full-body, upon a building or similar structure used for residential or commercial purposes and committing a crime for the purposes of personal gain in the absence of rightful guardian of that property” (SAPS, 2016). Investigations into the factors that contribute to the causes of burglary are undoubtedly required for crime prevention and reduction interventions by law enforcement institutions and the citizens who may be at risk. The incidence of burglary can be attributed to several reasons.

Scholars present various frameworks for understanding the occurrence of crime. One of the frameworks attempts to understand the psychological factors that make perpetrator’s commit illegal activities (Brantingham and Brantingham, 1975). Other scholars argue that crime may be explained by examining environmental factors (Bernasco and Block, 2009). Existing studies have empirically shown the importance of place and time in the understanding of crime events within the framework of environmental criminology (Cohen and Felson, 1979; Eck and Weisburd, 1995; Martin, 2002; Breetzke, 2012; Townsley *et al.*, 2015; Chen *et al.*, 2017).

Environmental criminology refers to a group of studies and theories that focus on the research of the environmental settings of locations where crimes occur. Environmental criminological research seeks to understand why perpetrators select an area to commit a crime (Bernasco and Luykx, 2003; Brown and Bentley, 1993; Cohen and Felson, 1979; Hipp, 2007). This review explored burglary victimisation studies from an environmental criminology point of view.

2.2 Environmental criminology: summary of the principal theories

In their study, Shaw and McKay (1942) as cited by (Sampson and Groves, 1989) found evidence that juvenile crimes were influenced by the environmental factors such as low economic status, racial heterogeneity and residential mobility (Sampson and Groves, 1989). Environmental factors discussed by Shaw and McKay (1942) led to a lot of studies that expanded on the environmental factors of communities and helped in the formulation of environmental criminology theories.

There are three theories related to environmental criminology. These are the routine activities (RA) theory, the rational choice (RC) theory and the crime pattern (CPT) (Brantingham and Brantingham, 1975; Brown and Bentley, 1993; Eck and Weisburd, 1995; Bernasco and Elfers, 2011; Zhang, Suresh and Qiu, 2012; Hiropoulos and Porter, 2014; Clack , 2015); Townsley *et al*., 2015). The social disorganisation (SOD) theory is another criminology theory closely related to the RA theory.

The RA theory was first presented by Cohen and Felson (1979). Under this theory, the authors argue that changes in everyday routine activities like going to work or school can magnify the risk of criminal exploitation (Cohen and Felson, 1979). Cohen and Felson (1979) conceptualise that the environment where crimes occur must include three elements; the shortage of a capable guardian, the motivated offender, and a suitable target. A capable guardian is defined to include any feature that discourages the motivated offender in committing a crime and may be classified into man-made features including but not limited to walls, security alarms and cameras, security personnel, guard dogs and the natural features such as weather and natural landscape (Cohen and Felson, 1979; Breetzke, 2012). A suitable target is a victim perceived by the motivated offender to be vulnerable, presenting financial gain and opportunity with least effort (Cohen and Felson, 1979; Bernasco and Luykx, 2003; Breetzke, 2012). The motivated offender is the perpetrator of crime who sees an opportunity to commit a crime for financial gain (Cohen and Felson, 1979). The RA theory is attractive from a spatial analysis perspective because it offers a simple and intuitive way of conceptualising the three elements under it compared to the RC and CPT theories (Breetzke, 2012).

The RA theory suggests that places perceived to be attractive and that lack capable guardians will be victimised by motivated offenders. However, one cannot possibly know if an area will have motivated offenders without assuming their existence in an area or study previous offenders that exist in an area. As such other theories that seek to understand the perpetrator's thought and travel processes were proposed. One such theory is the RC theory which states that the perpetrator of crime weighs up the opportunities at his or her disposal by assessing possible resources for personal gain against risks involved in committing the crime (Bernasco and Nieuwbeerta, 2005; Breetzke, 2012). Since perpetrators of crime study their options before committing crimes, scholars present the journey to crime studies. Journey to crime studies explore the proximity relationship between the offender's frequent location and crime location. The perpetrator's frequent location may include their place of residence. One of the components of the journey to crime studies is called distance decay pattern – which states that there will be a high concentration of crimes closer to offender's own home than further way (Bernasco and Nieuwbeerta, 2005; Breetzke, 2008; Rengert, Piquero and Jones, 1999).

While the RC theory focuses on the perpetrators thought process, the CPT theory explores the spatial setting of targets, offenders and opportunities, and how criminal events will occur at the intersection of awareness and opportunity spaces. Regions familiar to the perpetrator are referred to as awareness spaces while regions perceived to be easily accessible are referred to as opportunity spaces (Breetzke and Cohn, 2013; Hiropoulos and Porter, 2014). CPT offers an attractive way of studying crime patterns because it studies the relationship between the offender's routine activities and what informs them to commit a crime. However, it becomes difficult to materialise their routine activities without collecting data from past offenders.

The theories reviewed offer ways of understanding criminal events from an environmental perspective and can also explain crime events by studying the offender's thought processes. They, however, do not offer an explanation from a community based sociological perspective, which may be important to government institutions. Thusly the SOD theory was suggested by scholars.

The SOD theory states that communities that lack social cohesion are more at risk of crime victimisation (Bernasco and Block, 2009; Breetzke, 2010; Chastain, Qiu and Piquero, 2016). Social disorganisation theory is conceptualised around social community characteristics such that communities realise common goals and institute social controls (Sampson and Groves, 1989). These social controls include friendship ties and other informal social relationships such as community crime forums (Sampson and Groves, 1989). In simple terms, SOD theory studies social factors that increase the prevalence of crime because of a lack of social support structures that give communities a meaning. Racial heterogeneity, poverty, and residential mobility are some of the determinants identified to influence crime under the SOD (Sampson and Groves, 1989). SOD and RA theories are related because community characteristics may determine if regions are suitable targets under the RA theory (Sampson and Groves, 1989).

2.3 Crime research studies

Scholars that studied burglary in SA have conducted their research at police precinct level (Demombynes and Özler, 2003; StatsSA, 2014, 2017; Bhorat *et al.*, 2017) as well as at suburb level (Breetzke, 2010b, 2012; Breetzke and Cohn, 2013). Unemployment, income, social inequality and education levels were found to be influencers of burglary in these studies. Crimes were also influenced by demographic factors such as age, gender and racial composition (StatsSA, 2014). At the community level, racial heterogeneity and social cohesion measures were also selected by authors (Breetzke, 2012; Breetzke and Cohn, 2013). Scholars (Brown, 2001; Demombynes and Özler, 2003; StatsSA, 2014; Bhorat *et al.*, 2017) suggested that perpetrators commit burglaries because they are poverty-stricken. Breetzke (2010, 2012) and Breetzke and Cohn (2013) suggested that the nature of geographical features such as terrain may cause regions to be attractive and suitable to possible offenders. Geographic profiles of offenders were also studied in Breetzke (2008) to understand how offenders move and select regions for burglary. Geographic profiles of offenders are determined to include, but not limited to the way offenders move around their places of residence and criminal activity (Rengert, Piquero and Jones, 1999). A shortage of research in the area of environmental criminology in SA provides an opportunity to test these theories in the South African context.

In America, Europe, and Asia there is rich literature on crime and burglary providing extensive research employing environmental criminology theories.

In the United States, Cohen and Felson (1979) proposed the theory after realising that there was a lack of research that simultaneously examined certain factors that may affect crime; lack of capable guardian, a suitable target and motivated offender. Malczewski and Poetz (2005) in London, Ontario studied neighbourhood socio-economic factors that influence residential burglaries and they reported that neighbourhoods with minimally capable guardians were at greater risk of increased victimisation. Brown and Bentley (1993) investigated the use of crime deterring objects such as fences and locks to understand how they affect burglary. Their results have shown evidence of deterrence of burglary in households displaying these territorial features (Brown and Bentley, 1993). Social cohesion measures in an area were also found to deter crime. The idea is that burglars scan the area and assess the level of territoriality of households and the chances of being spotted by a bystander before they commit such crime (Brown and Bentley, 1993; Vélez, 2001; Breetzke, 2012). In China, Zhang, Messner and Liu (2007) found strong support for the RA theory, with factors such as affluence length of residence and measures of social cohesion; public control and collective efficacy affecting burglary in the cities of China. Townsley *et al*. (2015) found that the chances of an area being selected for burglary in the Netherlands, Australia, and the United Kingdom were consistently influenced by proximity to an offender's home, the volume of easily accessible targets and the targets available for selection.

Researchers have reported contradictory findings to what the RA theory states, particularly as it relates to a capable guardian. The presence of a capable guardian should deter criminal activities from happening, while, the lack of a capable guardian should increase burglary victimisation. In SA, Breetzke and Cohn (2013) reported unexpected results relating to the RA theory, namely they found that residing in a gated community (a form of capable guardian) did not deter burglary from occurring. Similar outcomes were found by Heiple (2010) in the USA where households that installed more security features/measures were likely to be victimised than households that did not install security features. Such mixed findings in burglary studies suggest that scholars have to look at various ways of explaining burglary particularly when RA is used as a backcloth for studying crime.

Bernasco and Nieuwbeerta (2005) proposed a method that assesses whether the choice of suitable target by a burglar is affected by affluence, accessibility and poor guardian in the City of Hague, in the Netherlands. They theorise that the choice of a place for burglary may be explained by both examining the target characteristics and offender characteristics simultaneously such as neighbourhood affluence, proximity to where offenders live and the expected likelihood of successful completion of a burglary. They found that the chances of an of burglary victimisation in an area are positively impacted by its perceived lack of suitable guardian, accessibility; as measured by the category of a single-family dwelling, as well as the number of expected objects to steal, such as electronic devices. The authors also studied the effect of social cohesion factors on community-level burglary.

In the United States of America, areas characterised by lack of social cohesion were found to be more vulnerable to burglary victimisation (Vélez, 2001). Vélez (2001) used the SOD theory to study the characteristics of public social control (perceived efforts by law enforcement in mitigating crimes) and their impact on criminal victimisation. Occurences of crime were more evident especially in more affluent areas that lacked social cohesion structures (Vélez, 2001).

3 Methodology

3.1 Study area

Gauteng is one of the 9 provinces in SA. It is bounded by Mpumalanga to the East, Free State to the South, North West province to the West and Limpopo to the North (Figure 3.1) and has an average height of about 1500 metres above sea level. Gauteng constitutes only 1.4% of the total land area of SA, making it the smallest province in SA. Economically, Gauteng contributes 34% to the country's gross domestic product (StatsSA, 2015). It has the largest population in the country, estimated at 12,271,707 and has a population density of 680 people/km² as per the 2011 census survey (StatsSA, 2014). There are 143 police precincts in Gauteng.

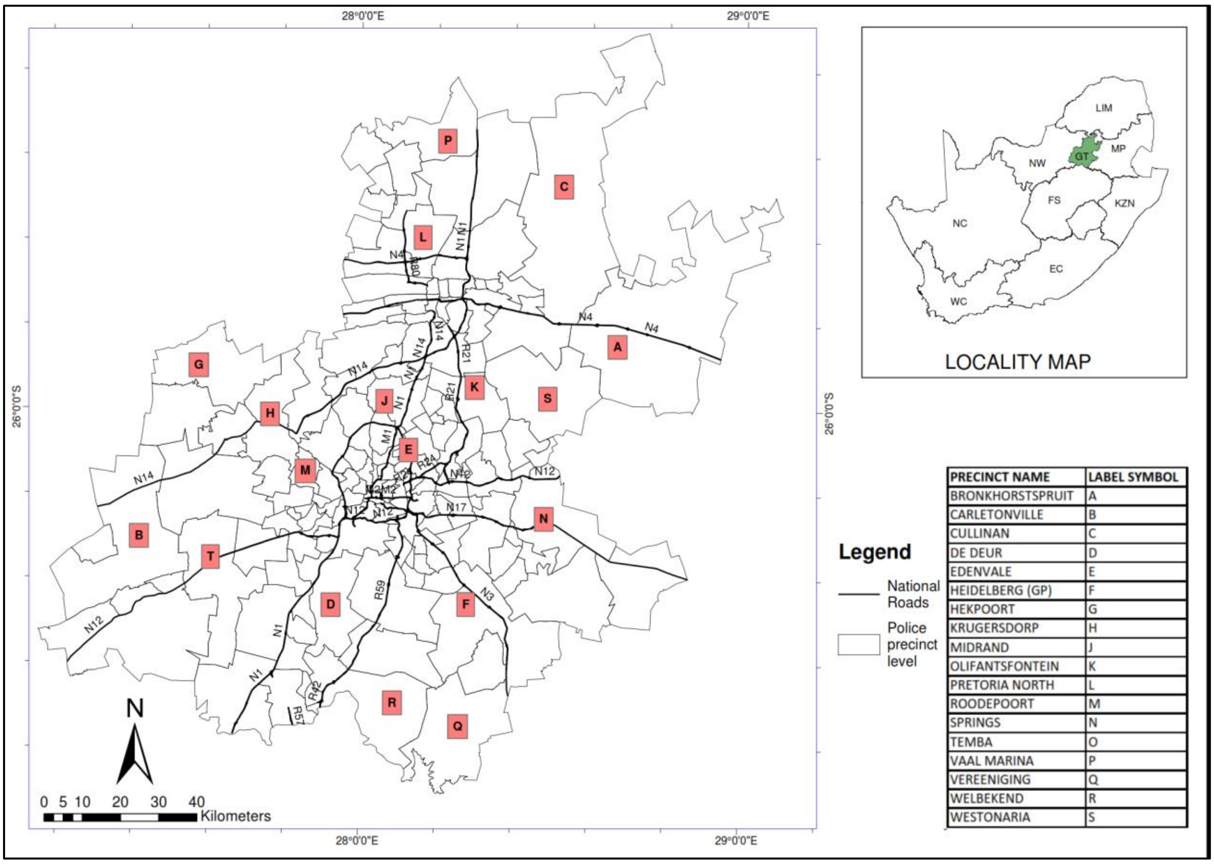


Figure 3.1 Study area map; a sample of police precincts labelled. The inset map shows the location of Gauteng in SA, 2011, (Source Open street map, StatsSA, SAPS)

3.2 Data

The variables that were considered in the report come from a variety of sources and are listed in Table 3.1. The variables were consolidated at the police precinct level consisting of 143 polygons and are of the lattice data type. Lattice data are represented by means of regular or irregular polygons and the process that generated the data is random (Cressie, 1991). Other types of spatial data include point pattern data and geostatistical data (Cressie, 1991) but these were not considered in the report. The choice of variables was informed by previous studies (Breetzke, 2012; StatsSA, 2014, 2017; Malczewski and Poetz, 2005; Zhang, Messner and Liu, 2007). The variables that capture suitable target are the percentage of the population with BSc degree, percentage of female-headed households, percentage of White-headed households, population density and roads density (a proxy for accessibility). Motivated offender and lack of capable guardian elements were captured using the set proxy variables in Table 3.1. This report conducted a province-wide analysis of residential burglary using a combination of 2011/2012 crime data for the 143 SAPS precincts (policing units) in Gauteng and data from the 2011 census. Additional information on the variables can be found in Appendix A.

Table 3.1: Variables explored in the report

Variable Measure	Variable	Source	Resolution	Choice Informed by:
Dependant				
Burglary	Burglary Rate (Per 1000 Households)	SAPS	Police Area and Police Cluster Area	Routine Activities Theory and Breetzke (2012)
Independent Variables				
Motivated Offender	Percentage of population Unemployed	StatsSA	Small area Level, 2011	Routine Activities Theory and Breetzke (2012)
	Percentage of Males aged 15 to 34 years	StatsSA	Small area Level, 2011	Routine Activities Theory and Breetzke (2012)
	Percentage of households with no income	StatsSA	Small area Level, 2011	Routine Activities Theory and Breetzke (2012)
Suitable Target	Female headed households	StatsSA	Small area Level, 2011	Routine Activities Theory and Breetzke (2012)
	Percentage of Population with BSc degree and above	StatsSA	Small area Level, 2011	StatsSA, 2017
	Percentage of white headed households	StatsSA	Small area Level, 2011	StatsSA, 2017
	Population density	StatsSA	Small area Level, 2011	Routine Activities Theory and Breetzke (2012)
	Accessibility in terms of Road Density(km/km^2)	Open Street Map	Primary up to tertiary level, 2017	Routine Activities Theory and Breetzke (2012)
Capable Guardian	Ethnic heterogeneity index	StatsSA	Small area Level, 2011	Bernasco and Lyukx (2003)
	Percentage of population that moved in the last 5 years	StatsSA	Small area Level, 2011	Zhang, Messner and Liu (2007) Poetz and Malczewski (2005
	Percentage of population with of no citizenship	StatsSA	Small area Level, 2011	Zhang, Messner and Liu (2007)

The burglary data are available at the police precinct level for Gauteng. The attributes linked to the 143 polygons are the published residential burglary counts for the 2011/2012 financial year (the period from 1 April 2011 until 31 March 2012). The dependant variable; burglary rates, was standardised by household counts from the census because burglary affects dwellings. In this report, all types of the dwelling; that is; flats/apartments, enclosed security estates or standalone dwellings were combined in the denominators used for the calculation of burglary rates. For the road densities, all road types intersecting police precincts level data were used.

The census data consisting of household and individual characteristics are available at the Small Area level (SAL) made up of 17 840 polygons. A total of 451 SAL polygons with no household information were removed to ensure correct calculation of road and population densities. Some of these SALs contained game reserves, mine dumps, and industrial areas.

Police precinct level and census level data are not aligned and thus aggregation of census data to police precinct-level data was necessary as detailed in Section 3.3. An example of this misalignment (Figure 3.2) shows the Kameeldrift police precinct polygon outlined in yellow, while the polygons outlined in black with pink hatching represent the SAL boundaries. The highlighted polygon in turquoise represents the overlapping SAL boundary overlapping between the Kameeldrift and Hammanskraal police boundaries.

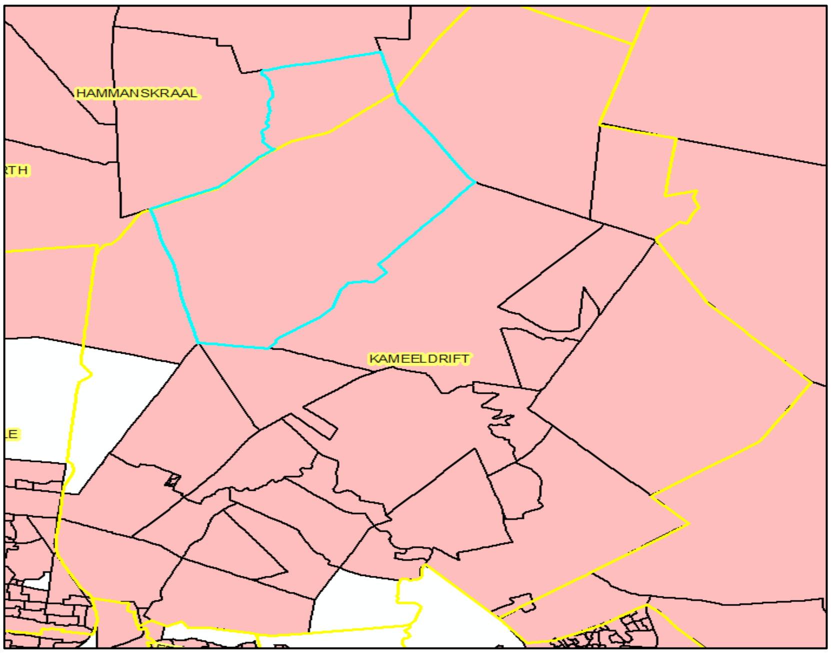


Figure 3.2 Misaligned police and census (SAL) boundaries (black with pink hatching), yellow = police precinct, turquoise = overlapping SAL boundary)

3.3 Aggregation of variables: area weighting of variables

A simple area weighting technique was used to assign the census variables at the SAL level to the police precinct level using ArcMap 10.5. The area weighting technique involves intersecting the source zones (SAL) by the target zones (police precinct) and estimating values at the target zones (police precinct) based on the areas formed by the intersected zones. Figure 3.3 shows the steps followed under this technique.

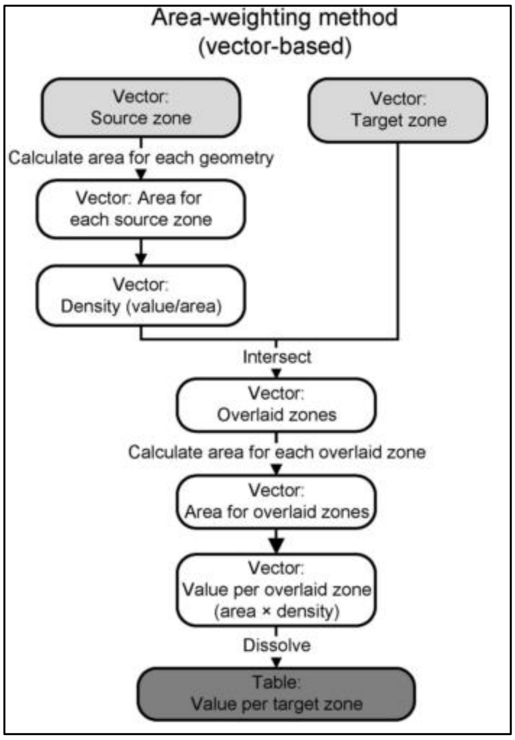


Figure 3.3 Simple Areal weighting aggregation flow chart (Source: Qiu, Zhang and Zhou, 2012, Figure 1, page 650)

The goal of the aggregation was to estimate values of the census (SAL) level variables at the police precinct level. To achieve this, the following formula was applied;

$$y_t = \sum_s \frac{y_s . A_{st}}{A_s}$$

Here, y_t is the interpolated value of the census level variable at police precinct level t, y_s is the census variable value at SAL layer s, A_s is the area of SAL layer s and A_{st} is the area of the intersection between SAL and police precinct layers (Li *et al.*, 2007). The calculation proportionally assigns values from SAL to the police precinct layer based on intersected areas. The method presents problems for the estimation of variables of interest because it assumes that households exist everywhere and are equally and randomly distributed within police precincts (Jega, Comber and Tate, 2017).

The Union and Dissolve tools in ArcMap 10.5 were used to achieve the aggregation represented by the equation above. Areas in square kilometres for each data set were calculated. The Union tool was used to create new boundaries based on the intersection of the regions in the feature datasets. The new regions allowed for the calculation of the proportional area. The proportional area was used to calculate the values of the variables for each new region. The Dissolve tool was subsequently used to get the sum of the values at the precinct level.

One of the issues that arose while aggregating was that some regions did not have households. These were mostly industrial areas. Aggregating data without removing the areas with no household leads to an assumption that households exist everywhere in the study area. To minimise the effects of this assumption, regions with no households were removed from the estimation. Removal of regions with no households did not entirely overcome the pitfalls introduced by this estimation technique. Another related issue is that this technique assumes a random distribution of households within a polygon. This may not be true because even within a polygon; households could be clustered in certain parts than others due to natural landscape or man-made features. The area weighting technique was however used for convenience purposes.

3.4 Method of analysis

A combination of exploratory and explanatory analysis techniques was employed. The commonly used processes for spatial analysis of crime data include calculating descriptive statistics and quantification of spatial autocorrelation. Spatial autocorrelation is a term used to assess how attribute values relating to geographic phenomena are dependant based on proximity between data points or spatial units (O' Sullivan and Unwin, 2010). This is usually followed with an explanatory analysis using global regression techniques using ordinary least squares (OLS) regressions to model the phenomena under study. The regression models are usually followed with diagnostics testing of the OLS regression residuals and reanalysing the data using spatial regressions methods that take spatial dependence into account when evidence of spatial dependence is found. Spatial heterogeneity is a concept that assesses whether variation in spatial attribute data is constant throughout a study region. This is usually investigated using geographically weighted regression (GWR). GWR can be regarded as an extension to OLS regression when there is evidence of spatial heterogeneity in the data. The different methods and techniques are discussed in Sections 3.4.1 through to 3.4.8.

Statistical analysis was done in the R statistical computing environment (Core Team, 2017), GWR4.0 (Nakaya, 2016) and GeoDa 1.12.1.59 (Anselin, Syabri and Kho, 2004). ArcMap 10.4 (Environmental Systems Research Institute, 2014) was used for mapping requirements.

3.4.1 Exploratory Data Analysis and Explanatory Spatial Data Analysis

Exploratory data analysis (EDA) is an important stage in the statistical analysis which follows data collection and happens before confirmatory statistical processes (Leinhardt and Leinhardt, 1980). It is an important stage because it provides an opportunity to discover unusual patterns in the data and therefore aids in the understanding of real-world phenomena (Leinhardt and Leinhardt, 1980). Summary statistics (such as mean, median), graphical aids (such as scatter plots, histograms, and quartile plots), bivariate techniques (such as correlation analysis) and spatial visualization (using maps) are typically employed at the EDA stage. Most common statistical methods make assumptions about normal distribution of the variables under study. The aim of these preliminary analyses is to guide in selecting suitable statistical techniques that can be used to make accurate conclusions about the variables under study.

A scatter plot is a diagram that shows the relationship between two variables by plotting the dependant variable on the x-axis and the independent variable on the y-axis (Friendly and Denis, 2005).

Histograms display similar values of the attributes by the height of the bars on the y-axis and the groups of values of the attribute on the x-axis of the graph (Andrienko, Andrienko and Savinov, 2001). Histograms allow for visualisation of the distribution of the data and for checking of symmetry as well as possible outliers in the data.

The aim of the bivariate correlation analysis is to indicate the set of variables that are significantly correlated or associated with each other. The Pearson correlation coefficient is a measure of a possible linear association between two continuous variables (Hauke and Kossowski, 2011). A value of 0 indicates no linear association between the two variables in question while a value of -1 (+1) indicates either a perfect negative (positive) correlation between the variables. Values closer to +/- 1 would have a stronger correlation than those closer to zero. Results of the correlations are usually presented as a symmetrical matrix with off-diagonal values representing values of correlation between pairs of variables. The correlation amongst sets of explanatory variables is referred to as multicollinearity (Chatterjee and Hadi, 2006). Checking for multicollinearity is very important in OLS regressions because severe multicollinearity can lead to incorrect signs of the coefficients in OLS regressions leading to incorrect inferences. (Chatterjee and Hadi, 2006). Multicollinearity implies that an explanatory variable included in the model can be predicted by one or more of the included explanatory variables in the model. This (multicollinearity) has an impact that some variables have no significant explanatory power in the model, thereby over-specifying the model (Graham, 2003; Chatterjee and Hadi, 2006). Pearson correlation is usually employed as an initial means of assessing multicollinearity.

Maps are an essential part of Geographic Information Systems (GIS) and are one of the ideal and effective tools used to explore the general patterns and trends in the data (Chang, 2006). Choropleth maps are common in the presentation of aggregated data stored in polygons. An important factor in Choropleth mapping is the selection of classification methods employed for presenting data. The most commonly used classification methods for choropleth maps include; Equal Interval, Equal Frequency, Standard deviation, and Natural (Jenk's) breaks methods (Evans, 1977; Chang, 2006; O' Sullivan and Unwin, 2010). The equal interval method divides the data into equal intervals based on the number of classes specified by the analyst. The equal frequency classification method divides the sum of data values by the number of classes chosen and ensures that each class contains the same quantity of data values. The standard deviation method divides the class breaks into equal units that are proportions of the standard deviations above or below the mean. Natural (Jenk's) Breaks is a method that optimises grouping of values by using an algorithm that minimises variance within a class and maximises the variance between class breaks (Evans, 1977; Chang, 2006).

Evans (1977) states that for normally distributed data, the standard deviation classification method may be used. For multimodal data, the Natural (Jenks) break classification method may be used, while the geometric classification method may be used for highly skewed distributions as they are suitable for distributions that closely follow a J-shaped geometric progression (Evans, 1977). In terms of the number of classes to use in the classification of data, Evans (1977) suggests that 7 or 8 classes may be acceptable in most cases.

3.4.2 Spatial Effects

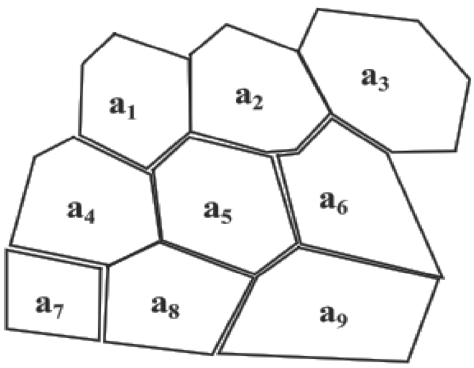
According to the first law of Geography; ‘everything is related to everything else, but near things are more related than distant things’ (Tobler, 1970). Bearing Tobler’s law of geography in mind, one aim of conducting exploratory spatial data analysis (ESDA) is to formally test if indeed patterns of clusters based on nearness are significant. The confirmation of the suspected patterns informs the selection of appropriate statistical tools to use.

There are two commonly known spatial effects relating to spatial lattice data. These are spatial autocorrelation and spatial heterogeneity. Spatial autocorrelation assesses how attribute values relating to geographic phenomena are dependant based on proximity between data points or spatial units (O’ Sullivan and Unwin, 2010). Spatial heterogeneity assesses whether the variation of spatial data is constant across the study area (Yan, 2007). Heteroskedasticity occurs when the variance in the error term of the OLS is non-stationary as the location of the observation changes (Osgood, 2000). The Breusch-Pagan (Anselin *et al.*, 1996) can be used to test heteroskedasticity.

The Moran’s I statistic, Moran’s I scatter plots (Figure 3.5), and the Local Indicators of Spatial Association (LISA) (Huo, Zhang, Sun, Li, Zhou, Li, 2011) are tools that were used for quantifying and visualizing spatial autocorrelation. With respect to the Moran’s I statistics, a positive value implies a spatial grouping of similar values while a negative value implies a spatial grouping of dissimilar observations (Voss *et al.*, 2006). Moran’s I statistic captures spatial autocorrelation at a global level; where the entire dataset is represented by a single value, while spatial autocorrelation at a local level highlights a specific area of clustering and is usually calculated using the LISA statistics and represented by LISA maps. Spatial weight matrices are used to specify the relationship between values based on how close spatial units are to each other.

3.4.3 Spatial weights matrix

The quantification of spatial autocorrelation depends on the relative effect neighbours have on each other (Voss *et al.*, 2006). This relative effect is referred to as the neighbourhood structure and is specified using a spatial weight matrix obtained either topologically (contiguity) or metric wise (distance-based) (Voss *et al.*, 2006, Okabe and Sugihara, 2012). Figure 3.4 shows the formation of the spatial weights matrix. For the contiguity based spatial weight matrix, neighbours in Figure 3.4 are defined to be polygons that share a boundary (such as a_1 and a_2) and a binary classification (with a value of 1 for neighbour and 0 for a non-neighbour) is used to specify this relationship. The weight matrix is usually row standardized making the element of each row sum to unity (Getis and Aldstadt, 2004, Voss *et al.*, 2006). The impact of row standardization is that it alters the size of the group impact on polygons with fewer neighbours (Kockelman, Wang and Wang, 2013).



	a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	**0**	1	0	1	1	0	0	0	0
a2	1	**0**	1	0	1	1	0	0	0
a3	0	1	**0**	0	0	1	0	0	0
a4	1	0	0	**0**	1	0	1	1	0
a5	1	1	0	1	**0**	1	0	1	1
a6	0	1	1	0	1	**0**	0	0	1
a7	0	0	0	1	0	0	**0**	1	0
a8	0	0	0	1	1	0	1	**0**	1
a9	0	0	0	0	1	1	0	1	**0**

Figure 3.4 Spatial contiguity matrix: queens case specification (Source: Bernasco and Elfers, 2011).

Distance-based weight matrices are formed by means of determining the distance between the centroid coordinates of the observation polygons (Getis and Aldstadt, 2004). The elements of the matrix, in this case, are the distances between the pairs of observation points. The elements along the diagonal are set to zero to indicate the distance of an observation point with itself. For inverse distance-based matrices, the distances are raised to the negative power of choice depending on how the neighbours are allowed to influence each other; for instance, in d_{ij}^{-6}, i and j are the pairs of neighbouring polygons within the threshold distance, d is the distance between the pairs of polygons (Getis and Aldstadt, 2004). The threshold distance is the distance between centroid coordinates of polygons whereby each polygon must have at least one neighbour.

Different specifications of weight matrices can lead to different inferences when modelling a geographical phenomenon (Kockelman, Wang and Wang (2013). Voss *et al.* (2006) tested different configurations of weight matrices and recommends selecting the matrix with the highest Moran's I value.

3.4.4 The Moran's I and the Moran's I Scatterplot

The Moran's I statistics capture spatial autocorrelation for the entire dataset as discussed in Section 3.4.2 above. The scatter plot shown in Figure 3.5 below is a tool that helps in visualising how observations cluster in the dataset.

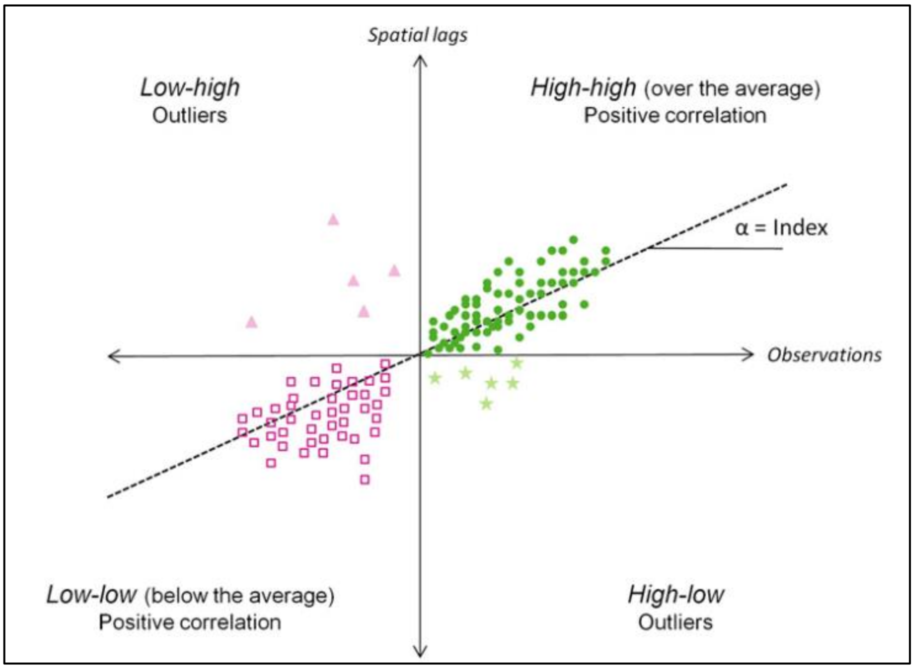


Figure 3.5: Moran's I scatter plot. (Source: Gómez, White and Wulder, 2011, page 1670)

The Moran's I scatter plot is a diagnostics tool that is used to assess spatial autocorrelation of an observed phenomenon. The observations are plotted on the x-axis and the spatial lags are plotted on the y-axis. It supplements the Moran's I statistic by allowing the visualisation of patterns in the clusters labelled low-high, high-high, high-low and low-low in the first, second, third and fourth quadrants respectively (Figure 3.5). In the Moran's I scatter plot, the sign of the slope of the graph shows whether there is positive (cluster of similar values) spatial autocorrelation or negative spatial autocorrelation.

3.4.5 Local Indicators of Spatial Association (LISA)

LISA is a local version of the Global Moran's I statistics and is used to indicate specific areas where spatial autocorrelation exists (Anselin, 1995). LISA is given by the following equation:

$$I_i = \frac{n(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

where *n* is the number of observations (spatial units), w_{ij} is the specified weights matrix for the definition of neighbours in the dataset at locations, $\imath$ and *j;* indexes for two neighbouring locations, $\bar{x}$ is the mean of *x* (Huo *et al.*, 2011). Results are presented using LISA map where different colours are used to indicate statistically significant clusters corresponding to; areas with high observations surrounded by other high observations (HH), low-high (LH), low-low (LL) and high-low (HL).

3.4.6 Ordinary least squares regressions and spatial regressions

Ordinary least squares (OLS) regression models the relationship between a dependant variable and a set of explanatory variables (Bernasco and Elfers, 2011). The basic formula for OLS is given by;

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for } i = 1 \ldots . n$$

Where y represents a vector of observations of the dependent variable and β_0 is the intercept parameter term and β_1 is the slope parameter for variable x_i, ε_i is the error term. The slope and intercept parameters are estimated in such a way that the value of; $\sum_{n=1}^{n}(y_i - \hat{y_i})^2$is minimised over the n observations. The term $(y_i - \hat{y_i})$ represents the residuals calculated by subtracting the estimated (fitted value) value from the i^{th} observation on the regression equation for that observation (Bernasco and Elfers, 2011). OLS models average relationship between a set of independent variables and a dependent variable. This relationship is generalised as an average and a single statistic (coefficient) is used (Matthews and Yang, 2012) on the dataset for modelling the association between the variables.

There are a couple of assumptions usually made about the OLS regression. Some of the common assumptions are:

a) The association between the dependant and independent variables is assumed to be linear.
b) The errors in the model are assumed to be normally distributed, uncorrelated or independent of each other; they have constant variance (homoscedastic) and they have a mean of zero.
c) The set of independent variables are linearly independent of each other (non-collinearity) (Voss *et al.*, 2006; Bernasco and Elfers, 2011).

When these assumptions are violated, conclusions made about the relationship and predictions will not be valid (Chatterjee and Hadi, 2006, Voss *et al.*, 2006). Violations can be tested using diagnostic tools; Cook's D plots scatter plots, histograms, box plots and various formal statistical tests such as the Moran's I correlation tests, the Breusch-Pagan tests of homoscedasticity, and the Jarque-Bera test statistics of normality of residuals (Chatterjee and Hadi, 2006, Voss *et al.*, 2006).

3.4.7 Diagnostics of regression residuals

Evidence of spatial autocorrelation in the residuals indicates that the variance of the dependent variable is not constant in the dataset. Therefore, OLS is not an appropriate model and an appropriate regression model that takes spatial autocorrelation into account may be required.

The Lagrange multiplier diagnostics tool (Figure 3.6) is commonly used as a guide for testing the type of spatial dependence present in the data.

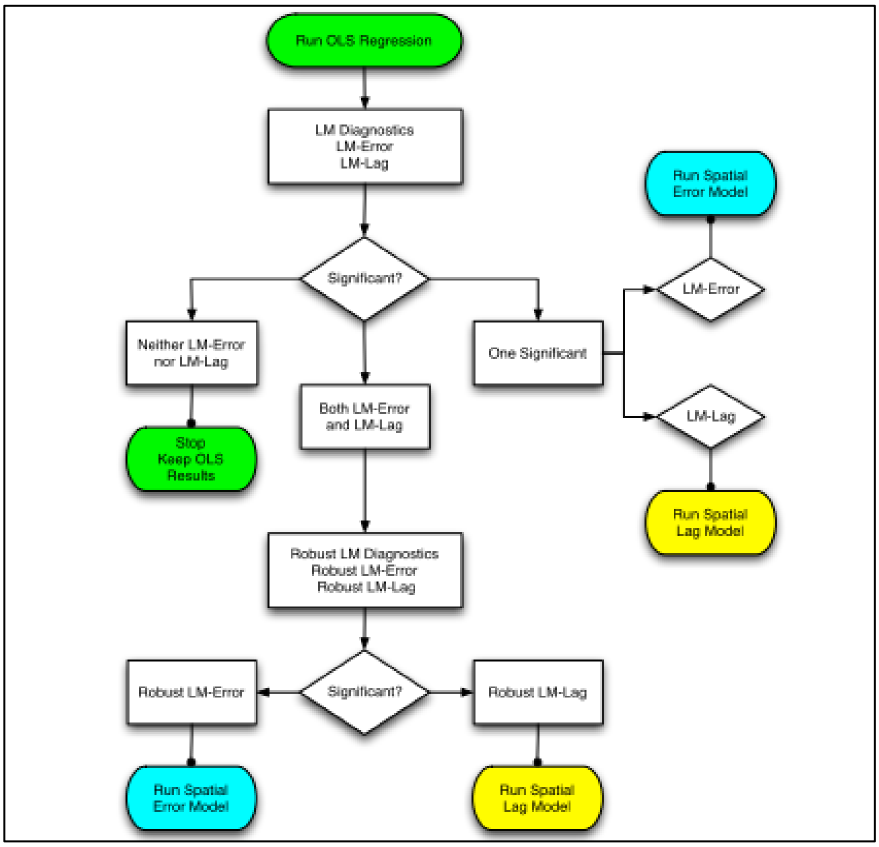


Figure 3.6 Spatial regressions diagnostics flow chart for decision making (Source: Anselin, 2005)

Firstly, the OLS regression is conducted followed by the Lagrange Multiplier (LM)-error and LM-lag diagnostics tests. When neither of the LM diagnostics is significant, the OLS regression results are retained. When one of the LM diagnostics is significant, the appropriate spatial regression is conducted. When these LM- error, and LM-lag are both significant, the robust LM-error and robust LM-Lag tests are conducted to select the best type of spatial regression model to explain the relationship between variables. The common forms of spatial regression models are Spatial Error Model (SEM) and Spatial Lag model (SLM) (Voss *et al.*, 2006).

The SEM model incorporates spatial effects (spatial autocorrelation) in the error term while the SLM incorporates the spatial effects in the lag term of the model (Anselin *et al.*, 1996, Voss *et al.*, 2006; Bernasco and Elfers, 2011). SEM regression caters for the spatial autocorrelation that is due to the set of unobserved explanatory variables (Bernasco and Elfers, 2011), for instance when crime at a location is being studied but not an exhaustive group of explanatory variables have been used in the equation of the regression (Voss *et al.*, 2006). When crime at a certain location is directly influenced by crime in the neighbouring location, the SLM regressions are appropriate (Voss *et al.*, 2006; Bernasco and Elfers, 2011).

3.4.8 Geographically Weighted Regressions (GWR)

GWR model is appropriate when there is significant non-stationarity in the data (Brunsdon, Fotheringham and Charlton, 2009; Bernasco and Elfers, 2011). GWR is a local regression technique that models spatial heterogeneity at every point in the study area by means of fitting a regression model for each spatial unit and taking neighbourhood structure into account - by means of weighting observations from neighbours as a function of distance from such point (Guo, Ma and Zhang, 2008).

GWR can be expressed as;

$$y_i = \beta_{0i}(u_i, v_i) + \sum_{n=1}^{k} \beta_{ni}(u_i, v_i) x_{ni} + \varepsilon_i$$

where y_i is the dependant variable at a certain location, (u_i, v_i) represent the coordinates of the location, *i* of the variable under study. For lattice data, this is usually the centroid coordinates of the polygon. β_{oi}, β_{ni} represent, respectively, the local computed intercept (β_{oi}) and effect (β_{ni}) of variable n for the explanatory variables on the dependant variable (Fotheringham, Brunsdon and Charlton, 2002). In the equation, *i* represents an index for a certain location and 0 represents the intercept value. Standard errors for each explanatory variable in the GWR are used to label a map to assess non-stationary or spatial "drift" across the study area.

The estimation of the term(β) in the model incorporates the neighbourhood structure in the weights matrix, *W*. (Brunsdon, Fotheringham and Charlton, 2009). The neighbourhood structure is expressed in terms of the kernel function and a bandwidth (Guo, Ma and Zhang, 2008). Each element in the weight matrix, when a Gaussian kernel is used, is expressed in the following manner (Brunsdon, Fotheringham and Charlton, 2009);

$$w_{i,j}(u_i, v_{j)} = e^{-0.5\left(\frac{d_{i,j}(u,v)}{h}\right)^2}$$

Where w is the weight of observation in the matrix, $d_{i,j}(u, v)$ a distance between i^{th} *and* j^{th} observations and h is referred to as the bandwidth. Another common kernel type is the bi-square kernel (Matthews and Yang, 2008).

The bandwidth is a distance along a radius length from the point of calculation or a number of points (centroid coordinate) (Guo, Ma and Zhang, 2008). The function can be chosen as fixed (nearest neighbour or distance) or adaptive (nearest neighbour or distance) (Brunsdon, Fotheringham and Charlton, 1998). Guo, Ma and Zhang (2008) indicate that GWR models with smaller bandwidths fit the data better. Guo, Ma and Zhang (2008) also indicate that a fixed kernel function produced smother spatial distribution than adaptive kernel functions. The two types of kernel functions, adaptive or fixed, may be used and the model which has the least AIC value is selected.

The results of GWR may be presented by *t*-values for the coefficients wherein statistical significance may be tested. The authors of the GWR software package did not include the calculation of *p*-value which is common for significance testing thus *t*-values are used instead. The *t*-values may be used in set of choropleth maps for the chosen variables classified to significance levels such as 1%, 5%, and 10% so that regions of statistical significance may be assessed (Fotheringham, Brunsdon and Charlton, 2002; Malczewski and Poetz, 2005; Breetzke, 2012; Chen *et al.,* 2017). The assessment of the significance of the variables in GWR was tested by comparing the critical values of the *t* distribution with more than 30 degrees of freedom at 5% and 1% significance levels. The critical values in the *t* distribution table with more than 30 degrees of freedom for a two-sided hypothesis testing correspond to $-1.96 \leqslant t \leqslant 1.96$ at 5% level and $-2.58 \leqslant t \leqslant 2.58$ at 1% level.

4 Results

4.1 Introduction

Summary statistics, plots, bivariate correlations results, and choropleth maps are presented in Sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4 respectively. The specifications of the spatial weight matrices are presented in Sections 4.3.1, followed by the results of Moran's I tests for spatial autocorrelation in Section 4.3.2 and LISA maps in Section 4.3.2.1.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Summary statistics and plots for dependant variables

The summary statistics allow for an assessment of the statistical properties (centre and spread) of variables. Summary statistics are presented in Table 4.1.

Table 4.1 Summary Statistics: Dependent variables

	Min.	Max.	1st Qu	3rd Qu	Mean	Median	Stdev	Skewness	Kurtosis
Burglary Rates	4.35	105.38	11.42	26.43	20.45	17.31	13.25	2.66	12.21
Log of Burglary Rates	1.47	4.66	2.43	3.27	2.85	2.85	0.57	0.07	-0.01
Square root of Burglary rates	2.09	10.27	3.38	5.14	4.34	4.16	1.28	1.10	2.73

Note 1: 1st Qu and 3rd Qu represent first and third quartiles respectively and Stdev represents standard deviation. The standard deviation measures the average dispersion (distance) of the observations from the sample mean. Burglary rates and log and the square root of burglary rates are measured per 1000 households.

Note 2: the Skewness value for a normal distribution is equal to zero and a skewness value greater than 2 is considered severe (Kim, 2013). A negative skewness implies that the distribution has a median value greater than the mean values (left-skewed) compared to a normal distribution while a positive skewness implies that distribution is right-skewed (median less than the mean values) compared to a normal distribution.

Note 3: Kurtosis is used to assess the magnitude of outliers in the distribution. Kurtosis for a normal distribution is equal to 3.

The difference between mean and median of the dependent variable (burglary rate per 1000 households) suggested a deviation from symmetry. In Table 4.1, the burglary rates in Gauteng were highly positively skewed because the skewness value was 2.7. Excess skewness of burglary rates is common (Breetzke, 2012). Excess positive skewness suggested that the sample mean value for burglary rates was greater than the sample mean value if observations were sampled from a normal distribution. This right-skewed distribution indicated that there were high (extreme) positive burglaries rates present in the data. The existence of these extreme values could also be an indication of outliers. The dependent variable (Burglary rates) was both log-transformed (using the natural log with a base of *e*) and square root transformed so that it conforms to the normal distribution and this is an assumption made in the OLS regression that will be carried out using the data. The kurtosis value for the untransformed burglary rate per 1000 households was very high (greater than 3) indicating severe kurtosis. In contrast, the log-transformed burglary rates had kurtosis value close to 0, indicating closeness to the normal distribution. Square root transformed burglary rates had higher kurtosis and skewness values than the log-transformed version of the variable. The log-transformed variable closely conformed to the normal distribution than the untransformed and square root transformed burglary rates.

Box and whisker plots, histograms and normal quantile-quantile plots that complement the summary statistics are presented in this section. The summary plots for the burglary rates per 1000 households are presented in Figure 4.1. The box and whisker plots on the top row in Figure 4.1 indicate the spread for the data. The whiskers are the dotted lines between quartile values and minimum and maximum values. The end of the bottom whisker represents the lowest or minimum observation that is not an outlier and similarly, the end of the whisker on top represents the maximum observation that is not an outlier. The line at the bottom of the box indicates the first quartile and the line at the top represents the third quartile. The bold line in the middle of the box represents the median value in the observations. Where the length (measured by the difference between the quartile value, 1st or 3rd and minimum or maximum) of the whiskers were equal it implied that the mean and median values were equal. The dots on the box and whisker plots depict outliers in the dataset.

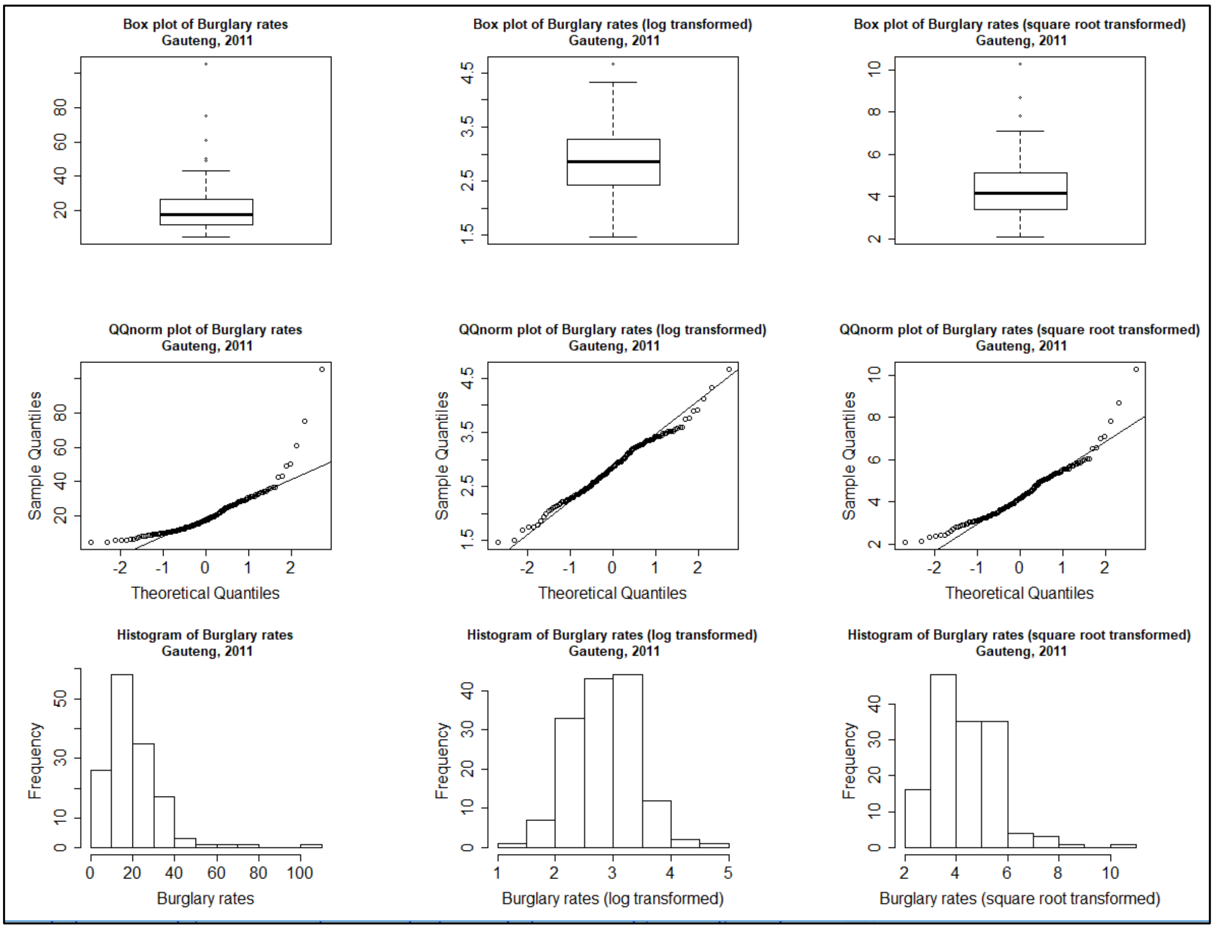


Figure 4.1: Box plots, normal quantile-quantile plots and histogram plots of the Burglary rates variables

The box and whisker plot for burglary rates per 1000 households displayed a lack of symmetry and the presence of outliers. The box and whisker plot for burglary rates per 1000 households (log-transformed) indicated symmetry and the presence of a single outlier. The box and whisker plot for burglary rates per 1000 households (square root transformed) indicated asymmetry suggesting the variable was skewed to the right. In addition, there were outliers on the box and whisker plots for the burglary rates per 1000 households.

The quantile-quantile plots on the second row in Figure 4.1 were also used to judge how closely the observations follow a normal distribution. The values on the x-axis are the theoretical quantile values of a normal distribution and the values on the y-axis indicate the sample or observation quantile values. Where the observations were normally distributed, the quantiles were on the 45-degree line (normal line).

The log-transformed burglary rates per 1000 households were more symmetric than the other two variables (untransformed and the square root transformed burglary rates per 1000 households) because most sample quantiles fell closer to the normal line especially in the tails compared to the other two cases.

The histograms are indicated in the third row in Figure 4.1. Burglary rates per 1000 households (transformed and untransformed) appeared to be skewed to the right because of the long tails on the right of the distributions. The histograms for the log-transformed burglary rates shown symmetry because the shape formed by the frequency bars closely followed the bell-shaped curve of a normal distribution curve.

The plots in Figure 4.1 provided preliminary evidence that the log-transformed burglary was better qualified for further analysis due to its symmetry as observed from the box and whisker plots as well as the quantile-quantile plots. Thus, the log-transformed burglary rate was chosen for further analysis.

Formal statistical tests using the Shapiro-Wilk Test were performed to check the validity of the suspected normality. The formal hypotheses for this test were set up as follows:

H_0: Log transformed burglary rate follows a normal distribution

H_1: Log transformed burglary rate does not follow a normal distribution

The Shapiro-Wilk Test statistic was 0.99 with a *p*-value = 0.5 which is greater than the level of significance alpha thus the null hypothesis was not rejected. The log-transformed burglary rate was therefore assumed to follow a normal distribution and chosen for further analysis.

4.2.2 Summary statistics and plots for the independant variables

Summary statistics are presented in Table 4.2.

Table 4.2: Summary statistics for independent variables

	Min.	Max.	1st Qu	3rd Qu	Mean	Median	Stdev
Percentage unemployed	1.58	20.16	6.60	16.39	11.24	11.03	5.39
Percentage of males aged 15 to 34 years	11.51	33.06	18.83	21.94	20.65	19.90	3.63
Percentage of households with no income	0.00	36.50	14.74	22.33	18.65	19.16	5.91
Percentage of female headed households	18.81	45.82	30.51	37.81	33.74	33.99	5.97
Percentage of population with BSc	0.08	10.83	0.52	3.35	2.48	1.42	2.66
Percentage of White-headed households	0.03	76.63	0.14	39.98	21.09	11.39	23.75
Population Density(people/km²)	5.16	24799.84	539.43	3941.53	3063.94	1803.04	3799.92
Percentage of people that moved in the last 5 years	0.90	23.10	2.21	5.90	4.66	3.43	3.75
Percentage of population with no citizenship	1.54	38.16	4.08	10.24	7.93	5.90	6.01
Ethnic Heterogeneity Index	0.01	0.76	0.02	0.66	0.39	0.50	0.29
Roads density(m/km²)	227.91	19558.00	2638.01	11162.65	7451.89	7741.97	4825.23

Histograms for the explanatory variables indicated that most variables were positively (right) skewed except for the percentage of males aged 15 to 34 years, percentage of households with no income and percentage of female-headed households which were more symmetric. Ethnic heterogeneity index was found to be bimodal; the percentage of the population unemployed and roads density was multimodal. Multimodal variables have samples from inherently different populations (Evans, 1977). The lack of symmetry and the multimodal nature of the independent variables were not addressed because OLS makes assumptions of symmetry of the dependant variable only.

4.2.3 Bivariate correlation statistics

The scatter plots showing the nature of the relationship between the log-transformed burglary rates and the independent variables are indicated in Figure 4.2. The red line fitted on the scatter plot represents the regression line of best fit based on the two variables indicated in the scatter plots (Figure 4.2). The importance of the best-fit line in scatter plots is that it helps to judge the sign (positive or negative) of the relationship between the two variables. The results of the correlations are presented in Table 4.3.

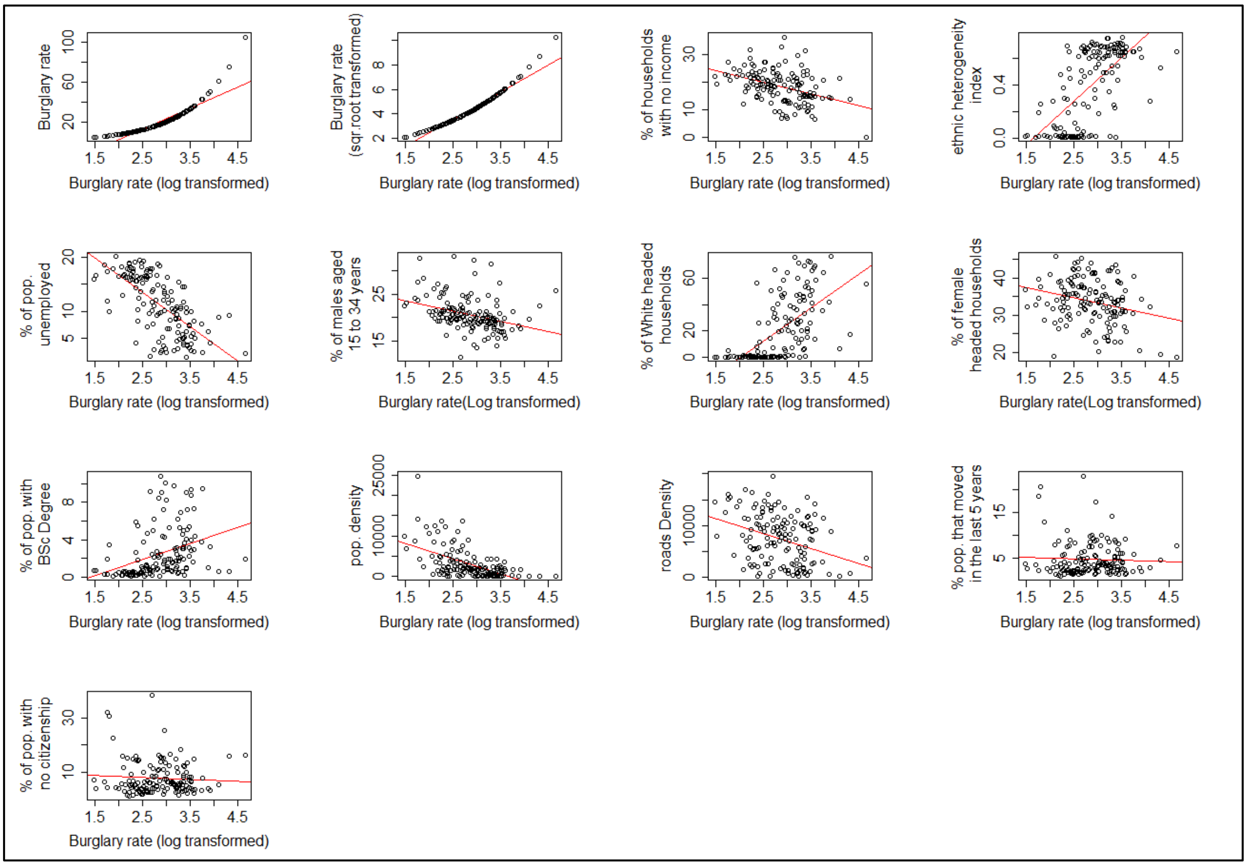


Figure 4.2 Scatter plot between burglary rates and independent variables

Table 4.3: Bivariate correlation matrix

	Variable Description	A	B	C	D	E	F	G	H	J	K	L	M	N	P
A	Roads Density	**1**													
B	Burglary Rates per 1000 households	**-0.29**	**1**												
C	Burglary Rates(Log transformed)	**-0.34**	**0.90**	**1**											
D	Burglary Rates(Square root transformed)	**-0.32**	**0.97**	**0.98**	**1**										
E	Percentage of population unemployed	0.15	**-0.57**	**-0.67**	**-0.64**	**1**									
F	Percentage of males aged 15 to 34 years of age	0.11	**-0.20**	**-0.34**	**-0.28**	**0.47**	**1**								
G	Percentage of households with no income	0.07	**-0.40**	**-0.40**	**-0.41**	**0.64**	**0.56**	**1**							
H	Percentage of female headed households	**0.46**	**-0.34**	**-0.27**	**-0.31**	**0.20**	-0.16	0.02	**1**						
J	Percentage of population with a BSc Degree	**0.17**	**0.25**	**0.37**	**0.33**	**-0.79**	**-0.39**	**-0.57**	**0.16**	**1**					
K	Percentage of White headed households	-0.06	**0.52**	**0.61**	**0.59**	**-0.88**	**-0.54**	**-0.58**	**-0.22**	**0.73**	**1**				
L	Population Density	**0.74**	**-0.42**	**-0.55**	**-0.50**	**0.51**	**0.34**	**0.34**	**0.27**	**-0.26**	**-0.41**	**1**			
M	Percentage of population that moved in the last 5 years	0.09	-0.02	-0.05	-0.04	**-0.23**	**0.49**	0.09	-0.14	**0.27**	0.02	0.01	**1**		
N	Percentage of population with no citizenship	0.09	0.01	-0.06	-0.03	**-0.21**	**0.44**	0.07	**-0.18**	**0.24**	0.02	0.02	**0.96**	**1**	
P	Ethnic heterogeneity index	**-0.19**	**0.51**	**0.65**	**0.60**	**-0.85**	**-0.36**	**-0.42**	**-0.35**	**0.64**	**0.79**	**-0.52**	**0.22**	**0.20**	**1**

There was a negative relationship between burglary rates (log-transformed) and percentage of the population with no citizenship and percentage of the population that moved in the last 5 years. There was a positive association between the burglary rates per 1000 households and percentage of the population with BSc degree, percentage of White-headed households and ethnic heterogeneity index. There was also a negative relationship between the burglary rates (log-transformed) and percentage of female-headed households, percentage of the population unemployed, percentage of males aged 15 to 34 years of age, population density, percentage of households with no income and roads density. The scatter plot for ethnic heterogeneity index shows 2 clear clusters in the graph. The possible cause for these two clusters may be the fact that in Gauteng, there were many regions which were completely comprised of only one race, where the index will be 0 and other regions that have mixed races where the ethnic heterogeneity index will be approaching 1. Pearson correlation coefficient was obtained to quantify the bivariate relationships indicated in Table 4.3 and to test the significance of the correlation between a set of variables.

In Table 4.3, values indicated in bold text indicated statistically significant Pearson correlation coefficient between the pair of variables (p-value less than 0.05 significance level). The correlation coefficient is on a scale between 0 (no linear) and 1 (perfect linear association) and the positive and negative sign indicate the nature of the linear association.

The correlation coefficients presented in Table 4.3 show a moderately (+/- 0.5) positive correlation between log-transformed burglary rates per 1000 households and ethnic heterogeneity index as well as the percentage of White-headed households. There was a strong negative correlation between log-transformed burglary rates per 1000 households and the percentage of population unemployment and population density. The strong negative correlation between burglary rates (log-transformed) and population density imply that high population density in a neighbourhood was a deterring factor for burglary to occur in an area. Female-headed households and the percentage of males aged 15 to 34 years of age have a low negative correlation coefficient with the burglary rates per 1000 households (log-transformed). There was no statistically significant correlation between log-transformed burglary rates per 1000 households with the percentage of the population with no citizenship and percentage of the population that moved in the last 5 years.

The correlation matrix also gives an indicator of the nature of the relationship between pairs of independent variables. The pairs of independent variables that had strong positive significant correlation coefficients with ethnic heterogeneity index were the percentage of White-headed households, percentage of the population with a BSc degree. Ethnic heterogeneity index had a strong negative correlation coefficient with the percentage of the population unemployed and a moderate negative correlation coefficient with population density. Percentage of the population unemployed had a strong significant negative correlation coefficient with the percentage of White-headed households, percentage of the population with a BSc degree and ethnic heterogeneity index. There was a moderate positive correlation between the percentage of the population unemployed and population density.

Roads density had a strong positive correlation with population density. The strong positive correlation (0.74) between roads and population densities was expected (Andresen, 2006).

Percentage of population with no citizenship had a strong positive correlation coefficient with the percentage of the population that moved in the last 5 years of the census survey. This correlation implied areas where people had no citizenship were also regions where people migrated to in the last 5 years of the census survey. Percentage of population with BSc degree had a positive correlation coefficient with the percentage of White-headed households. This correlation implied that White-headed households were mostly associated with higher education levels.

In summary, independent variables that were moderate to highly (greater than +/- 0.5 to +/- 1) correlated to the log-transformed burglary rates were; the percentage of people unemployed, percentage of White-headed households, population density and ethnic heterogeneity index. These variables were also correlated with each other. The correlation between a set of independent variables may change the signs of the parameter estimates of the OLS regression, thus one must be careful in selecting the variables for regression purposes (Graham, 2003). Andresen (2006) states that the extent of correlation exceeding 0.8 between independent variables used in regressions is cause for major concerns. In the results presented, the percentage of the population unemployed has a strong negative (-0.88) correlation with percentage of White-headed households and ethnic heterogeneity index (-0.85).

4.2.4 Visualization of spatial patterns

Choropleth maps were employed to explore the data further. Choropleth maps aid in assessing the existence of non-randomness or spatial clustering of the phenomena being studied. Randomness is evident when there are no visible clusters of similar or dissimilar values. The choropleth maps are presented in Figures 4.3 to 4.6. The maps showing the spatial patterns relating to the independent variables are presented in units of counts per police precinct area, otherwise so indicated by the standard deviation (SD).

Authors suggest several methods for presenting information on Choropleth maps. These maps may be produced by taking into consideration of two properties of the data; the statistical properties of data or statistical and spatial properties of data simultaneously. Spatial properties relate to the proximity between polygons while statistical properties relate to the nature of the distribution of the underlying data (Evans, 1977; O' Sullivan and Unwin, 2010). For this report, the statistical distributions were considered without the spatial properties only for convenience purposes.

The maps presented in this section have been classified by means of the various classification methods: Standard Deviation for normally distributed data and the Natural (Jenks) break classification method for multimodal data. Variables that were not significantly different (visually on the histograms presented in Section 4.2.2) from a 'bell-shaped' curve were classified using the SD method, these variables were; the percentage of female-headed households, percentage of households with no income, percentage of males aged 15 to 34. Seven variables; percentage of population that moved in the last 5 years, percentage of population with a BSc degree, percentage of population with no citizenship, ethnic heterogeneity index, percentage of population unemployed, percentage of White-headed households, population density, and roads density were classified using the Natural (Jenk's) break method because some appeared to be bimodal, and some were skewed to the right. In terms of the number of classes to use in the classification of data, Evans (1977) suggests that 7 or 8 classes are acceptable; hence the maps presented here have up to 8 classes.

Figure 4.3 shows the choropleth map for burglary rates for the untransformed and log-transformed values. The numbers in brackets in Figure 4.3 and all maps in this section represent the count of values per range/class. The year 2011 was used in the caption for all the maps presented in this research to signify when the data used was captured (by SAPS and StatsSA).

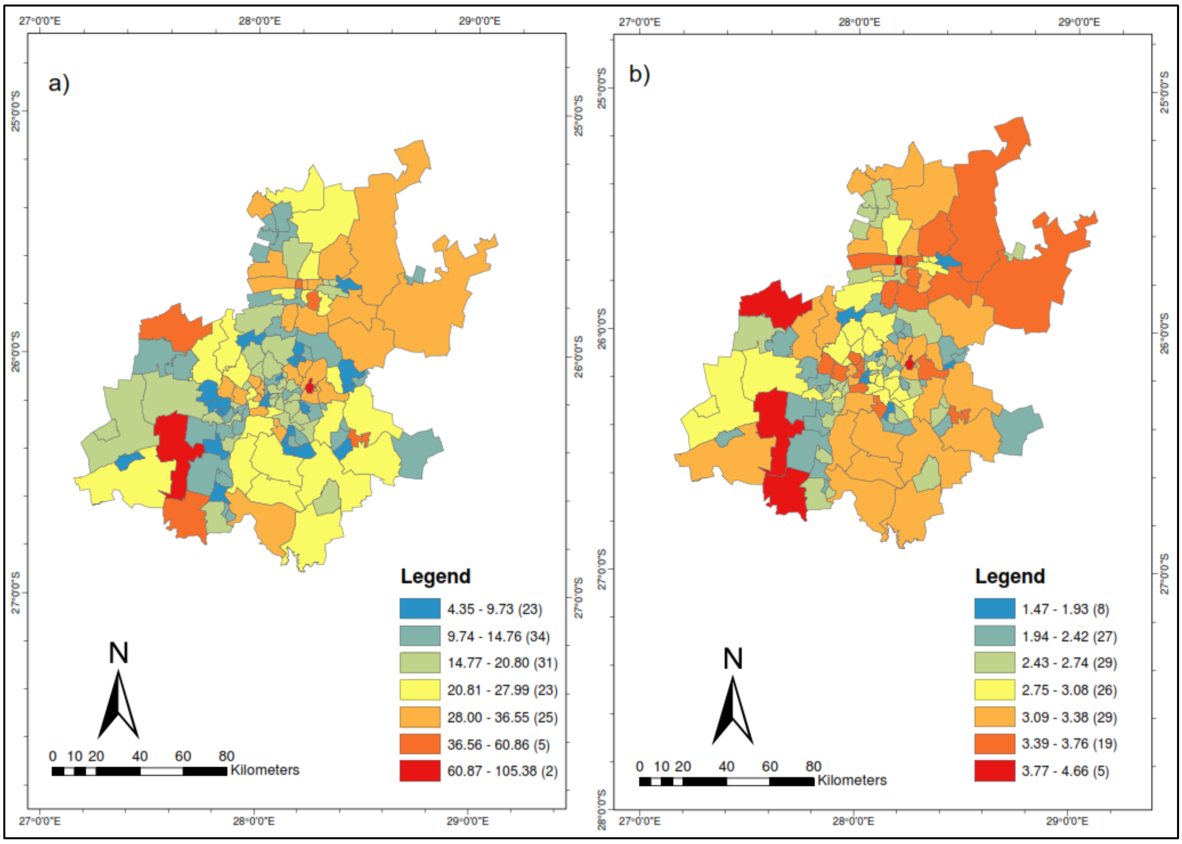


Figure 4.3: Choropleth map of a) burglary rates (untransformed), b) burglary rates (log transformed) per 1000 households in 2011. (Source: StatsSA and SAPS)

Burglary rates of the similar cluster in the North-Eastern, Southern and Eastern parts of the study region were observed. Two regions had a standard deviation greater than 2.5 from the mean; these are outliers in the dataset. It was deduced from the map in Figure 4.3 that burglary rates were not random; there were similar values clustering together.

The maps showing the spatial patterns relating to the independent variables are presented in Figures 4.4 to 4.6 in percentages or otherwise so indicated by SD for standard deviation. The maps labelled (a) to (d) in Figure 4.4 suggested that patterns were not random for all the variables. Specifically, map (a) showing the percentage of people that moved in the last 5 years, clustering of values in the central parts of the study region were observed. A similar pattern was observed in the map (d); the percentage of the population with a B.Sc. degree, with the clustering of values in the central parts of the study region. Maps (b) and (d) in Figure 4.4 show clustering in the central and South-Western regions of the study region. In the map (b) there were clusters of similar values on the Eastern, South-Eastern and Western parts of the study region. The percentage of the population unemployed, map (c) depicted highest percentages

(20.16 %) of people unemployed mostly in the South-Western, North-Western regions of the study area as It appeared that the central parts of the study region had the highest percentages of people with a degree and low percentages of people with no income. These regions in the central parts of the study area also had high percentages of people that moved in the last 5 years; between 6.79% and 13.01%.

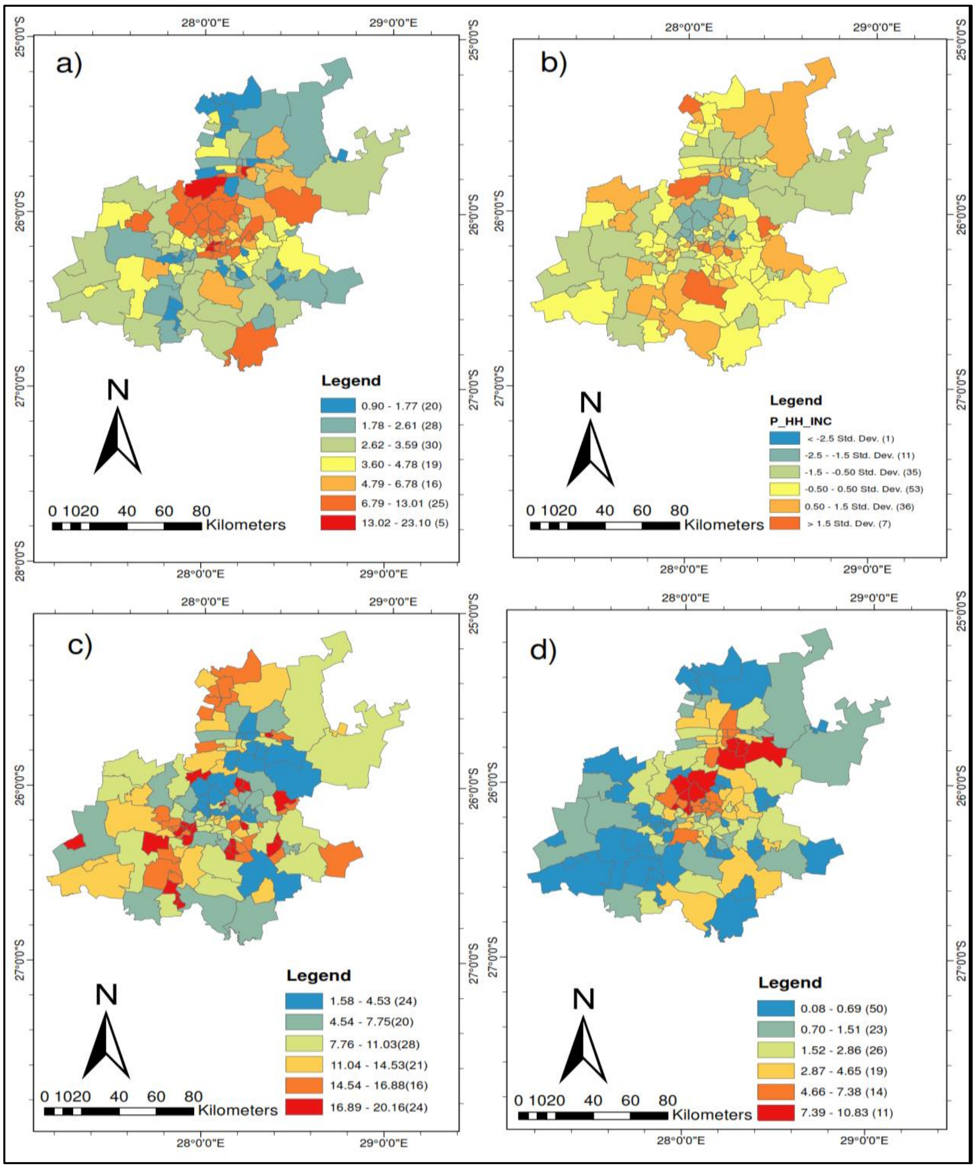


Figure 4.4: Choropleth maps of percentage of (a) people that moved in the last 5 years, (b) households with no income, (c) people unemployed and (d) population with BSc. Degree in 2011. (Source: StatsSA and SAPS)

Maps in Figure 4.5 show the spatial patterns for the percentage of White-headed households, (b) population density, (c) percentage of female-headed households and (d) road density. There were similar patterns of clusters for maps (b) and (d) where similar values were found in the central and upper central parts of the study region. There were low percentages of White-headed households in the South-Western, Eastern and northern parts of the study region. These (South-Western, Eastern and northern) regions also had a high concentration of percentages of households with no income in the map (b). This implied that regions with low percentages of white-headed households were characterised by high levels of no income and unemployment. Female-headed households (map (c)) depicted random patterns.

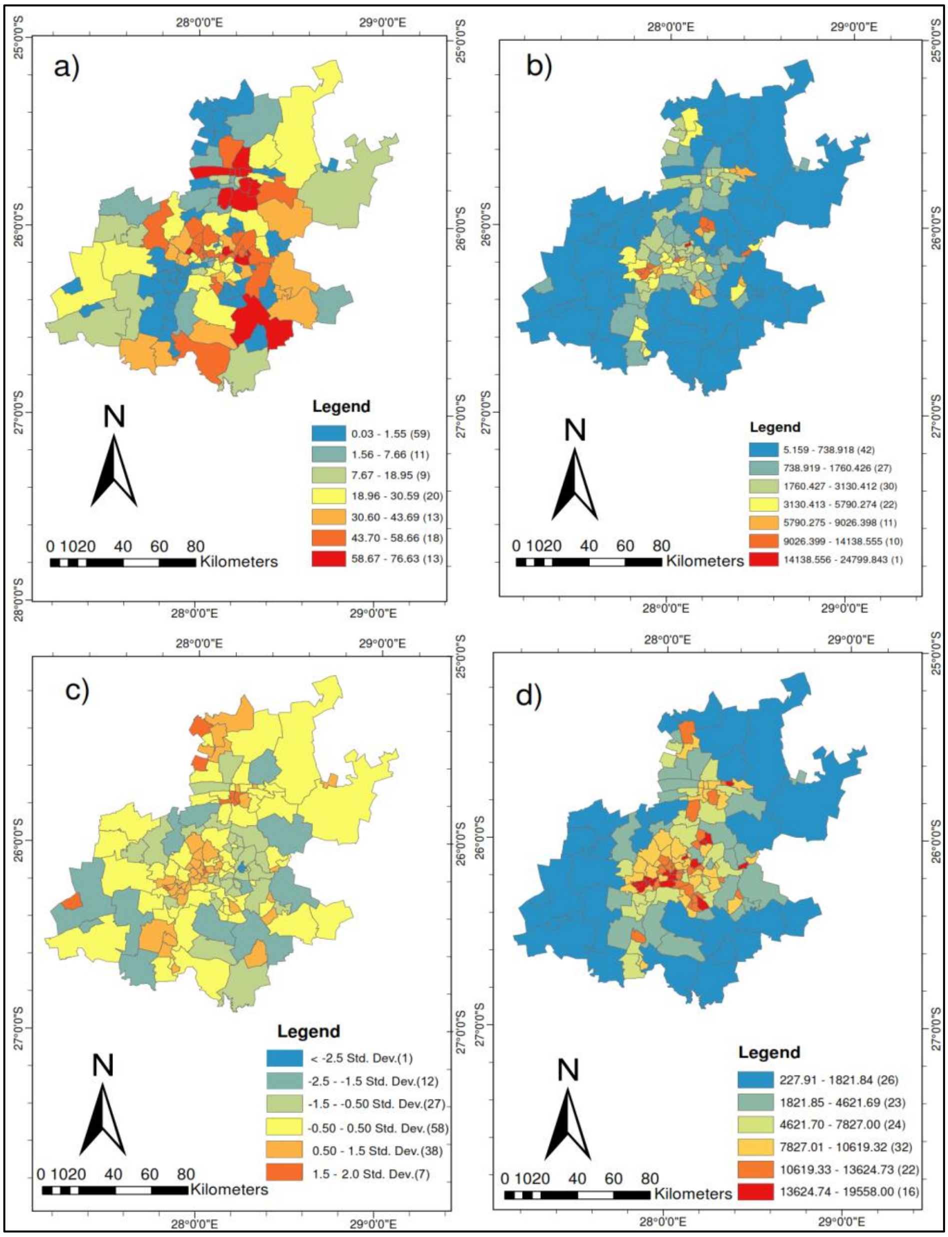


Figure 4.5: Choropleth maps for percentage of (a) White headed households, (b) population density (people/km^2), (c) percentage of female headed households (represented as SD) and (d) road density (m/km^2), 2011 (Source: StatsSA and SAPS)

Choropleth maps for ethnic heterogeneity index, percentage of the population that moved in the last 5 years and percentage of the population with no citizenship are indicated in Figure 4.6. There were clusters of similar values in the central and South-Eastern parts of the study area for ethnic heterogeneity index. Similar types of clusters were observed in maps (b) and (c) in Figure 4.6. These were clusters of similar values in the central parts of the study region.

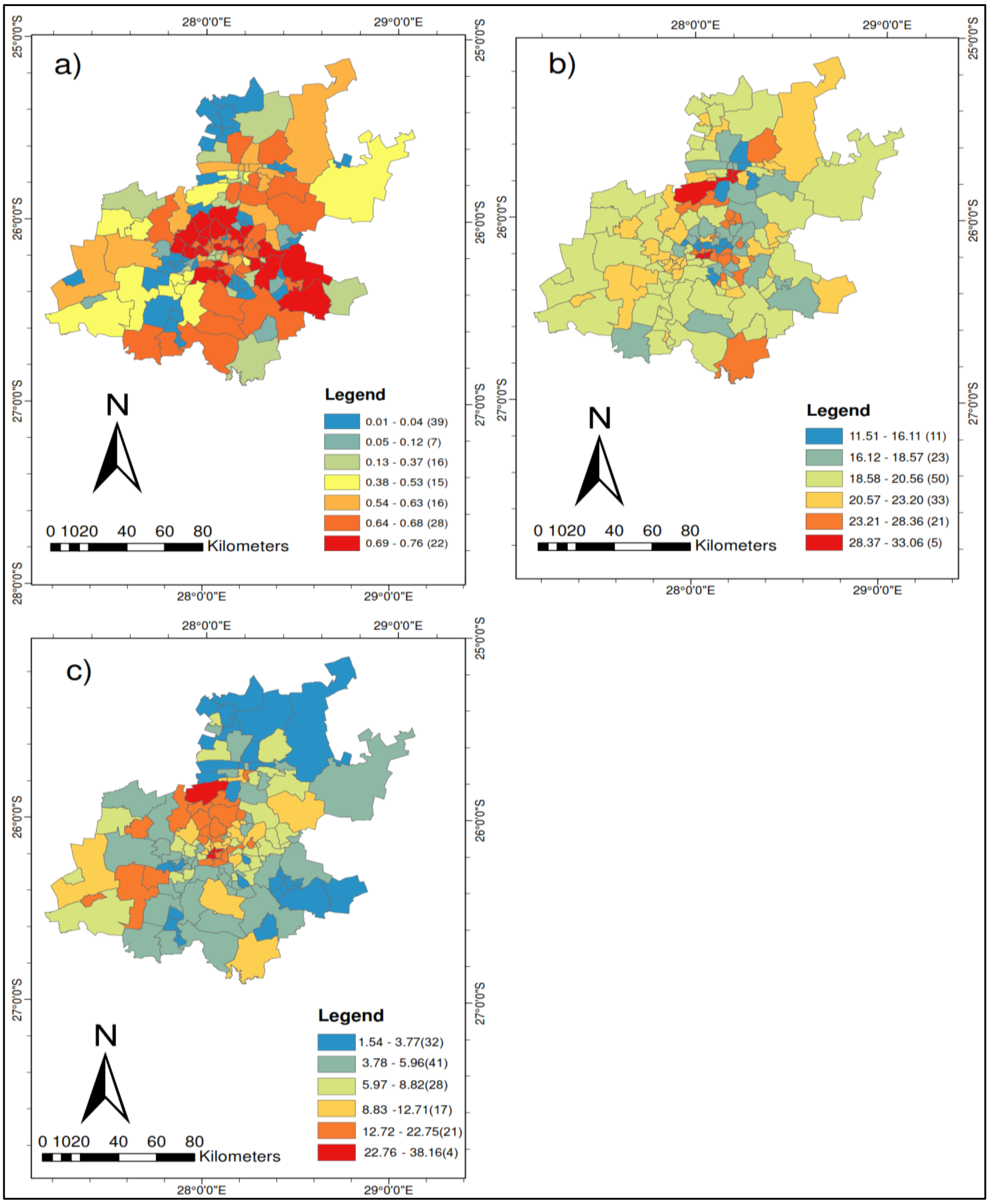


Figure 4.6: Choropleth maps for (a) ethnic heterogeneity index, (b) percentage of males aged between 15 and 34 years and (c) percentage of population with no citizenship, 2011 (Source: StatsSA and SAPS)

The maps presented in this section depicted the spatial distribution of the variables at the police precinct level. Generally, the results indicated that the variables chosen in this report were not random in Gauteng. There were clear variations of employment and income variables. Areas that had high concentrations of white-headed households had low unemployment concentrations. Road densities were lower on the periphery of the study region compared to the central regions which were also characterised by high population density.

4.3 Explanatory Spatial Data Analysis (ESDA)

4.3.1 Spatial weights matrix

As in Voss *et al.* (2006), two specifications of the spatial weight matrix were formulated and tested; queens case contiguity specification and inverse distance weighting.

For the distance-based weight matrix, a power of negative 6 and a threshold distance of 25.27 km was chosen to increase the influence nearest neighbours have on each other (Voss *et al.*, 2006). In a study, an inverse distance weighting can be formulated, with many possible combinations, however, an optimal threshold distance whereby neighbours are defined is important for a variety of reasons. One of the reasons is the computing power and another is theoretical, such as Tobler's First Law of Geography (Tobler, 1970). The negative power of 6 was chosen because "the higher the exponent the greater the influence of neighbouring cells as opposed to a greater distance between cells" (Getis and Aldstadt, 2004). The threshold distance is the distance between centroid coordinates of polygons whereby each polygon in the dataset must have at least one neighbour. In simple terms, a threshold distance is the maximum cut off distance whereby neighbouring polygons will not be influenced by polygons located beyond this search radius of 25.27 km. The lack of prior knowledge regarding how burglaries in neighbourhoods of Gauteng were affected by distance influenced the decision to use the default value determined by in ArcMap as well as GeoDa. The matrix was row standardised to minimise the group effect neighbours have on regions with lesser neighbours (Kockelman, Wang and Wang, 2013).

The queen's case first-order contiguity matrix was computed where polygons that touch on vertices and or boundaries were considered as neighbours and a binary classification with 1 for the neighbour and 0 for the non-neighbour was used for the matrix.

Summary statistics for the queen's case contiguity and inverse distance weight matrices are presented in Table 4.4. The queen's contiguity matrix specification had a smaller number of neighbours compared to the inverse distance weight matrix. Griffith (1996) as quoted by Getis and Aldstadt (2004) states that it is better to underspecify the weights matrix (fewer neighbours) in order to increase the power of the tests. Power of the tests refers to the "probability of rejecting the null hypothesis when an alternative hypothesis is true" (Waller, Hill and Rudd, 2006). The queen's case contiguity matrix was used for this report. Although the suggestion is that the weight matrix with fewer neighbours must be used, the weight matrix that offers the highest Moran's I value and the strongest statistical significance were recommended (Voss *et al.*, 2006). The results of the Moran's I test results are presented in Section 4.3.2.

Table 4.4 Spatial weight matrices summary statistics

	Contiguity (Queens's case) Matrix Summary	Distance (Inverse Distance Weight) Matrix Summary
Number of Features:	143	143
Percentage of Spatial Connectivity:	3.86	19.26
Average Number of Neighbours:	5.52	27.54
Minimum Number of Neighbours:	1	1
Maximum Number of Neighbours:	10	57

4.3.2 Moran's I Tests for Spatial Autocorrelation

The Moran's I statistic test was performed on the log-transformed burglary rate using both the queen's case contiguity matrix and the inverse distance weight matrix.

Hypothesis tests were set up as follows;

H_0: There is no spatial autocorrelation for log-transformed burglary rate

H_1: There is spatial autocorrelation for log-transformed burglary rate

The Moran's I index under the queen's contiguity matrix is 0.12 with a z-score of 2.40, and a *p*-value of $p=0.016$. The *p*-value for the test statistic was less than $\alpha=0.05$. Thus, the null hypothesis was rejected, and it was concluded that there was spatial autocorrelation for log-transformed burglary rate. The Moran's I value of 0.12 indicated positive spatial autocorrelation, that is, similar values clustered together in space.

The test for spatial autocorrelation using Moran's I test under the inverse distance weight matrix specification was also conducted. The Moran's I statistic was found to be -0.04 with a z-score of -1 and a *p*-value p =1.6. The null hypothesis was not rejected because the *p*-value was greater than the significance value of 0.05. The conclusion under this specification was there was no spatial autocorrelation of burglary in Gauteng.

The *z*-score for the queens-based weight matrix for the Moran's I statistic was greater than the *z*-score for the inverse distance-based weight matrix. The higher *z*-score indicated higher statistical significance (Voss *et al.*, 2006). The queen's case weight matrix was thus chosen for further analysis due to the higher *z*-score value and the higher Moran's I statistic.

4.3.2.1 Moran's I scatter plot of spatial autocorrelation

Moran's I plot allows for an assessment of types of clusters in the study region. Types of clusters can be such that similar values cluster together or dissimilar values cluster together in space.

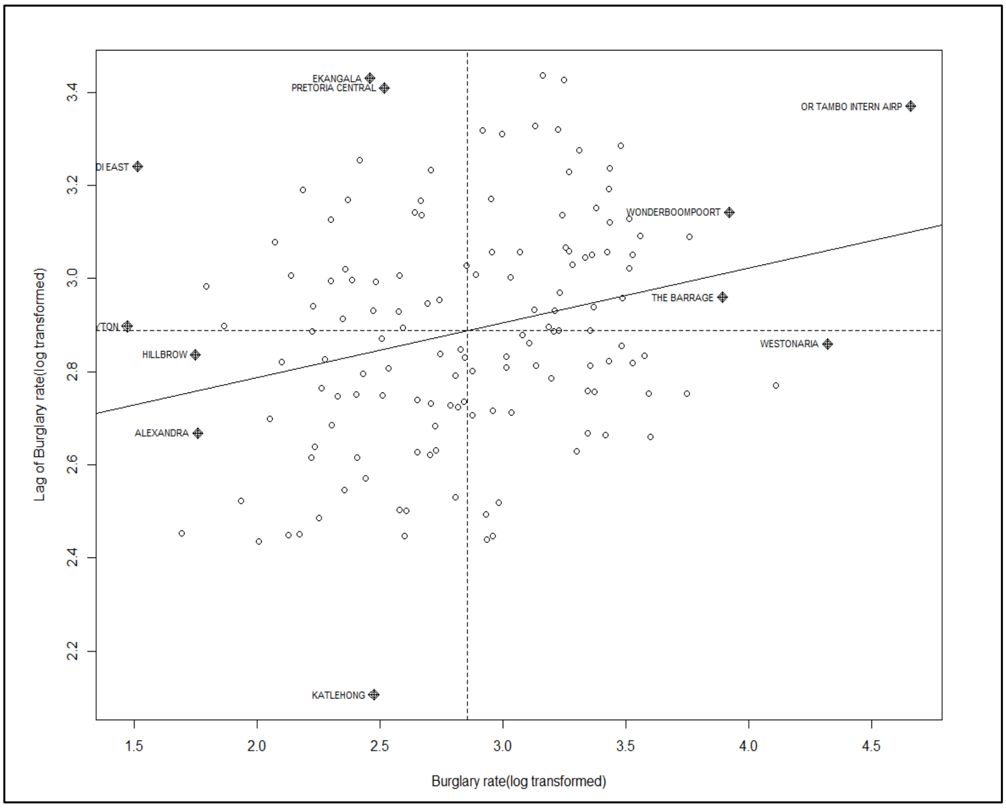


Figure 4.7 Moran's I Plot for log transformed burglary

Figure 4.7 presents the Moran's I plot based on the queen's weighting criteria. Figure 4.7 illustrates the association between the attribute values (horizontal axis) and the local mean attribute value (the mean value of the adjacent locations). The positive slope of the line depicted a global positive spatial autocorrelation. Police stations OR Tambo Airp (Airport), The Barrage and Wonderboompoort were outliers in the High-High region while police stations Alexandra, Hillbrow, and Katlehong were outliers in the Low-low region. There were outlier values in the Low-High and High-Low regions of the Moran's I scatter plot, and these were, respectively, Lyteltton, Enkangala, Pretoria Central Mamelodi East and Westonaria (Figure 4.7).

Although the tests for spatial autocorrelation using the Moran's I statistic implied clustering of burglary in Gauteng, it was important to discriminate clusters of significant spatial autocorrelation (hot spots and cold spots) in the study area. The importance of this was that it can assist in policymaking and effective allocation of resources. LISA map for the dependant variable, burglary rates (log transformed) is presented in Figure 4.8.

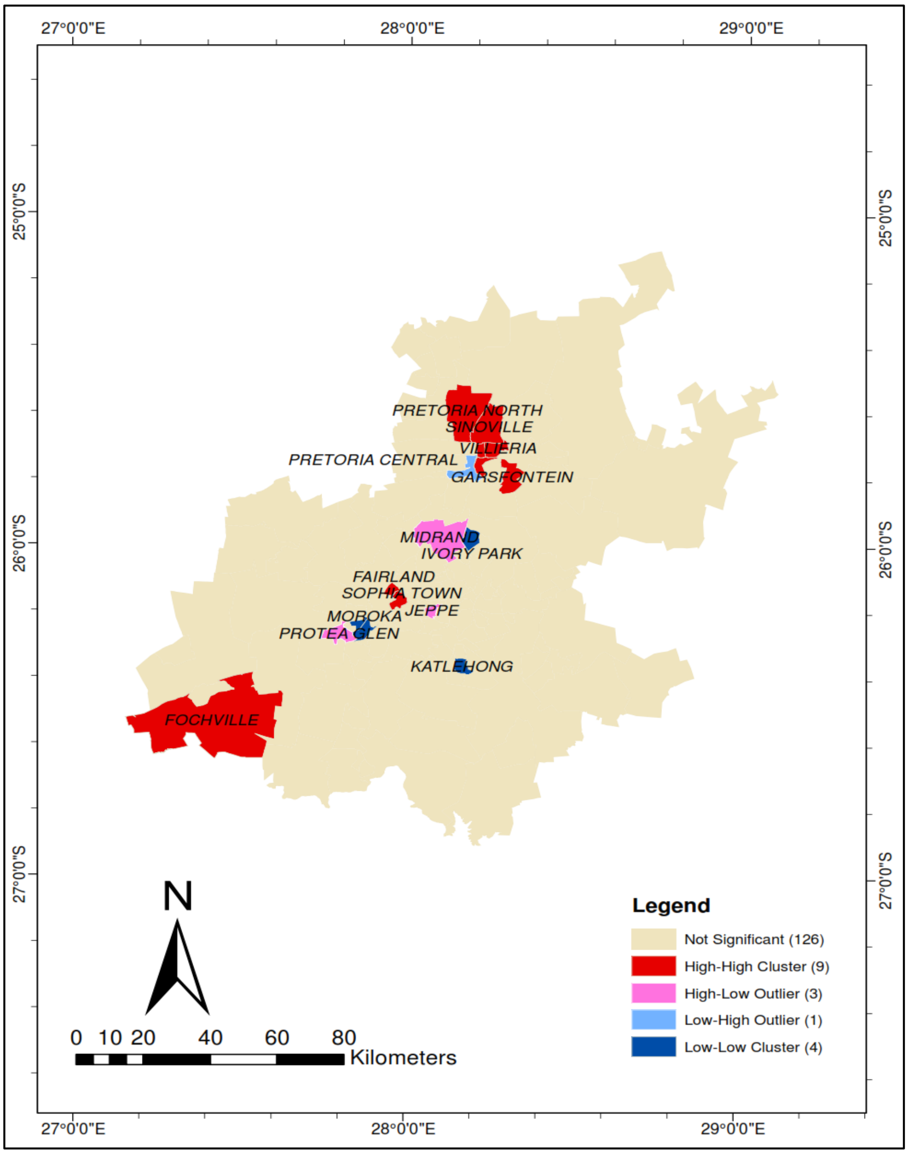


Figure 4.8: LISA Map and LISA significance map: Gauteng burglary, 2011 (Source: SAPS)

Figure 4.8 shows the hot spots (high-high) and cold spots (low-low) for burglary based on the queen's case first-order contiguity weights matrix. Regions associated with high-high clusters were Garsfontein, Pretoria Moot, Sunnyside, Villeria in the City of Tshwane municipality and Sophia Town in the City of Johannesburg. In the West Rand, Fochville was surrounded by other high regions. Low-low cluster regions were in Katlehong, Jabulani, Ivory park and Moroka regions in the City of Johannesburg Municipality. Midrand, Protea Glen, and Noorwood regions had high values surrounded by low values. Regions associated with high values surrounded by low values were outliers. Hotspots were also found in the central and Western parts of the study area.

4.4 Results for Ordinary Least Squares regression (OLS)

Five different models were considered (Table 4.4). The first model included all the variables in the model and examined the variables for statistical significance to determine if variables can be eliminated. Three models were formulated to assess the separate contribution of the 3 groups of variables specified in Table 3.1 (motivated offender, a suitable target and lack of capable guardian) and to test the hypotheses postulated in Section 2.4. Finally, the variable selection for OLS regression analysis was performed by employing a series of iterative steps. Iteratively, previously eliminated variables were successively added in the model and the reduction of the Akaike Information Criterion (AIC) was assessed. This iterative process is referred to as a Step-Wise Method (Chatterjee and Hadi, 2006). The variable inflation factors (VIF) was also observed with every iterative step to ensure that the model is not over-specified or redundant (Environmental Systems Research Institute, 2014). The recommended VIF value must be less than 7.5 (Environmental Systems Research Institute, 2014). The goodness of fit of the OLS of the model was assessed using R^2 (Getis and Aldstadt, 2004) and the AIC that compares performance between any two competing models (Akaike, 1974).

Table 4.5: OLS results

Dependent variable: Burglary Rates (log transformed)					Final model		
	All Variables	Motivated Offender Variables	Suitable Target Variables	Capable Guardian Variables	Final Model	t value	p-value
	Model 1	Model 2	Model 3	Model 4	Model 5		
Percentage of population unemployed	-0.077*** (0.019)	-0.072*** (0.009)			-0.048*** (0.011)	-4.22	0.000043 ***
Percentage of males aged between 15 and 34 years	0.035** (0.016)	-0.010 (0.012)					
Percentage of households with no income	-0.002 (0.008)	0.006 (0.008)					
Percentage of female headed households	0.010 (0.009)		-0.002 (0.007)				
Percentage of population with a BSc Degree	-0.072*** (0.026)		-0.023 (0.022)				
Percentage of White-headed households	0.002 (0.003)		0.013*** (0.003)				
Population density	-0.00003 (0.00002)		-0.00005*** (0.00002)				
Roads Density	-0.00000 (0.00001)		-0.00000 (0.00001)		-0.00002*** (0.00001)	-3.49	0.00064 ***
Percentage of people who moved in the last 5 years	-0.036 (0.036)			-0.019 (0.035)	-0.029*** (0.009)	-3.29	0.00129 ***
Percentage of people with no citizenship	-0.003 (0.021)			-0.008 (0.022)			
Ethnic heterogeneity index	0.478* (0.247)			1.373*** (0.126)	0.530** (0.215)	2.47	0.01467 **
Constant	2.939*** (0.555)	3.752*** (0.208)	2.876*** (0.239)	2.462*** (0.071)	3.500*** (0.219)	15.98	0***
Observations	143	143	143	143	143		
R^2	0.603	0.45	0.495	0.462	0.557		
Adjusted R^2	0.569	0.438	0.477	0.451	0.544		
Residual Std. Error	0.373 (df = 131)	0.426 (df = 139)	0.411 (df = 137)	0.421 (df = 139)	0.384 (df = 138)		
F Statistic	18.070*** (df = 11; 131)	37.860*** (df = 3; 139)	26.900*** (df = 5; 137)	39.840*** (df = 3; 139)	43.340*** (df = 4; 138)		
VIF	2.517						
Note:	*p<0.1; **p<0.05; ***p<0.01, df= degrees of freedom, standard errors in parenthesis						

Table 4.5 shows the results of the OLS models that were fitted. Model 1 was fitted to examine the variables that predict burglary in Gauteng. Out of the eleven candidate variables, only ethnic heterogeneity index, percentage of the population with BSc degree, percentage of the population unemployed and percentage of males aged between 15 and 34 years were significant in the model. Only the ethnic heterogeneity index had a positive parameter estimate, although it was only significant at 10%.

In Model 2, where motivated offender variables were tested in the model, only the percentage of population unemployed was significant in predicting burglary when other variables were held constant. The sign of the parameter estimate was negative. Model 1 was better in terms of performance compared to Model 2 based on the adjusted R^2 value.

Percentage of white-headed households and population density were significant in Model 3 while other suitable target variables were not included in the model. A positive sign for the percentage of white-headed households was found while the population density indicated a negative sign. Model 3 performed better than Model 2 (motivated offender) and indicated 48% of the variance in the dependent variable was accounted for by the independent variables.

With respect to the variables that capture the lack of capable guardian (Model 4), only the ethnic heterogeneity index was significant ($p<0.01$) and the sign of the parameter estimate was positive.

Over specification of an OLS model occurs when insignificant variables are included in the model such that they contribute no explanatory value in the model (Baybak, 2004). The inclusion of variables that do not contribute any explanatory power has an effect of increasing the R^2 value. The increased R^2 value leads to a misleading interpretation of the explanatory power of the model (Baybak, 2004). As such, Model 1 was over-specified based on the adjusted R^2 compared to the final model (Model 5). Over specification was diagnosed by the fact that iteratively removing most insignificant variables in Model 5 resulted in relatively the same adjusted R^2 and AIC values with fewer variables and higher degrees of freedom. The residual plots for Model 1 indicated distinctive non-random pattern which suggested that the model was not specified correctly (Chatterjee and Hadi, 2006). Therefore Model 5 was chosen as the final model in this study. Choosing Model 5 resulted in fewer variables (4) used to explain relatively the same amount of variation (adjusted R: 54%) in the model compared to 11 variables in model 1.

The results for the final model fitted using the step-wise method is presented under Model 5 in Table 4.5. The variables that were selected after the stepwise procedure were the percentage of the population unemployed, roads density, percentage of the population that moved in the last 5 years and ethnic heterogeneity index. The stopping criteria was a high adjusted R^2 value and least AIC. Except for ethnic heterogeneity index which had a positive parameter estimate, the other variables had negative parameter estimates. The variation that was explained by the final model was 54.4% based on the adjusted R^2. The AIC for the final model was 139.

The hypotheses testing the overall significance for Model 5 was set up as follows;

H_0: The model is not significant vs H_1: The model is significant

The *F*-statistic = 43.34 and the *p*-value for the final model was significant with $p < 0.01$. The null hypothesis was rejected, and it was concluded that the model was significant in explaining burglary in Gauteng.

The hypothesis testing for the coefficients of each of the independent variables as presented in the selected Model 5 in Table 4.5 was stated as follows;

H_0: The coefficient of the independent variable in the fitted OLS model was not significant

H_1: The coefficient of the independent variable in the fitted OLS model was significant

The *p*-value for each of the independent variables was less than 0.05 and thus the null hypothesis was rejected in each case and the conclusion was that the coefficients of the independent variables indicated in Table 4.5 were all significant.

The model did not show evidence of multicollinearity because the VIF of the model was 2.404, which was less than 7.5 for severe multicollinearity as recommended by (Environmental Systems Research Institute, 2014).

The signs of the parameter estimates relating to all the variables mentioned in Section 1.4 were unexpected in all the models, except for the ethnic heterogeneity index which was found to be positive in all the models. The hypotheses in Section 1.4 were as follows, population density and roads density was expected to increase burglary, female-headed households were expected to increase crime and lastly, the percentage of the population that moved in the last 5 years and ethnic heterogeneity index was also expected to increase crime. The population density was only significant in Model 3 and its parameter estimate was negative. Percentage of female-headed households was not significant in all the models; however, it had a positive parameter estimate in the full model and the sign changed to negative in Model 3. Ethnic heterogeneity index was significant and positive in Models 1, 4 and 5, while the percentage of the population that moved in the last 5 years was only significant on the final model.

4.5 OLS Diagnostics

Assumptions of the OLS detailed in Section 3.4.5 were tested using diagnostics plots. These plots were used to detect patterns in the residuals of the fitted OLS models. Clustering was diagnosed by means of Moran's I statistic on the residuals. A significant *p*-value leads to rejecting the null hypothesis of no dependence of the error term and indicates spatial regressions must be fitted (Martin, 2002; Voss *et al.*, 2006). Normality of the residuals was diagnosed by the Jarque-Bera test statistic.

Figure 4.9 presents the residuals versus fitted values for Model 5, the studentised scatter plots in the top row. The normal quantile-quantile plot and the Cook's distance plots are respectively presented in the last row. The scatter plots (indicated on the top row in Figure 4.9) allow the visualisation of patterns exhibited by the residuals. The idea was to judge if the regressors were linear unbiased estimators of the phenomenon being investigated (Chatterjee and Hadi, 2006). Non-random patterns in the scatter plots give an indication of violations of the OLS assumption. The normal quantile-quantile plot gives an indication of how well the residuals fit a normal distribution. For normally distributed residuals, most observations will be very close to the 45-degree line of best fit. The Cook's distance plot was used to indicate outlier observations.

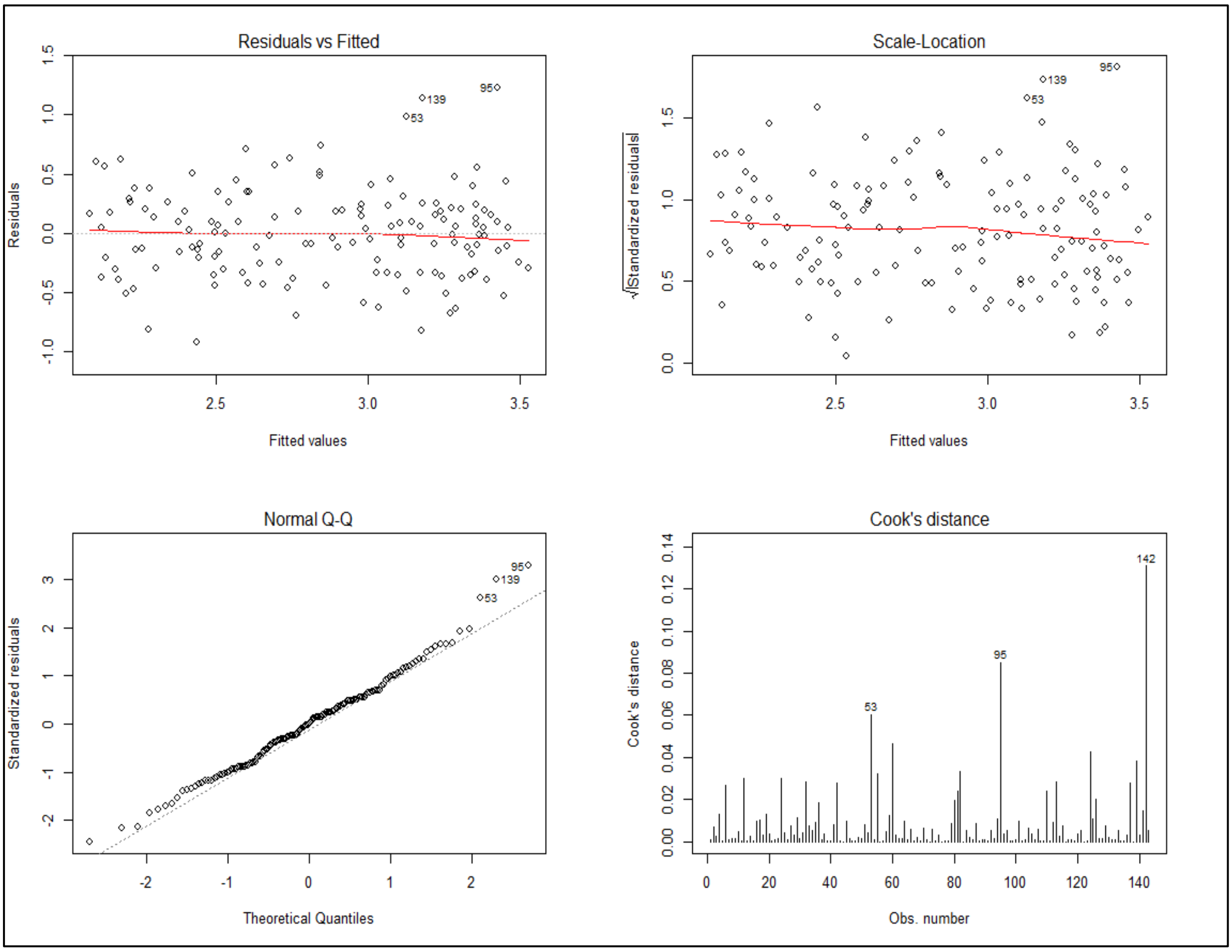


Figure 4.9: OLS diagnostics plots; Residuals vs fitted value and normal quantile-quantile plot

These plots in Figure 4.9 indicated that there were no severe violations to the OLS assumptions. The residuals versus fitted scatter plot did not show evidence of non-random pattern because the residuals were fairly dispersed with no discernible non-random cluster. The standardised versus fitted values scatter plot also did not show evidence of non-randomness. The normal quantile-quantile plot indicated that most observations fall on the 45-degree line of best fit. However, there were observations (outliers) that were significantly greater than the theoretical values at locations 95, 124 and 139. These areas were OR Tambo International Airport, Sunnyside and Westonaria respectively. These outliers signified greater influence in the determination of the slope on the regression modelling compared to other observations.

It was preliminarily accepted that the residuals were independent and normally distributed. Formal statistical tests of the assumed normality are presented in Section 4.5.1.

4.5.1 Normality test of OLS residuals for the final model

Diagnostics plots in Figure 4.9 suggested that the residuals may be accepted to follow a normal distribution. The formal statistical test was performed to check this assumption. The hypotheses were formulated as stated below;

H_0: The residuals follow a normal distribution,

H_1: The residuals do not follow a normal distribution

The *p*-value for the Jarque-Bera test statistic (4.0) was found to be $p = 0.13$. The null hypothesis was not rejected thus the residuals follow a normal distribution.

4.5.2 Moran's I on the error term of OLS

A formal test of spatial autocorrelation on the residuals of the OLS model was conducted to confirm the suggested patterns on the residual plots.

The hypotheses for this were set up as follows:

H_0: The residuals were not spatially auto-correlated vs H_1: The residuals were spatially auto-correlated

Using the Moran's I test statistic (0.046) for spatial autocorrelation on the residuals of the OLS model, the *p*-value was found to be 0.2. The null hypothesis was not rejected. Thus, the residuals were not spatially autocorrelated.

4.5.3 Test for homoscedasticity

The test conducted in Section 4.5.2 was used for checking spatial autocorrelation on the residuals of the final model; to check violations of the assumptions of OLS as discussed in Section 3.4.5. One of the assumptions made about regression residuals is that they must have constant variance throughout the study region (Chatterjee and Hadi, 2006). The Breusch-Pagan (BP) test was conducted to check if the variance of the residuals was constant.

The hypotheses for the formal test of homoscedasticity were set up as follows;

H_0: The residuals of the OLS model have constant variance vs H_1: The residuals of the OLS model do not have a constant variance

The value p = 0.057 for the Breusch-Pagan test statistic (9.19) was greater than 0.05, however, the test was significant at 10%. Although the homoscedasticity test indicated that the residuals were random at 5% significance level, the p-value was marginally greater than the rejection level (0.05) and this suggested that relationship could be modelled better using GWR.

4.5.4 Tests for the type of spatial dependence present in the data

Moran's I test in Section 4.5.2 indicated that there was no spatial dependence present in the model. The LM tests were however conducted to further show evidence of no spatial dependence present in the model. Results are presented in Table 4.5.

Table 4.6: Results of LM tests for spatial type of spatial clustering

Statistic	df	Value	p-value
LagrangeMultiplier (lag)	1	0.6738	0.412
RobustLM(lag)	1	0.6716	0.413
LagrangeMultiplier (error)	1	0.1332	0.715
RobustLM(error)	1	0.131	0.717
df = degrees of freedom			

The hypotheses for the LM tests were set up as follows:

> H_0: The Lagrange multiplier for each of the **spatial regression** model was not significant
>
> H_1: The Lagrange multiplier for each of the **spatial regression** model was significant

The p-values for all the tests further shown that there was no significant spatial dependence in the data and thus no spatial regression models were fitted.

4.5.5 Model residuals map

In Sections 4.5.2 above, the spatial autocorrelation of the residuals was tested, and the results indicated no spatial autocorrelation of the residuals, however, the test of spatial dependence (homoscedasticity in Section 4.5.3) indicated that the residual variance was not constant throughout the study region at the 10% significance level.

Residuals of the final model were mapped to visualise how the residuals cluster and to check which regions were outliers based on Model 5 in Table 4.4. The choropleth map is presented in Figure 4.10. The values of the choropleth map were classified using the standard deviation method because the residuals follow a normal distribution. The patterns observed in Figure 4.10 indicated that the residuals were random.

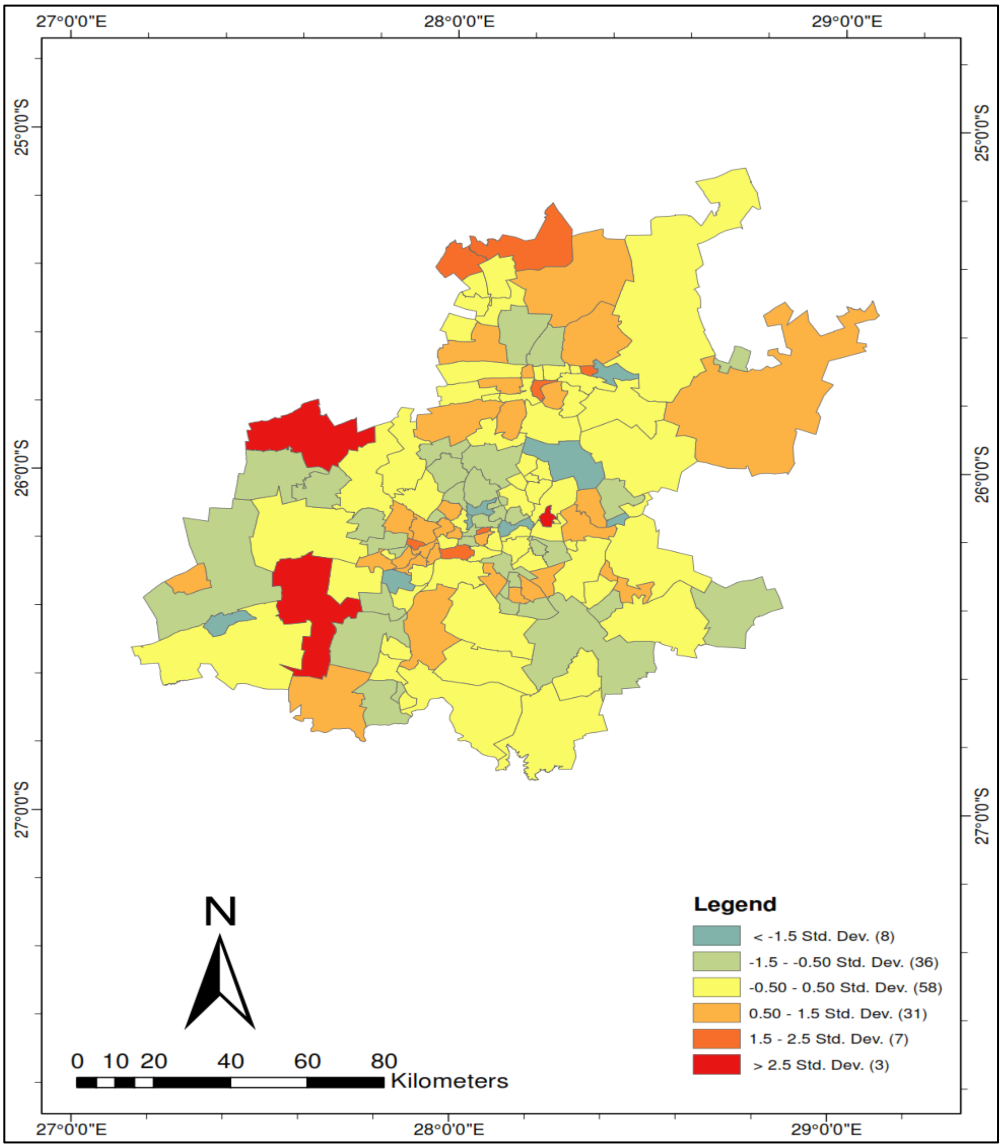


Figure 4.10: Choropleth map of standardised residuals based on the fitted OLS model

The residuals were tested to have no spatial autocorrelation in Model 5. The residuals were showing evidence of heteroscedasticity based on the Breusch-Pagan test, although only at 10% which violates OLS assumptions regarding the variance of OLS. Thus, the GWR was fitted and results are presented in Section 4.6 below. It was decided to conduct GWR only because the rejection level 0.057 was not substantially far from the rejection level of 0.05.

4.6 GWR Results

A statistically significant test for non-homoscedasticity at 10% significance level in Section 4.5.3 suggested that GWR modelling may better explain the local dynamics of the determinants of burglary in the study area. Two calibrations of the kernel functions; Gaussian and bi-square adaptive kernels were tested. The calibration method chosen was the nearest neighbour bi-square adaptive kernel AIC (131) over the adaptive Gaussian kernel method (AIC=139). The optimal bandwidth was calibrated using GWR4.0 (Nakaya, 2016). The optimal bandwidth was found to be 50 nearest neighbours. The same set of variables used under Model 5 were also used in the GWR so that comparisons between the OLS and GWR models can be conducted.

The adjusted R^2 and AIC for the final model were, respectively, 54.4% and 139. The GWR model improved the results slightly because the adjusted R^2 increased to 59% and the AIC, reduced to 131 based on the same variables.

4.6.1 GWR summary statistics

The mean values for the intercept of the GWR model versus Model 5 were found to be relatively the same at 3.2 (for model 5 this was 3.5). Comparison of parameter estimates (mean) values between GWR and Model 5 revealed the following; a) the ethnic heterogeneity index was greatly improved in the GWR (higher mean estimate value, 0.775) compared to 0.53 in Model 5, b) the percentage of the population unemployed increased slightly while the density of the roads did not change, c) the signs for all parameter estimates remained unchanged. The minimum and maximum *t*- values for all the estimates are shown in the last two columns in Table 4.7.

Table 4.7 GWR parameter estimates summary statistics

	Parameter basic statistics						t value basic statistics	
	Min.	1st Quartile	Mean	Median	3rd Quartile	Max	Min.	Max
Intercept	1.69149	2.52239	3.20808	3.31313	3.81435	4.79874	2.606824	9.177365
Roads Density(m/km²)	-0.00010	-0.00003	-0.00002	-0.00001	0.00001	0.00003	-4.64705	1.72757
Percentage of people unemployed	-0.09012	-0.06943	-0.03889	-0.03923	-0.01442	0.02595	-3.72424	1.029865
Percentage of people that moved in the last 5 years	-0.07382	-0.03095	-0.02300	-0.02328	-0.01255	0.01204	-3.11748	0.292166
Ethnic heterogeneity index	-0.54731	0.29424	0.77590	0.72356	1.27840	1.90192	-0.92865	3.956332

The regions indicated in darker shades of black in Figure 4.11 indicated high values of explained variation based on the GWR model with the chosen variables. The GWR model performed better in terms of R^2 (Figure 4.11) in the upper central, Eastern and Southern parts.

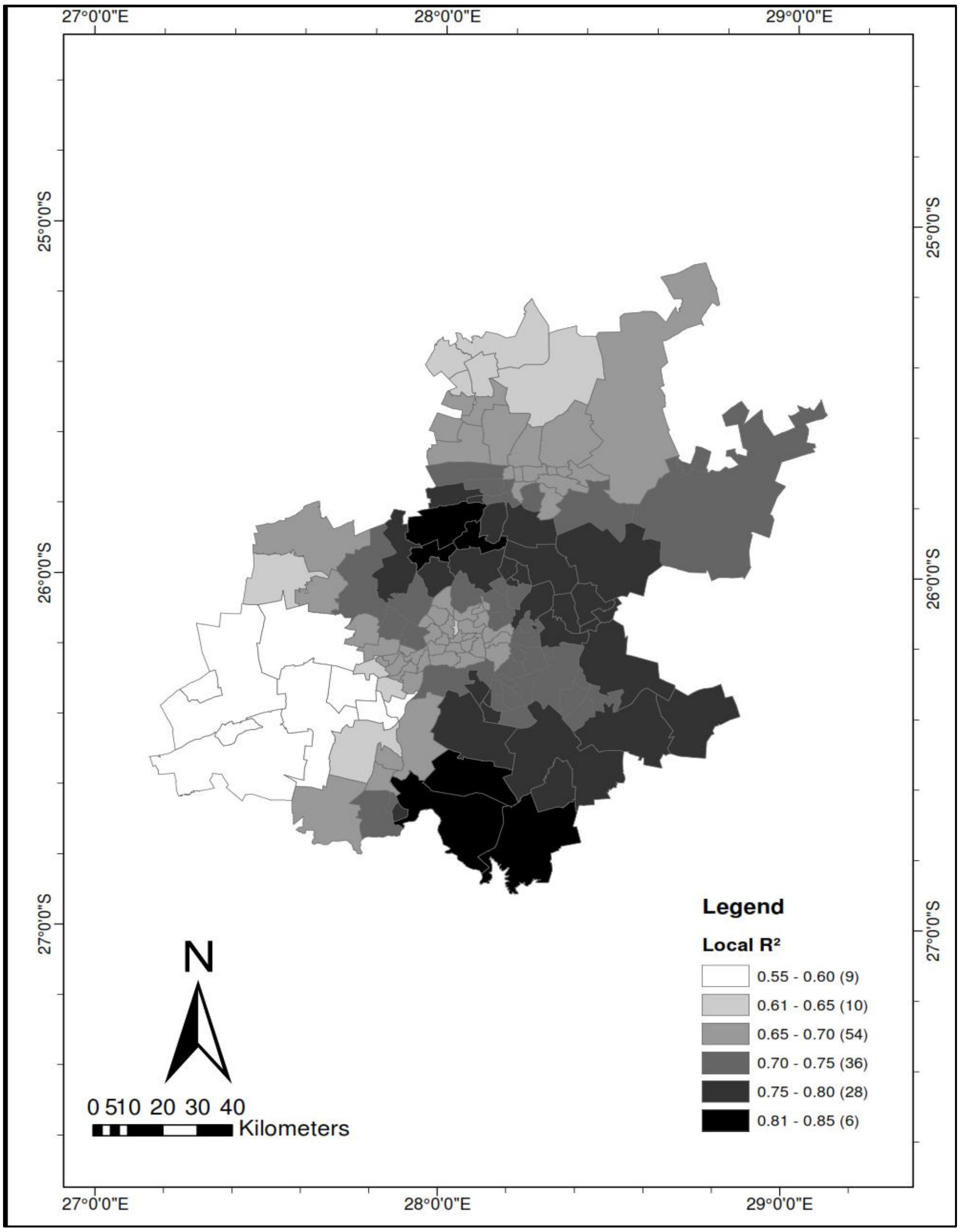


Figure 4.11 Local R^2 map for GWR, 2011

Figure 4.12 indicated that the residuals under the GWR model were random.

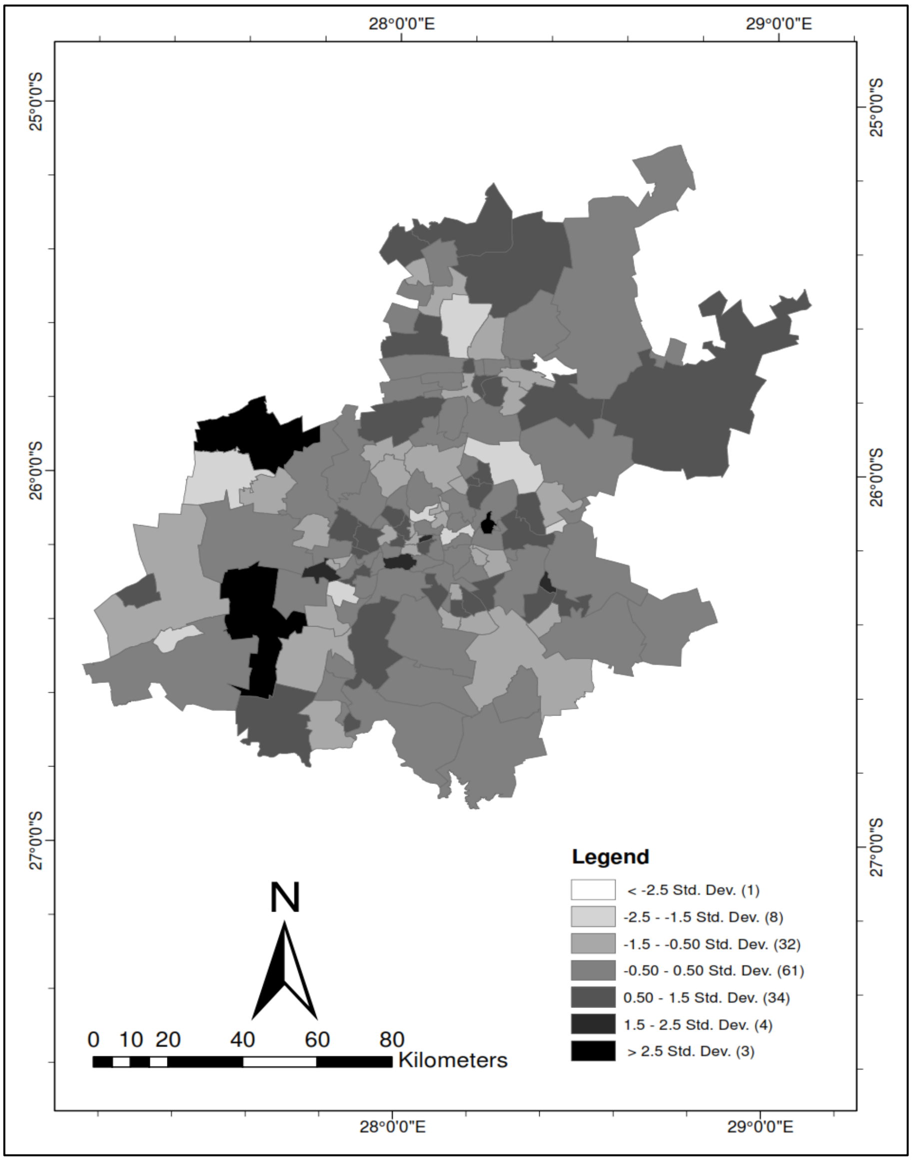


Figure 4.12: Choropleth map GWR standardised residuals, 2011

4.6.2 GWR maps for selected variables

The *t*-values for GWR maps were calculated by dividing the estimated value by the standard error for every point in the study area. As stated in Section 3.4.8, the statistical significance of the variables in the GWR model was tested by comparing *t*-values to the critical values of the *t* distribution at 5% and 1% significance levels with more than 30 degrees of freedom. 30 degrees of freedom are preferred because a *t*-distribution with larger degrees of freedom closely resembles a standard normal distribution (Grami, 2019). Since there were 143 regions corresponding to 143 *t*-values to be tested, maps were plotted, and the values were classified according to the critical values at 5% and 1% levels. The parameter estimates choropleth maps together with t-values are presented in Figures 4.13 to 4.16, respectively for the percentage of the population unemployed, percentage of the population that moved in the last 5 years, roads density and ethnic heterogeneity index.

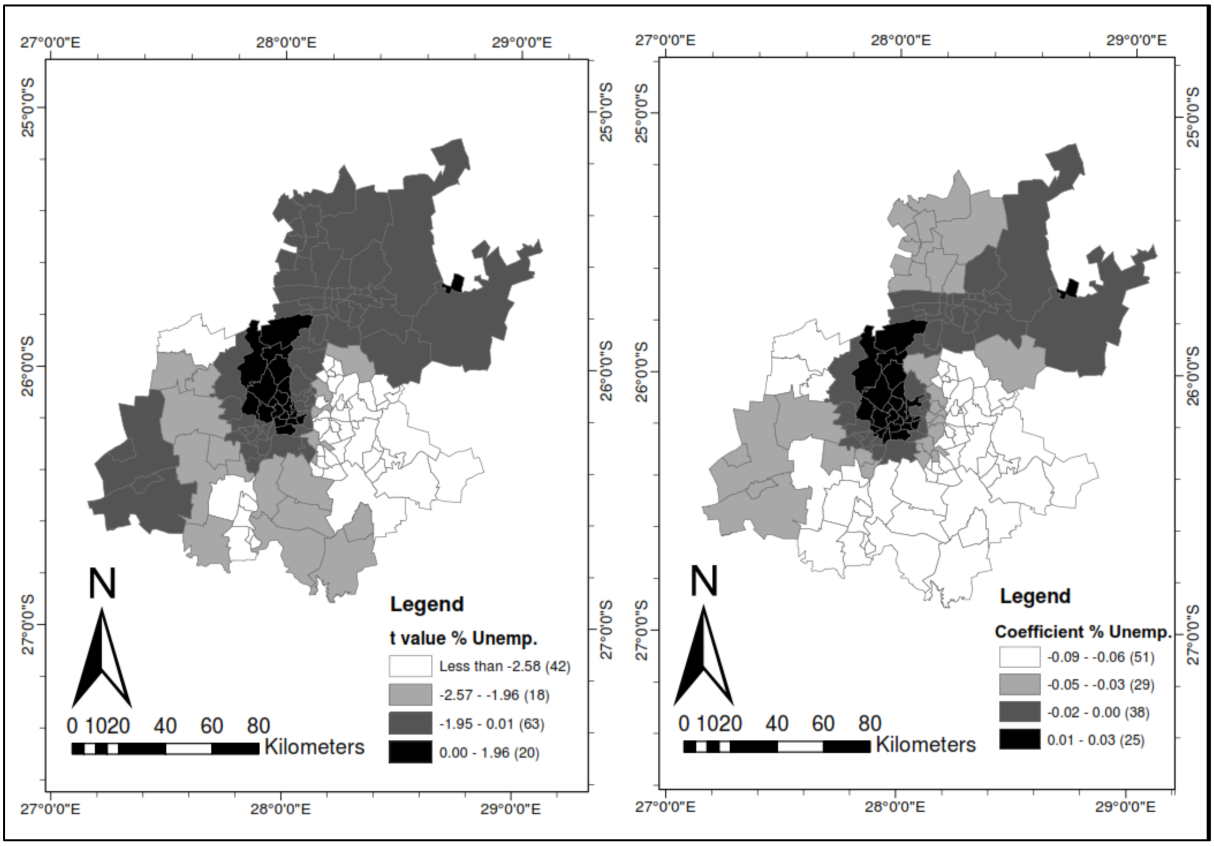


Figure 4.13: Choropleth map of GWR for t-values and coefficients estimates for the percentage of the population unemployed, 2011

In the map in Figure 4.13, the regions shown in white represent the regions with negative significant values (at 1% level, less than -2.58 critical value) for the percentage of the population unemployed and its effect on burglary. There were distinct clusters of the estimated parameters (coefficients) in the central, Southern and Eastern parts of the study area. The values for the coefficients of the percentage of the population unemployed had positive and negative values. This implied that locally, the GWR revealed a complex phenomenon compared to the global OLS which does not reveal this complexity.

The percentage of the population that moved in the last 5 years had negative significant values at 1% level in the central and Eastern parts of the study area, shown by white in Figure 4.14.

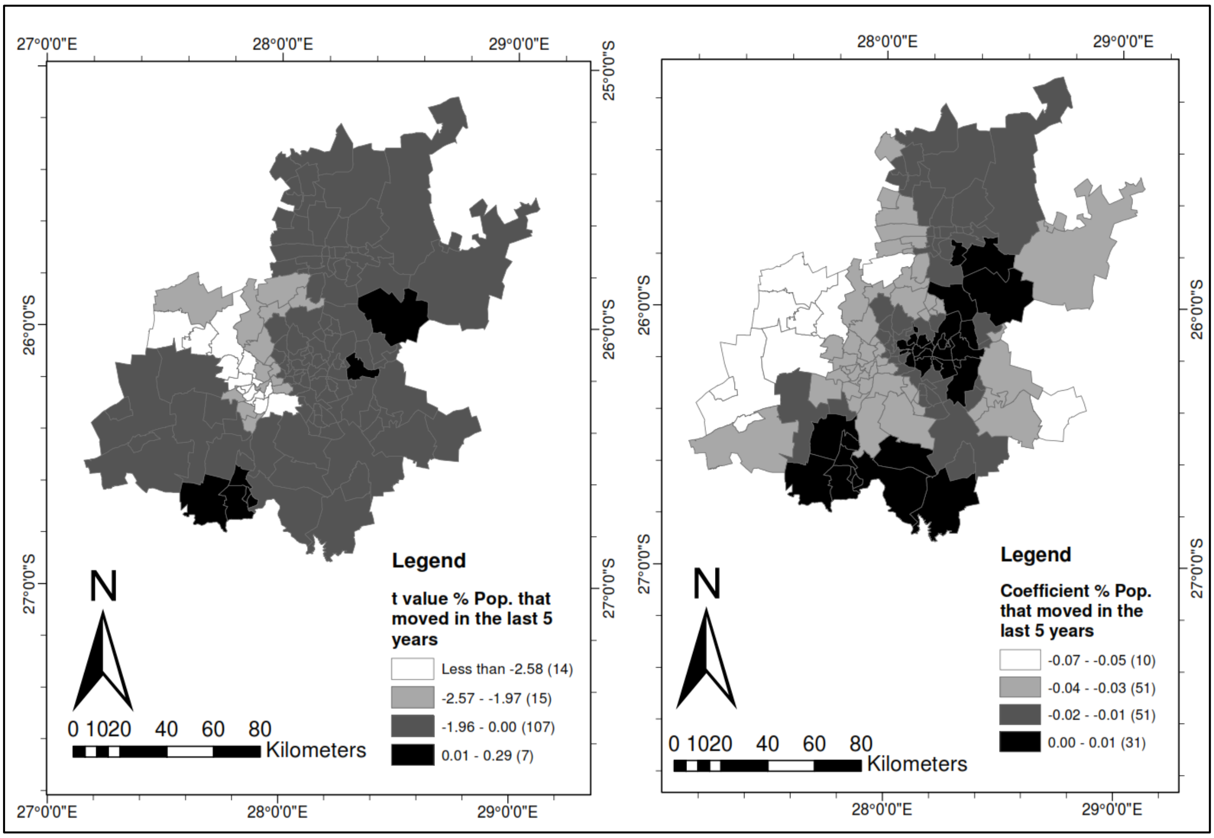


Figure 4.14: Choropleth map of GWR for t-value and coefficient of percentage of population that moved in the last 5 years, 2011

Roads density variable affected burglary in a non-random manner. In Figure 4.15 roads density had negative significant values in the central-Eastern parts of the study region as revealed by the cluster of values coloured white in the map.

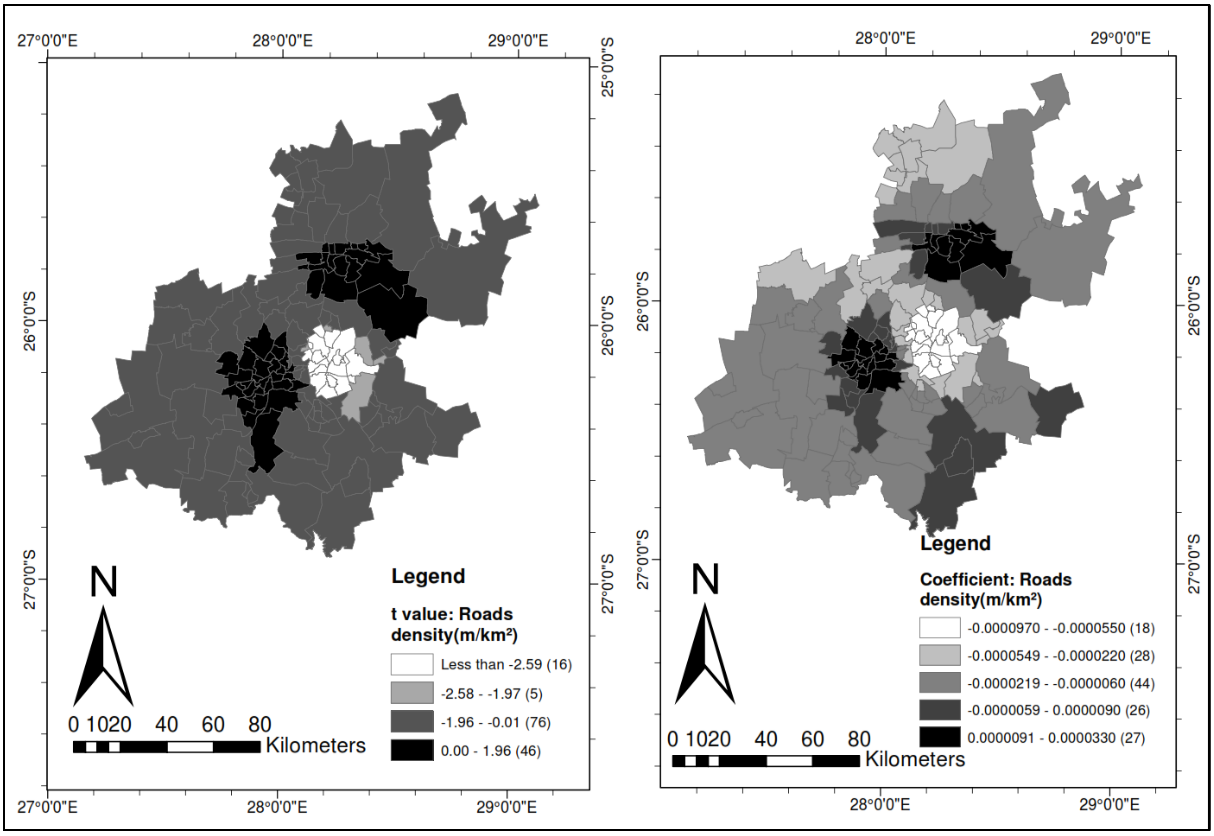


Figure 4.15: Choropleth map of coefficients and t-values for roads density

Similarly, to the maps presented in Figures 4.13 to Figure 4.14 above, the relationship between burglary and roads density was positive and negative in different parts of the study region.

Ethnic heterogeneity (Figure 4.16) index also shown evidence of variability within the study area in terms of the relationship depicted by the coefficient values (negative and positive).

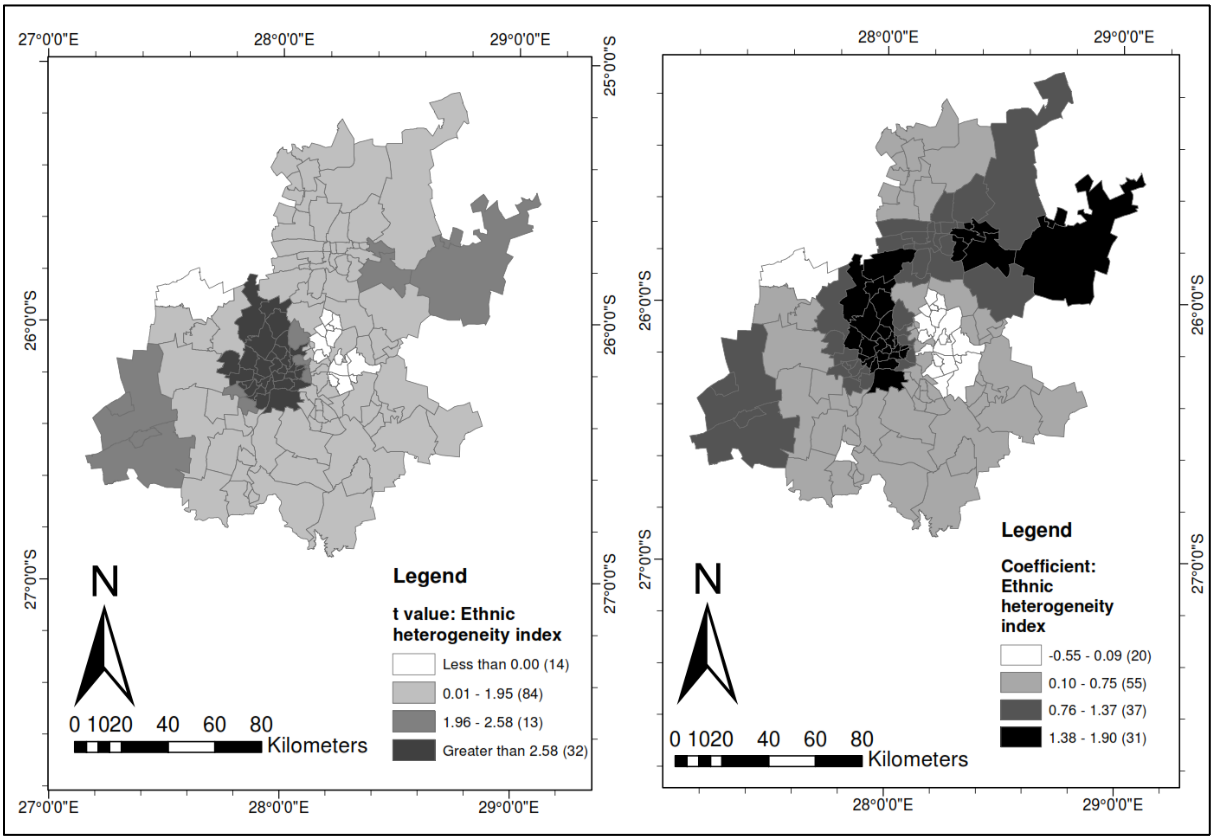


Figure 4.16: Choropleth map for coefficient and t-value for ethnic heterogeneity index

The *t*-values for ethnic heterogeneity index indicated that there were significant (at 1% level) positive values in the central parts of the study area as shown by darker grey colour in Figure 4.16 above.

The positive and negative parameters for all the selected variables provided evidence to the robustness of using local regressions for modelling of burglary. This can assist in conducting research at the local level so that the cause of these patterns can be understood.

4.7 Comparison between police cluster and police precinct level results

4.7.1 Introduction

One of the objectives of this report was to check if the results change when the unit of analysis was changed. It was important to check the results at a different scale because aggregation may have an effect on the results when spatial lattice data are used for modelling (Dark and Bram, 2007). To achieve this, SAL attributes were aggregated to SAPS cluster regions (23 regions) using the same technique described in Section 3.3. The cluster-level dependant variable was standardised using 100 000 households because the mean number of households was 169813. This standardisation does not affect the comparison of the results because this is constant.

The exploratory analysis results between cluster level and police precinct-level results were conducted because the crime rates at cluster level could not be transformed to normality using the same log transformation. This implied that the parametric OLS could not be performed and that only non-parametric OLS was required. Non-parametric OLS were beyond the scope of this report and thus they were not conducted.

4.7.2 Police cluster-level summary statistics

The summary statistics at cluster level are found in Appendix B. In terms of the bivariate correlation (Appendix C), the variable that had a strong positive correlation (0.65) with burglary rates was the percentage of households headed by Whites. This size of correlation (0.65) between burglary and percentage of White-headed households was found to be relatively similar even at police precinct level (0.52). There was a strong, negative correlation between burglary rates and percentage of households with no income (-0.89). Ethnic heterogeneity index had weak positive, correlation (0.33) in the cluster level results compared to the police precinct-level results (0.51). Further summary statistics plots are found under Appendices D to F.

The normality was further confirmed by the Shapiro Wilk test found under Appendix E, common forms (log and square root) of transformations that were attempted did not conform to normality. There was no spatial autocorrelation present in the chosen variable using both the queen's case criteria and distance (inverse) based criteria (Appendix E).

The independent variable summary plots did not indicate significant deviations from the normal line of the quantile-quantile plots and there were no significant outliers on all the summary plots (Appendix D).

4.7.3 Visualisation of results at the police cluster level

Choropleth maps showing the spatial patterns between the two scales; cluster and police precinct are presented in Figure 4.17. The patterns on both maps indicated higher crime rates were found on the North-Eastern parts of Gauteng. The higher concentration levels of burglary in the southern central part of the study at police precinct level were not evident at the cluster level. This can be attributed to the effect of aggregation that reduces variability in the data.

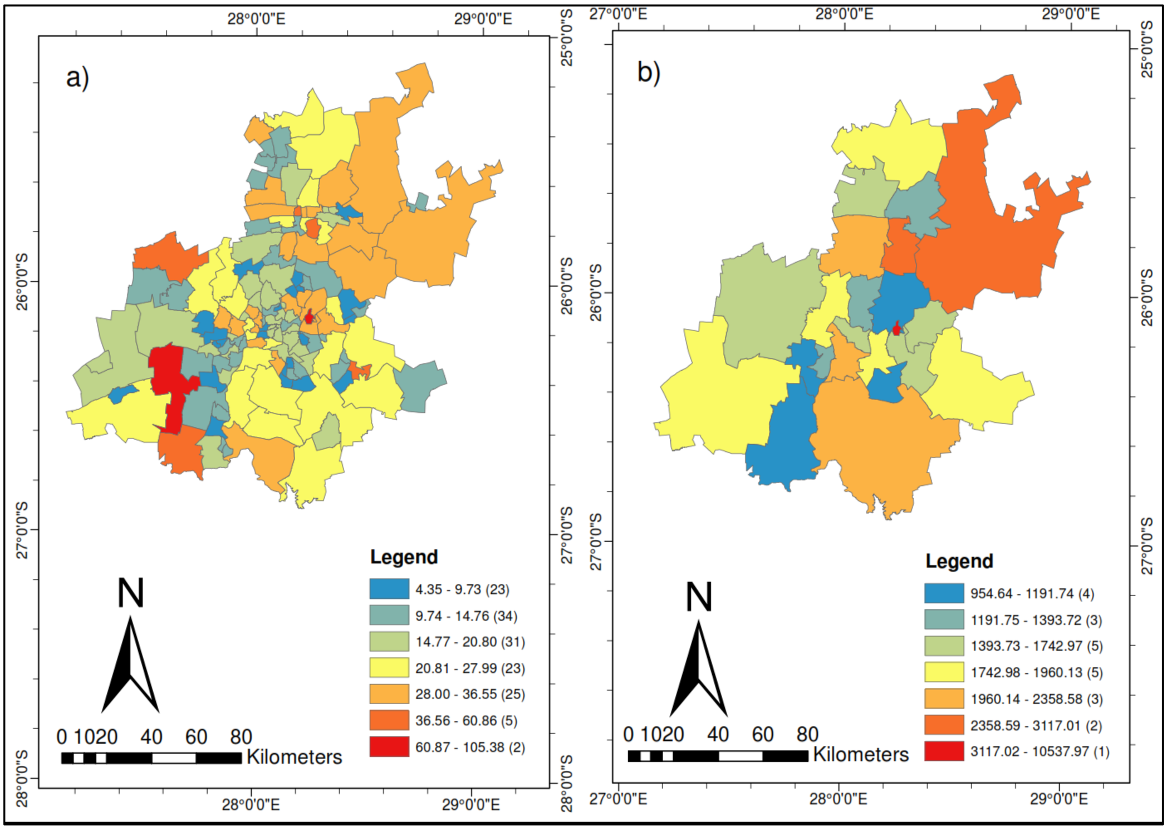


Figure 4.17: Choropleth maps for cluster and precinct-level data

4.7.4 Cluster level spatial autocorrelation tests

The first order contiguity and inverse distance weights matrices were formulated like the analysis done at the precinct level. There was no spatial autocorrelation in the cluster level data as the Moran's I statistic was not significant as indicated in the summary table in Appendix E.

5 Discussion

Crime studies in SA, despite it being ranked amongst the highest in terms of crime, are limited (Bhorat *et al.*, 2017; StatsSA, 2017). Studies of crime in SA are undoubtedly required. This report has adopted methodologies and selection of some common determinants of burglary from existing literature to achieve the aim, objectives and research hypotheses mentioned in the introductory section.

The results for Model 5 were fitted by step-wise OLS regression technique because, although Model 1 was performing better as suggested by the adjusted R^2 value, it seemed to be over-specified because most of the variables were not significant. Over specification was diagnosed by the fact that iteratively removing most insignificant variables resulted in relatively the same adjusted R^2 and AIC values in the final model; with fewer variables. Therefore Model 5 in Table 4.4 was chosen as the final model in this study. The significant variables were roads density, percentage of the population unemployed, the percentage of people that moved in the last 5 years and ethnic heterogeneity index. This model included at least one variable of the elements of RA theory. Although there was a strong negative correlation (-0.85) between ethnic heterogeneity index and the percentage of people unemployed, the variables were added in the model because the VIF were all below 7.5 for severe multicollinearity.

5.1 Crime and explanatory variables

The hypotheses were; a) population density and roads density were expected to increase crime b) percentage of people moved within a period of 5 years prior to the data collection and ethnic heterogeneity index of a neighbourhood were expected to increase the chances of burglary and lastly c) female-headed households have higher chances of burglary. These variables were selected because they have been previously shown to affect crime (Bernasco and Luykx, 2003; Demombynes and Özler, 2003; Breetzke, 2012; StatsSA, 2014, 2017; Bhorat *et al.*, 2017).

It has been argued that greater permeability in terms of roads leads to least effort for the burglar in committing the crime (Davies and Bishop, 2013). Cozens (2011) states that changing grid-based roads networks using road closures has been shown to successfully reduce crime. Burglaries are not merely opportunistic (Bhorat *et al.*, 2017); criminals judge the risks of being apprehended and apprehension may not totally be unrelated to permeability. Although road densities were previously found not be significant in the North-Western parts of Gauteng (City of Tshwane Metropolitan Municipality) (Breetzke, 2012), this study demonstrated an opposite finding; that the density of the roads had a significant but negative correlation with burglary rates (Table 4.3). The regression analysis (Section 4.4) conducted also shown this significant negative relationship with burglary. This relationship suggested that in Gauteng, crimes decreased with higher roads density at a police precinct. In support of this claim, map d) in Figure 4.5 had clusters of higher roads density values in the central region. In the same region, the GWR map (Figure 4.15) revealed negative significant values at 1% level. Therefore in Gauteng road density can be regarded as a capable guardian instead of an enabler for burglary as it has been previously reported in other studies (Johnson and Bowers, 2010; Davies and Bishop, 2013).

The population density was hypothesised to have a positive relationship with burglary because the higher the population density the greater there will be households and thus items to steal (Harries, 2006). The negative correlation between population density implied that an increase in population density resulted in decreased burglary rates in Gauteng. This can be attributed to the fact that there may be a greater chance of spotting a criminal activity in a high-density area (Demombynes and Özler, 2003). The idea is that there may be surveillance, whether directly (in the form of security guards) or indirectly (by means of high population and activity) in high populous areas. Therefore, for burglaries, which are not opportunistic, the population density, much like roads density, were found to have a deterring effect in areas being targeted. Population density may be regarded as a capable guardian when crimes are investigated with the use of RA.

In a study by Malczewski and Poetz (2005) in London, Ontario, the percentage of people that moved in the previous 5 years, was not found to be a significant correlate with burglary. In this study, the findings of the correlations suggested that areas characterised by high residential mobility may be less attractive to burglars because there was a non-significant correlation with burglary. This can be attributed to the argument that people in these regions usually do not have a property (Breetzke, 2010). Variables that measure the lack of capable guardian are expected to have a positive relationship with crime (Cohen and Felson, 1979). Correlations are not causal (Shmueli, 2010). This study used predictive analysis, which implied that not only exploratory results were needed. Therefore, although there was no significant correlation between burglary and percentage of the population that moved in the last 5 years, it did not imply that its effects on burglary are not significant (Shmueli, 2010). Thus, in the OLS model, the percentage of the population that moved in the last 5 years was found to be significant and has a negative relationship with burglary. This relationship suggested that burglary decreased with an increase in residential mobility. Literature in SA suggested that residential mobility (measured by the percentage of the population that moved in the previous 5 years) were also significant determinants of violent crimes (Breetzke, 2010b).

The results for ethnic index had a strong positive correlation suggesting that police precinct level racial heterogeneity was a significant determinant of burglary. Ethnic heterogeneity index was previously shown to affect burglary in a positive manner (Bernasco and Nieuwbeerta, 2005). The ethnic heterogeneity index captures the lack of capable guardian in the sense that areas characterised by substantially diverse races have an effect that people socialise minimally. Non-socialising has the negative effect that people will not create social bonds and informal social controls that make residential areas less attractive to crime (Sampson and Groves, 1989). In Figure 4.16 ethnic heterogeneity index was shown to be significant at 1% significance level in the central parts of Gauteng. Similarly, the central parts had higher index values in the central parts on the choropleth map (Figure 4.6). Its explanatory power (R^2) was found to be 42% when it was considered alone in the regression analysis (Appendix G). The positive parameter estimates for ethnic heterogeneity index implied that regions that were characterised by mixed races were likely to be affected by burglary than other regions. This supported the SOD theory that postulates that areas lacking social cohesion are more vulnerable to crime than areas that have high social cohesion. Social cohesion has been measured by the use of ethnic heterogeneity index in the literature (Hipp, 2007).

The literature on various types (violent and property crime) of crimes constantly makes reference to gender playing an important role in crime (Osgood, 2000; Brown, 2001; Demombynes and Özler, 2003; Lancaster and Kamman, 2016; StatsSA, 2017). Some of the pertinent issues relating to gender to crime are the fact that females are perceived to be more vulnerable to criminal victimisation (Smith, Torstensson and Johansson, 2011). In this study, the results relating to female-headed households shown a weak and negative correlation with burglary rates. This relationship was also found in SA by Breetzke (2010); StatsSA, (2017) and Brown (2001). StatsSA (2017) provides a possible explanation for this negative relationship between burglary and female-headed households and they state that the female-headed households have a guardian most of the time. Brown (2001) states that female-headed households are perceived to be characterised by low economic status and ultimately less attractive. Conversely, Demombynes and Özler (2003) found a positive relationship between female-headed households and property crime in SA.

The percentage of the population unemployed was previously shown in the literature to have a positive relationship with burglary (Breetzke, 2012). However, Demombynes and Özler (2003) found no significant relationship between property crimes and unemployment rates. In this report, percentage of the population unemployed had a negative relationship with burglary rates, particularly; a unit increase in the percentage of population unemployed resulted in a decrease of burglary rates (log) by a factor of 0.08 while holding other variables constant. This relationship suggested that areas that had high unemployment percentages had lower crimes. Literature suggest that one of the main driving forces of burglary is for monetary gain (Brown, 2001; Demombynes and Özler, 2003). Therefore areas with high unemployment rates are expected to be less attractive than areas that have high employment percentages; which may be highly correlated with high-income levels (Bernasco and Luykx, 2003; Bernasco and Nieuwbeerta, 2005). In fact, in this report, it was demonstrated on the correlation statistics that there was a moderate positive (0.64), a significant correlation between the percentage of the population unemployed and the percentage of households with no income. However, it is worthwhile to state that correlations do not imply causality (Breetzke, 2012). In the regression results, the percentage of the population unemployed was shown to be a highly significant (1% level) but the negative determinant of burglary compared to other significant variables.

It is rare to find studies in criminology that do not mention inequality and its effects on crime. In this study, percentages of households with no income were investigated and it was found that there was a negative relationship with burglary. The possible explanation for this relationship in the South African context could be linked to the apartheid practices that marginalised the majority of non-Whites in terms of income (Breetzke, 2012). The situation in SA is that there are high levels of inequality (Demombynes and Özler, 2003; Chibba and Luiz, 2011). High levels of inequality, measured by household's mean expenditure and mean per capita expenditure of the area of concern, lead to crime occurrence (Demombynes and Özler, 2003). However, areas that have high levels of inequality are the formerly segregated (racially, spatially and economically) communities (Bhorat *et al.*, 2017). These formerly segregated areas may not be attractive to burglars due to the same reasons they may be perceived to contain no assets to steal.

Brown (2001) stated that crimes were mostly committed by men. In Model 1 in this study, a unit increase in the percentage of males aged between 15 and 34 resulted in an increase of burglary by a factor of 0.035 when other variables remained constant. This finding supported the RA theory, that is motivated males aged between 15 and 34 at police precinct level will commit burglary. But most importantly, it must be noted that this statement may not be applicable at a different scale such as at an individual level, because individual-level characteristics may not be applicable to neighbourhood level characteristics (Gotway and Young, 2002). This phenomenon is referred to as ecological fallacy and it is said that one cannot deduce individual-level characteristics from aggregated data.

Models 2, 3 and 4 were used to test the effects of the variable groups relating to the elements of RA individually. The aim was to see if they supported the hypotheses proposed in this report. As previously mentioned in this report, the motivated offender variables, suitable target and lack of capable guardians all had negative relationships with burglary except for ethnic heterogeneity index which had a positive relationship in Models 1, 4 and 5 while the percentage of White-headed households was positive in Model 3.

5.2 Spatial patterns of data

Studies on crime generally report on spatial effects such as spatial autocorrelation and spatial heterogeneity (Demombynes and Özler, 2003; Andresen, 2006; Kelejian and Prucha, 2010; Bernasco and Elfers, 2011; Zhang, Suresh and Qiu, 2012; Breetzke and Cohn, 2013; Hiropoulos and Porter, 2014). It was mentioned in Chapter 3 that these factors can lead to incorrect inferences relating to crime if they are not diagnosed particularly when parametric OLS regressions are used for prediction and modelling (Chatterjee and Hadi, 2006). Therefore, diagnostics tools such as choropleth maps, Moran's I, Jarque-Bera and Breusch-Pagan statistical tests on the residuals of the fitted (Model 5) OLS were conducted. In addition to these (choropleth maps, Moran's I, Jarque-Bera and Breusch-Pagan statistical tests on the residuals), LISA maps were created to examine areas where spatial autocorrelation is significant.
The results presented in this report found no statistical significance of residuals being non-normal based on the Jarque-Bera statistical tests (Section 4.5.1). This suggested that the

residuals did not violate the assumptions of OLS regression modelling. In support of the residuals being normally distributed, the choropleth map was drawn to visualise the patterns portrayed by the residuals. The model (Model 5) residuals were random with no discernible pattern as seen in Figure 4.11. The random patterns of residuals suggested burglary in an area could be estimated with accuracy based on the significant variables. Similar findings were reported by Breetzke (2012). The results were probably similar because most of the variables, the theoretical backcloth, and analysis techniques were very similar between these studies. The Moran's I on the residuals was found to be insignificant (Section 4.5.3); which suggested that there was no need to conduct either spatial error or spatial lag regression models. However, the formal tests for the heteroscedasticity (Breusch-Pagan) of the regression residuals indicated that GWR would improve the results to capture the local dynamics of the chosen explanatory variables (Chen *et al.*, 2017). Although the statistical test for homoscedasticity in the final model was only indicating heterogeneity at 10% significance level, the GWR (Section 4.6) improved the results and the choropleth maps indicated that the local modelling may be better in explaining burglary in Gauteng. Heteroscedasticity of the OLS residuals variance was also found in other studies (Martin, 2002; Breetzke, 2012; Chen *et al.*, 2017).

The GWR model was fitted with the same variables that were used in the OLS (Model 5). The GWR model improved the results in terms of explained variation in the model and AIC. The adjusted R^2 increased to 59% while there was a slight improvement in the AIC value to 131. The GWR t-value maps (Figures 4.14 to 4.17) in Section 4.6 provided evidence that determinants of burglary were not completely random because the test for heteroscedasticity, although not significant at 5% and the residuals plots showing no visual patterns, at 10% there was heteroscedasticity. Although GWR is an exploratory tool, it provided proof that local relationships may be explained better with the use of GWR and support for this argument was found in the literature (Fotheringham, Brunsdon and Charlton, 2002; Malczewski and Poetz, 2005; Breetzke, 2012; Chen *et al*., 2017)

5.3 Issues with units of analysis

It is acknowledged that crimes aggregated to different spatial scales may not be accurate in modelling the relationships between dependant and independant relationships (Zhang, Suresh and Qiu, 2012). Some of the issues associated with aggregation include the boundary issue when a point in polygon aggregation is conducted. This aggregation technique overlooks the influence of criminal events committed at or the boundaries of polygons used for analysis (Zhang, Suresh and Qiu, 2012). Some authors do not indicate this issue as a potential limitation in their studies such as Bhorat *et al.* (2017). This research sought to test if there were going to be any statistical changes in the results (correlations and patterns of the data). This was achieved by aggregating the census small area level to police cluster level and conducting the comparisons. The goal was to check if the results changed and to conclude on the appropriate or inappropriateness of aggregating to larger units.

Indeed, in this study evidence of concealment of relationships between variables was found (Section 4.7, Figure 4.17), particularly, there was no spatial autocorrelation present in the police cluster level data. This was opposite to the relationship explained at length in the previous chapters relating to police precinct-level data.

6 Conclusion, limitations and recommendations

This report has provided insight into some of the determinants of burglary in Gauteng in for the 2011/2012 financial year. These determinants included the roads density, percentage of the population unemployed, ethnic heterogeneity index and percentage of people that moved/ migrated in the last 5 years of the 2011 census report. The research employed the RA theory. The proposed hypotheses were proven to be untrue as they related to crime. The relationships between burglary and the chosen explanatory variables were negative except for the ethnic heterogeneity index variable. This negative relationship suggested that further studies need to be conducted in order to fully understand the crime situation in Gauteng.

Although testing RA was not the main objective of this research, the regressions models presented in this report provided contradictory results to some of the requirements of RA theory. However, support to the social disorganisation theory was found, specifically ethnic heterogeneity index provided the most explanatory power (higher coefficient estimates and higher *t-values*) in the models. Thus, RA should be studied more comprehensively to understand its applicability in the South African context.

This report also provided some insights into how the chosen variables vary across the study region in terms of how they affect burglary. Despite the heterogeneity significant only at the 10% level, there were significant improvements when a GWR was used to model the relationship. The GWR provide insights into how interventions can be made at a local level.

It has also been shown in this report that crimes are complex in nature because of the mixed results in literature with different explanations. It proves that modelling human behaviour is complex in nature (Brantingham and Brantingham, 1993).

6.1 Limitations

The official SAPS crime numbers may not be truly representative of the burglaries that are committed (StatsSA, 2014). While most crime reports against property are normally reported for insurance claims purposes, Hiropoulos and Porter (2014), stated that the factors for non-reporting burglaries to the officials include among others, the distance from the police stations, police response time and general dissatisfaction with the police services (StatsSA, 2017). On the other hand, Victimisation Surveys are only based on samples. Despite the coverage issue, SAPS data are the better source. Burglary crimes still contribute the highest to the category of property-related crime in SA despite the gradual reduction of the reported counts to the SAPS. However, the reduction of the counts may not possibly be attributed to the success of the SAPS in combating this crime alone. One of the reasons for the reduced numbers could be the perceived lack of trust in the authorities in successfully apprehending and prosecuting the perpetrators of criminal activities which leads to under-reporting (StatsSA, 2017).

The adoption of a specific theory to explain crime for a particular geographic location may not be useful in all areas because crime is complex and *"many socio-economic and psychological conditions may result in someone engaging in theft*" (Brantingham and Brantingham, 1993).

6.2 Recommendations

This study found significant indicators of ethnic heterogeneity index contributing positively to burglary and contributing well to explained variation of burglary. Therefore, it is recommended that if governments are to combat crime, then they should engage with people to promote social cohesion and they must encourage and recognise other crime prevention strategies such as community crime forums. Governments must implement and promote policies on social cohesion through social and economic development. Social and economic policies must also expand to include collaborations from all professionals and individuals who have an interest in socio-economic development.

Reports of lack of trust in the SAPS must be studied carefully by the SAPS so that police can engage with the community members in a manner that promotes collaborative solutions to crime. Stricter sentences on convicted burglars and other similar types of crimes must be implemented so that choice to commit a crime can have more costing effects than gains.

Ethnic heterogeneity index may be used (mapped) to focus more police patrols in areas that have significantly higher index heterogeneity of races. Deploying more police resources in areas with higher racial heterogeneity may significantly reduce burglary.

It is argued that RA is not suitable for crime studies in SA based on this study and support from findings in other studies in SA (Breetzke, 2010; Breetzke, 2012; Breetzke and Cohn, 2013). Therefore, research on SOD must be expanded in SA so that crime dynamics can be understood and modelled better.

Technically, it is recommended that point data at police precinct level be used to fully investigate burglary because aggregated data may not be appropriate for individual-level inferences (Gotway and Young, 2002). Different techniques of aggregation such as the Dasymetric and Areal kriging technique aggregation techniques may be used to aggregate census level data to police level data so that the method that offers better estimates be used. Weir-Smith (2016) provided an attempt on Areal kriging technique and found more accurate interpolations in rural areas than in urban areas. Another aggregation technique was presented by (Bhorat *et al.*, 2017). Bhorat *et al.*, 2017 utilised a random point in polygon algorithm to determine SAL points that fell within police precinct-level data and made an aggregation based on the resulting points. Hence extension in these techniques is recommended for further crime studies.

7 Appendices

7.1 Appendix A- Summary of independent variable calculations

Variable	Calculation / Definition
Ethnic heterogeneity index	$$1 - \frac{A^2 + B^2 + C^2 + D^2 + E^2}{(A + B + C + D + E)^2}$$ Where A, B, C, D, and E are, respectively, the number of black people, number of coloured people, number of Indian/Asian people, number of White people and the number of other people as defined by StatsSA (2011). The index indicated the likelihood that two randomly selected neighbourhoods are of different ethnic origins (Bernasco and Luykx, 2003). The index ranges from 0 to 1, with 0 indicating no ethnic heterogeneity (a neighbourhood with no mixed race), and 1 indicating complete heterogeneity of race.
Roads density	Sum of the total street length in metres per precinct boundary divided by the total area of the police boundary measured in square kilometres.
Population Density	Sum population count at the police precinct divided by the area in square kilometres of the police boundary zone (after considering the regions that do not have households).

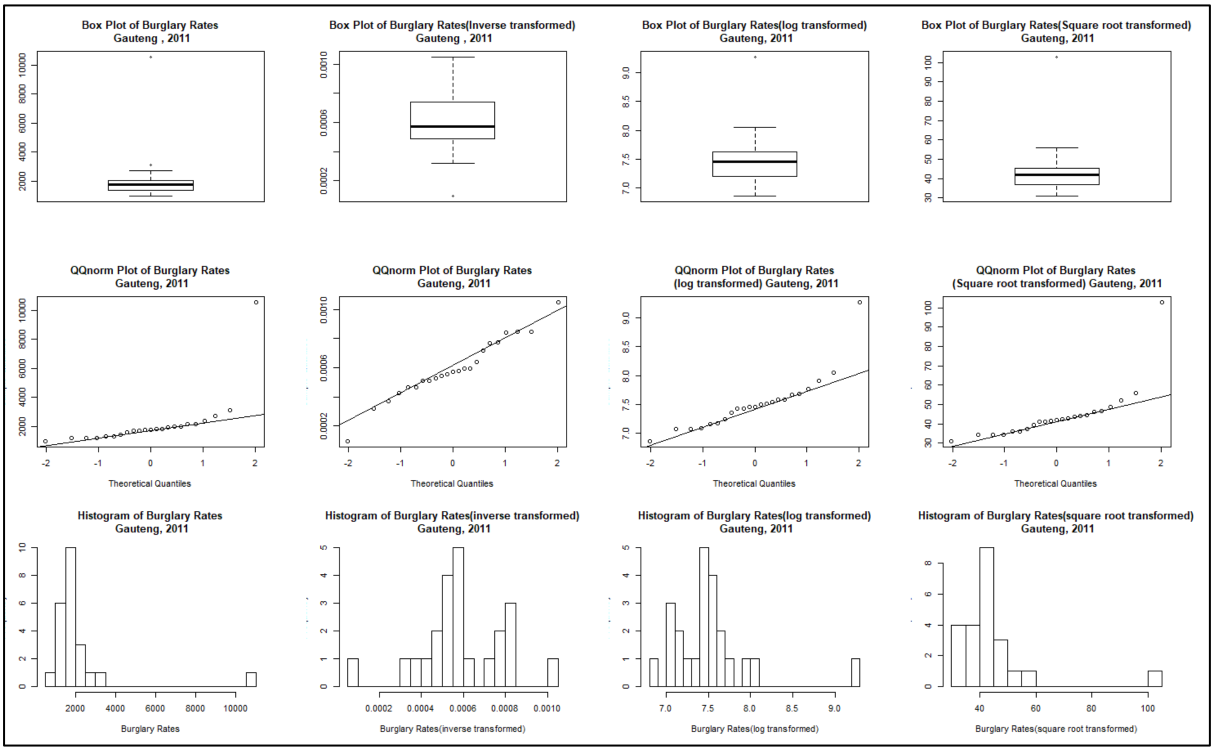
7.2 Appendix B- Cluster Summary Statistics

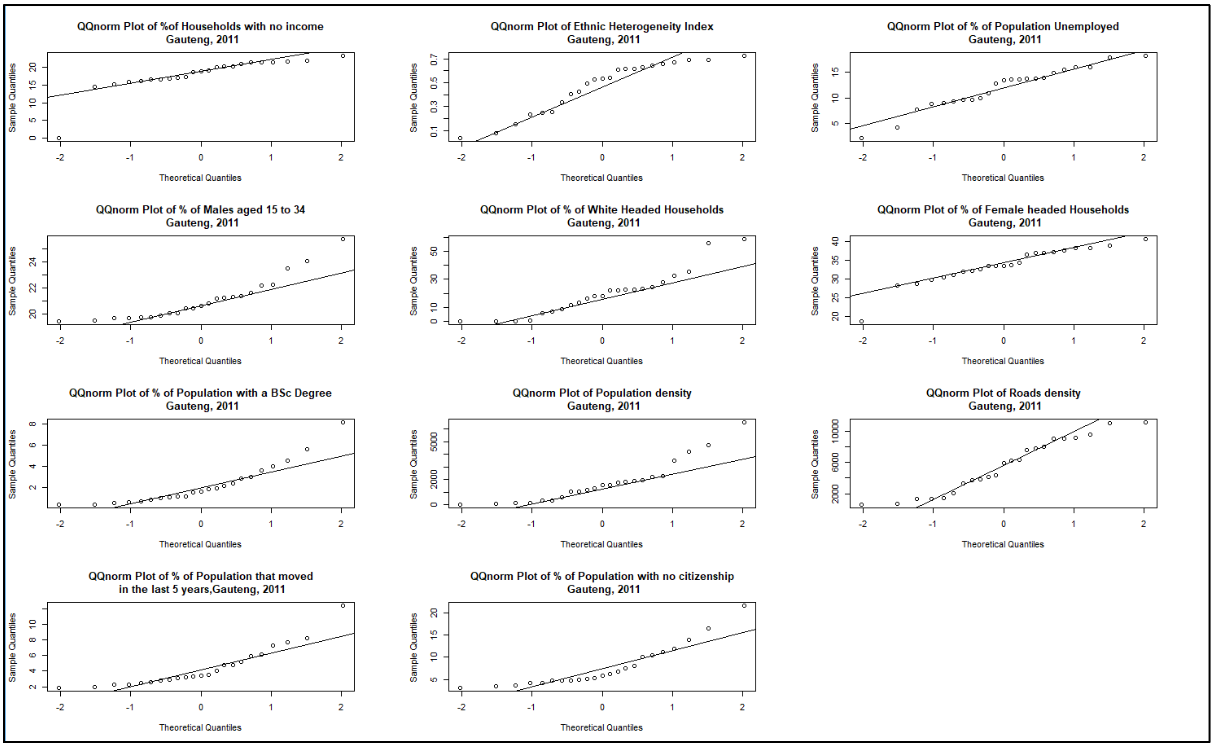
Name of Variable	n	Min.	Max.	1. Quartile	3. Quartile	Mean	Median	Variance	Std dev	Skewness	Kurtosis
Burglary rates per 100 000 households	23	954.64	10537.97	1349.88	2050.60	2145.92	1742.97	3608359.86	1899.57	3.71	13.59
Burglary rates per 100 000 households (log transformed)	23	6.86	9.26	7.21	7.62	7.51	7.46	0.23	0.48	2.00	5.31
Burglary rates per 100 000 households (square root transformed)	23	30.90	102.65	36.74	45.27	44.24	41.75	197.60	14.06	3.05	10.07
Population density	23	10.26	6506.00	442.83	2049.81	1735.33	1525.51	2702681.18	1643.98	1.27	1.08
Roads Density	23	647.36	11039.57	2704.77	8518.15	5546.06	5909.30	11541191.56	3397.23	0.05	-1.45
Percentage of female headed households	23	18.81	40.59	31.58	37.12	33.67	33.57	22.34	4.73	-1.15	1.82
Percentage of White-headed households	23	0.09	58.39	7.87	23.69	19.41	17.98	248.18	15.75	0.90	0.38
Percentage of population unemployed	23	2.25	18.09	9.46	14.38	11.94	13.48	16.10	4.01	-0.58	-0.28
Percentage of males aged between 15 and 34 years	23	19.42	25.74	19.78	21.49	21.05	20.63	2.60	1.61	1.29	1.08
Percentage of households with no income	23	0.00	23.20	16.60	21.15	18.07	18.86	21.62	4.65	-2.41	7.09
Percentage of population with a BSc Degree	23	0.42	8.08	0.96	2.96	2.23	1.65	3.58	1.89	1.45	1.72
Percentage of people who moved in the last 5 years	23	1.86	12.36	2.69	5.60	4.43	3.44	6.55	2.56	1.39	1.60
Percentage of people with no citizenship	23	3.12	21.54	4.73	10.24	7.75	5.86	21.98	4.69	1.35	1.13
Ethnic heterogeneity index	23	0.04	0.72	0.30	0.63	0.47	0.53	0.04	0.21	-0.65	-1.00

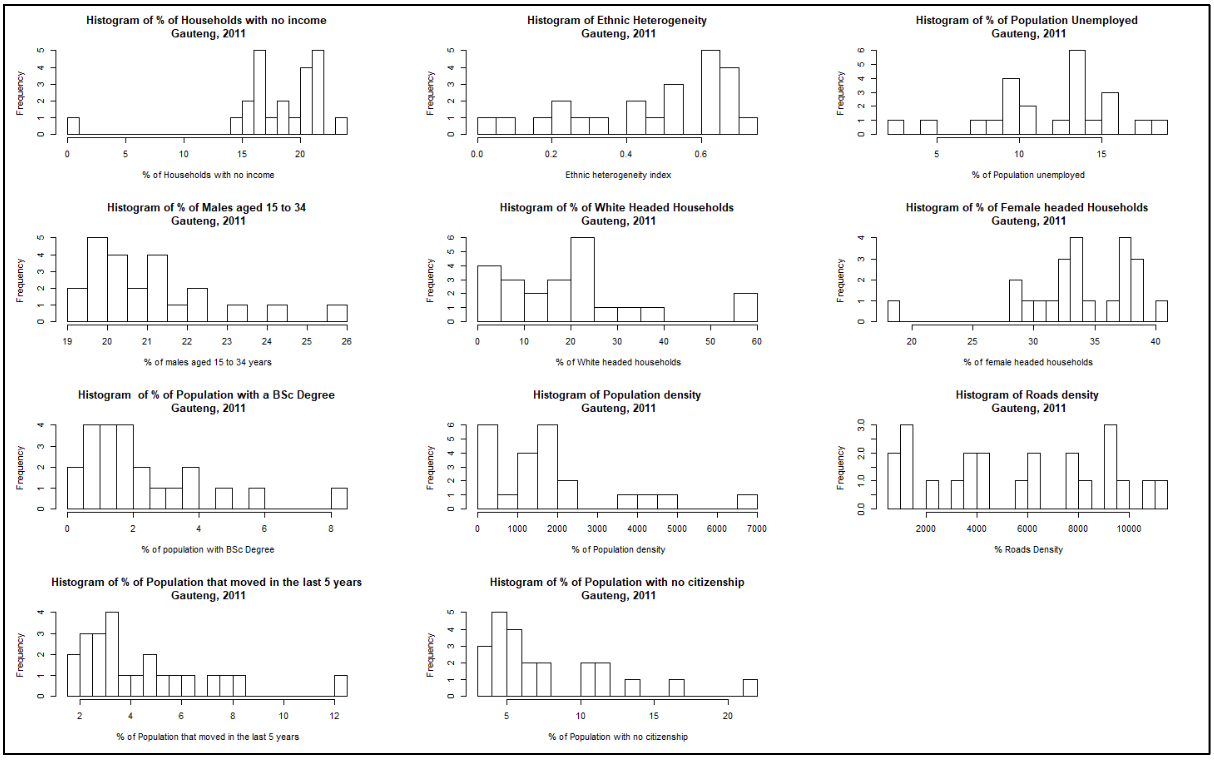
7.3 Appendix C- Cluster Correlation Matrix

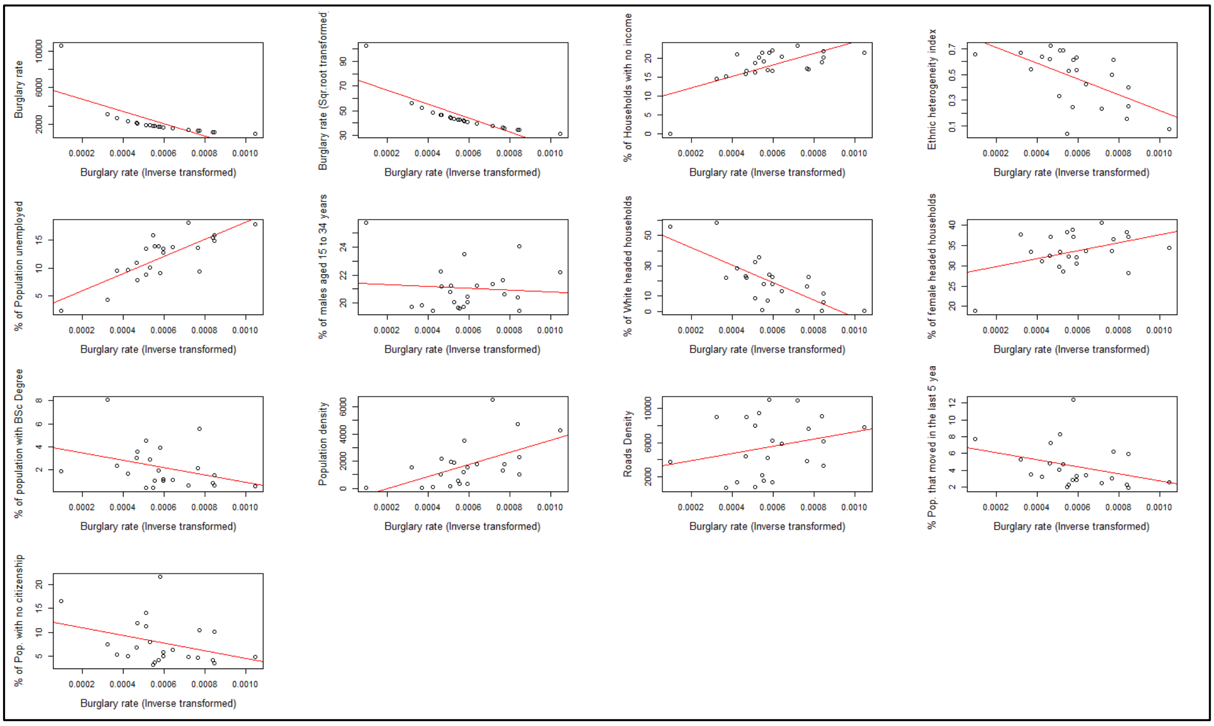
	Variable	A	B	C	D	E	F	G	H	J	K	L	M
A	Population density	1											
B	Roads Density	**0.81**	1										
C	Percentage of female-headed households	0.48	0.34	1									
D	Percentage of White-headed households	-0.38	0.1	**-0.5**	1								
E	Percentage of population unemployed	0.41	-0.07	**0.42**	-0.93	1							
F	Percentage of males aged between 15 and 34 years	0.19	0.25	**-0.52**	0.23	-0.27	1						
G	Percentage of households with no income	**0.42**	0.2	**0.56**	-0.66	**0.73**	**-0.44**	1					
H	Percentage of population with a BSc Degree	-0.03	**0.43**	0.15	0.69	**-0.71**	0.01	-0.28	1				
J	Percentage of people who moved in the last 5 years	0.06	**0.45**	-0.2	0.52	**-0.65**	**0.6**	-0.31	**0.58**	1			
K	Percentage of people with no citizenship	0.04	0.36	-0.32	0.48	**-0.61**	**0.67**	-0.38	**0.43**	**0.96**	1		
L	Ethnic heterogeneity index	-0.36	0.1	**-0.41**	0.81	**-0.8**	0.14	-0.39	**0.63**	**0.58**	**0.5**	1	
M	Burglary rates per 100 000 households	-0.34	-0.16	**-0.67**	0.65	**-0.68**	**0.53**	**-0.89**	0.11	0.31	**0.42**	0.33	1
N	Burglary rates per 100 000 households (log transformed)	-0.48	-0.22	**-0.59**	0.76	**-0.8**	0.33	**-0.84**	0.26	0.34	**0.41**	**0.49**	**0.93**
P	Burglary rates per 100 000 households (square root transformed)	-0.41	-0.19	**-0.65**	0.72	**-0.75**	**0.45**	**-0.89**	0.18	0.33	**0.42**	**0.41**	**0.99**

7.4 Appendix D: Cluster summary statistics plots






Histogram of % of Households with no income
Gauteng, 2011
Histogram of Ethnic Heterogeneity
Gauteng, 2011
Histogram of % of Population Unemployed
Gauteng, 2011
Histogram of % of Males aged 15 to 34
Gauteng, 2011
Histogram of % of White Headed Households
Gauteng, 2011
Histogram of % of Female headed Households
Gauteng, 2011
Histogram of % of Population with a BSc Degree
Gauteng, 2011
Histogram of Population density
Gauteng, 2011
Histogram of Roads density
Gauteng, 2011
Histogram of % of Population that moved in the last 5 years
Gauteng, 2011
Histogram of % of Population with no citizenship
Gauteng, 2011

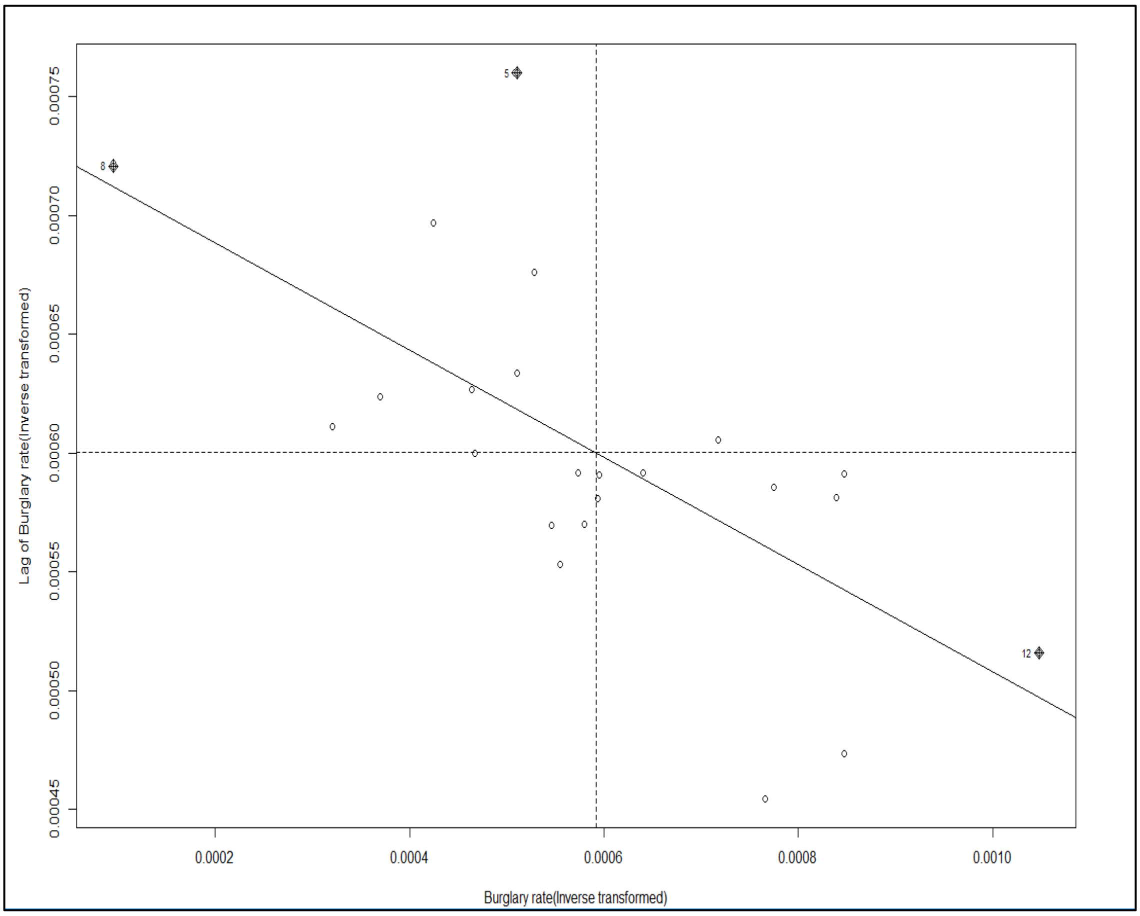

Burglary rate (Inverse transformed)

7.5 Appendix E: Cluster; dependant variable normality tests summary and Moran's I significance p-values.

	Shapiro-Wilk Test Statistic	P-value
Burglary rates per 100 000 households	0.45	0.00000003
Burglary rates per 100 000 households (log transformed)	0.8	0.0004
Burglary rates per 100 000 households (square root transformed)	0.62	0.0000

	Moran's, I statistic	p-value
Queen	-0.2256	0.9
IDW (-6)	-0.35742	1

7.6 Appendix F: Cluster Moran's I scatter plot



7.7 Appendix G: R^2 comparison between significant variables in the model

	Dependent variable:			
	Log (Burglary Rates per 1000 households)			
	(1)	(2)	(3)	(4)
(1) Ethnic Heterogeneity index	1.289*** (0.127)			
(2) Roads Density		-0.00004*** (0.00001)		
(3) Percentage of people that moved in the last 5 years			-0.008 (0.013)	
(4) Percentage of population unemployed				-0.070*** (0.007)
Constant	.347*** (0.062)	3.152*** (0.083)	2.891*** (0.076)	3.647*** (0.082)
Observations	143	143	143	143
R^2	0.423	0.115	0.003	0.446
Adjusted R^2	0.419	0.108	-0.004	0.442
Residual Std. Error (df = 141)	0.433	0.537	0.570	0.424
F Statistic (df = 1; 141)	103.400***	18.280***	0.378	113.700***

8 References

Akaike, H. (1974) 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723. doi: 10.1109/TAC.1974.1100705.

Andresen, M. A. (2006) 'Crime measures and the spatial analysis of criminal activity', *British Journal of Criminology*, 46(2), pp. 258–285. doi: 10.1093/bjc/azi054.

Andrienko, G., Andrienko, N. and Savinov, A. (2001) 'Choropleth Maps : Classification revisited'. German National Research Center for Information Technology. Available at: Url: http://ais.gmd.de/KD/.

Anselin, L. (1995) 'Local Indicators of Spatial Association-LISA', *Geographical Analysis*, 27(2), pp. 93–115.

Anselin, L. *et al.* (1996) 'Simple diagnostic tests for spatial dependence', *Regional Science and Urban Economics*, 26(1), pp. 77–104. doi: 10.1016/0166-0462(95)02111-6.

Anselin, L. (2005) 'Exploring Spatial Data with GeoDa: A Workbook', *Geography*, p. 244.

Anselin, L., Syabri, I. and Kho, Y. (2004) 'GeoDa : An Introduction to Spatial Data Analysis'. Urbana, Champaign, pp. 1–18.

Baybak, M. (2004) 'What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models', *Psychosomatic Medicine*, (131), pp. 4–5. doi: 10.1097/01.psy.0000127692.23278.a9.

Bernasco, W. and Block, R. (2009) 'Where offenders choose to attack: A discrete choice model of robberies in Chicago', *Criminology*, 47(1), pp. 93–130. DOI: 10.1111/j.1745-9125.2009.00140.x.

Bernasco, W. and Elfers, H. (2011) 'Handbook of Quantitative Criminology', *handbook of Quantitative Criminology*, (Chapter 33), pp. 699–724. DOI: 10.1007/978-0-387-77650-7.

Bernasco, W. and Luykx, F. (2003) 'Effects of Attractiveness, Opportunity and Accessibility to Burglars on Residential Burglary Rates of Urban Neighborhoods', *Criminology*, 41(3), pp. 981–1002. doi: 10.1111/j.1745-9125.2003.tb01011.x.

Bernasco, W. and Nieuwbeerta, P. (2005) 'How do residential burglars select target areas? : A new approach to the analysis of criminal location choice', *British Journal of Criminology*, 45(3), pp. 296–315. doi: 10.1093/bjc/azh070.

Bhorat, H. *et al.* (2017) *The Socio-Economic Determinants of Crime in South Africa: An Empirical Assessment*. Cape Town. Available at: https://ideas.repec.org/p/ctw/wpaper/201704.html.

Brantingham, P. L. and Brantingham, P. J. (1975) 'Spatial Patterning of burglary crime 1975', *Environmental Criminology*, pp. 11–23.

Brantingham, P. L. and Brantingham, P. J. (1993) 'Nodes, paths and edges: Considerations on the complexity of crime and the physical environment', *Journal of Environmental Psychology*, 13(1), pp. 3–28. doi: 10.1016/S0272-4944(05)80212-9.

Breetzke, G. (2010a) 'A socio-structural analysis of crime in the city of Tshwane, South Africa', *South African Journal of Science*, 106(11–12). DOI: 10.4102/sajs.v106i11/12.223.

Breetzke, G. (2010b) 'Modeling violent crime rates: A test of social disorganization in the city of Tshwane, South Africa', *Journal of Criminal Justice*. Elsevier Ltd, 38(4), pp. 446–452. DOI: 10.1016/j.jcrimjus.2010.04.013.

Breetzke, G. D. (2008) *Geo-analysis of offenders in Tshwane : toward an urban ecological theory of crime in South Africa*. Pretoria.

Breetzke, G. D. (2012) 'The effect of altitude and slope on the spatial patterning of burglary', *Applied Geography*. Elsevier Ltd, 34, pp. 66–75. doi: 10.1016/j.apgeog.2011.10.017.

Breetzke, G. D. and Cohn, E. G. (2013) 'Burglary in Gated Communities: An Empirical Analysis Using Routine Activities Theory', *International Criminal Justice Review*, 23(1), pp. 56–74. doi: 10.1177/1057567713476887.

Brown, B. B. and Bentley, D. L. (1993) 'Residential burglars judge risk: The role of territoriality', *Journal of Environmental Psychology*, 13(1), pp. 51–61. doi: 10.1016/S0272-4944(05)80214-2.

Brown, K. (2001) ‘The Determinants of Crime in South Africa’, *The South African Journal of Economics*, 69(2), pp. 269–299.

Brunsdon, C., Fotheringham, S. and Charlton, M. (1998) ‘Geographically Weighted Regression-Modelling Spatial Non-Stationarity’, *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3), pp. 431–443. doi: 10.1111/1467-9884.00145.

Brunsdon, C., Fotheringham, S. and Charlton, M. (2009) ‘Geographically Weighted Regression’, *National Centre for Geocomputation*. doi: 10.1111/1467-9884.00145.

Chang, K. (2006) *Introduction to Geographic Information Systems*. Fourth Edi. McGraw-Hill Higher Education.

Chastain, B., Qiu, F. and Piquero, A. R. (2016) ‘Crime Theory Evaluation Using Simulation Models of Residential Burglary’, *Southern Criminal Justice Association*. doi: 10.1007/s12103-016-9336-8.

Chatterjee, S. and Hadi, A. S. (2006) *Regression Analysis by Example*. doi: 10.1002/0470055464.

Chen, J. *et al.* (2017) ‘Modeling Spatial Effect in Residential Burglary: A Case Study from ZG City, China’, *ISPRS International Journal of Geo-Information*, 6(5), p. 138. doi: 10.3390/ijgi6050138.

Chibba, M. and Luiz, J. M. (2011) ‘Poverty, Inequality and Unemployment in South Africa: Context, Issues and the Way Forward’, *Economic Papers: A journal of applied economics and policy*, 30(3), pp. 307–315. doi: 10.1111/j.1759-3441.2011.00129.x.

Clack, W. (2015) ‘Environmental criminology theories : An analysis of livestock theft cases’, *Southern African Journal of Criminology*, 28(2), pp. 92–106.

Cohen, L. E. and Felson, M. (1979) ‘Social Change and Crime Rate Trends : A Routine Activity Approach.’, *American Sociological Association*, 44(4), pp. 588–608.

Core Team, R. (2017) ‘R: A Language and Environment for Statistical Computing’. Available at: http://www.r-project.org/.

Cozens, P. M. (2011) ‘Urban planning and environmental criminology: Towards a new perspective for safer cities’, *Planning Practice and Research*, 26(4), pp. 481–508. doi: 10.1080/02697459.2011.582357.

Dark, S. J. and Bram, D. (2007) ‘The modifiable areal unit problem (MAUP) in physical geography’, *Progress in Physical Geography*, 31(5), pp. 471–479. doi: 10.1177/0309133307083294.

Davies, T. P. and Bishop, S. R. (2013) ‘Modelling patterns of burglary on street networks’, *Crime Science*, 2(1), p. 10. doi: 10.1186/2193-7680-2-10.

Demombynes, G. and Özler, B. (2003) *Crime and Local Inequality in South Africa*, *Policy Research Working papers*. doi: 10.1596/1813-9450-2925.

Eck, J. E. and Weisburd, D. L. (1995) ‘Crime Places in Crime Theory’, *Crime Prevention Studies*, 4, pp. 1–33. doi: 10.1086/229477.

Environmental Systems Research Institute (2014) ‘ArcGIS Desktop’. Carlifonia: Environmental Systems Research Institute. Available at: http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/what-is-areal-interpolation.htm.

Evans, I. S. (1977) ‘The selection of class intervals’, *Transactions of the Institute of British Geographers*, 2(1), pp. 98–124. doi: 10.2307/622195.

Fanaroff, B. *et al.* (2004) *Social Fabric Crimes: Implications for Johannesburg’s Human Development Agenda*, *Human Development*.

Fotheringham, S., Brunsdon, C. and Charlton, M. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex, England: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Friendly, M. and Denis, D. (2005) ‘The early origins and development of the scatterplot’, *Journal of the History of the Behavioral Sciences*, 41(2), pp. 103–130. doi: 10.1002/jhbs.20078.

Getis, A. and Aldstadt, J. (2004) ‘Constructing the Spatial Weights Matrix Using a Local Statistic’, *Geographical Analysis*, 36(2), pp. 90–104. doi: 10.1007/978-3-642-01976-0_11.

Gómez, C., White, J. C. and Wulder, M. A. (2011) 'Characterizing the state and processes of change in a dynamic forest environment using hierarchical spatio-temporal segmentation', *Remote Sensing of Environment*. Elsevier B.V., 115(7), pp. 1665–1679. doi: 10.1016/j.rse.2011.02.025.

Gotway, C. A. and Young, L. J. (2002) 'Combining Incompatible Spatial Data', *Journal of the American Statistical Association*, 97(458), pp. 632–648. doi: 10.1198/016214502760047140.

Graham, M. H. (2003) 'Confronting multicollinearity in ecological multiple regression', *Ecology*, 84(11), pp. 2809–2815. doi: 10.1890/02-3114.

Grami, A. (2019) *Probability, random variables, statistics, and random processes : fundamentals & applications*. 1st edn. Hoboken, USA: Wiley & Sons.

Griffith, D. (1996) 'Some guidelines for Specifying the Geographic Weights Matrix Contained in Spatial Statistical Models', in *Practical Handbook of Spatial Statistics*. Boca Raton: CRC.

Guo, L., Ma, Z. and Zhang, L. (2008) 'Comparison of bandwidth selection in application of geographically weighted regression: a case study', *Canadian Journal for Research*, 38, pp. 2526–2538. doi:10.1139/X08-091.

Harries, K. D. (2006) 'Property Crimes and Violence in the United States: an analysis of the influence of population density', *International Journal of Criminal Justice Sciences*, 1(2), pp. 24–34.

Hauke, J. and Kossowski, T. (2011) 'Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data', *Quaestiones Geographicae*, 30(2), pp. 87–93. doi: 10.2478/v10117-011-0021-1.

Heiple, E. (2010) *Routine Activities Theory : An Empirical Test in a Rural Setting*. Southern Illinois University Carbondale.

Hipp, J. R. (2007) 'Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point', *American Sociological Review*, 72(5), pp. 659–680. doi: 10.1177/000312240707200501.

Hiropoulos, A. and Porter, J. (2014) 'Visualising property crime in Gauteng: Applying GIS to crime', *SA Crime Quarterly*, (47), pp. 17–28.

Huo, X. *et al.* (2011) 'Spatial Pattern Analysis of Heavy Metals in Beijing Agricultural Soils Based on Spatial Autocorrelation Statistics', *International Journal of Environmental Research and Public Health*, 8, pp. 2074–2089. doi: 10.3390/ijerph8062074.

Jega, I. M., Comber, A. J. and Tate, N. J. (2017) 'A Comparison of Methods for Spatial Interpolation across Different Spatial Scales', *SSRG International Journal of Geo informatics and Geological Science*, 4(3), pp. 12–22. doi: 10.14445/23939206/IJGGS-V4I3P102.

Johnson, S. D. and Bowers, K. J. (2010) 'Permeability and burglary risk: Are cul-de-sacs safer?', *Journal of Quantitative Criminology*, 26(1), pp. 89–111. doi: 10.1007/s10940-009-9084-8.

Kelejian, H. H. and Prucha, I. R. (2010) 'Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances', in *Journal of Econometrics*. Elsevier B.V., pp. 53–67. doi: 10.1016/j.jeconom.2009.10.025.

Kim, H.-Y. (2013) 'Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis', *Restorative Dentistry & Endodontics*, 38(1), p. 52. doi: 10.5395/rde.2013.38.1.52.

Kockelman, K. M., Wang, Y. and Wang, X. (2013) 'The impact of weight matrices on parameter estimation and inference: A case study of binary response using land-use data', *Journal of Transport and Land Use*, 6(3), pp. 75–85.

Lancaster, L. and Kamman, E. (2016) 'Risky localities: Exploring a methodology for measuring socio-economic characteristics of high murder areas', *South African Crime Quarterly*, (56), pp. 27–35. doi: 10.17159/2413-3108/2016/v0n56a51.

Leinhardt, G. and Leinhardt, S. (1980) 'Exploratory Data Analysis : New Tools for the Analysis of Empirical Data', *Review of Research in Education*, 8(1980), pp. 85–157.

Li, T. *et al.* (2007) 'A comparison of spatial disaggregation techniques as applied to population estimation for South East Queensland (SEQ), Australia', *Applied GIS*, 3(9). doi: 10.4225/03/57E9AECEBA789.

Malczewski, J. and Poetz, A. (2005) 'Residential Burglaries and Neighborhood Socioeconomic Context in London Ontario Global and Local Regression Analysis', *The Professional Geographer*, 57(September 2004), pp. 516–529.

Martin, D. (2002) ‘Spatial Patterns in Residential Burglary’, *Journal of Contemporary Criminal Justice*, 18(2), pp. 132–146. doi: 10.1177/1043986202018002002.

Matthews, S. A. and Yang, T.-C. (2008) ‘Mapping the results of local statistics: Using geographically weighted regression’, *North*, pp. 1883–1889. doi: 10.3174/ajnr.A1256.Functional.

Matthews, S. A. and Yang, T.-C. (2012) ‘Mapping the results of local statistics’, *Demographic Research*, 26, pp. 151–166. doi: 10.4054/demres.2012.26.6.

Nakaya, T. (2016) ‘GWR4.09 User Manual’, *GWR 4 Development Team*, p. 40.

O’ Sullivan, D. and Unwin, D. J. (2010) *Geographic Information*. Second Edi. New Jersey: John Wiley and Sons, Inc. doi: 10.1007/978-3-319-22789-4.

Okabe, A. and Sugihara, K. (2012) ‘Network Moran’s I Statistics’, in *Spatial Analysis Along Network: Statistical and Computational Methods*.

Osgood, D. W. (2000) ‘Poisson-Based Regression Analysis of Aggregate Crime Rates’, *Journal of Quantitative Criminology*, 16(1), pp. 21–43.

Qiu, F., Zhang, C. and Zhou, Y. (2012) ‘The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study’, *GIScience & Remote Sensing*, 49(5), pp. 644–663. doi: 10.2747/1548-1603.49.5.644.

Rengert, G. F., Piquero, A. R. and Jones, P. R. (1999) ‘Distance Decay Reexamined’, *Criminology*, 37(2), pp. 427–446. doi: 10.1111/j.1745-9125.1999.tb00492.x.

Sampson, R. J. and Groves, W. B. (1989) ‘Community Structure and Crime: Testing Social-Disorganization Theory’, *American Journal of Sociology*, 94(4), pp. 774–802. doi: 10.1086/229068.

Shmueli, G. (2010) ‘To Explain or To Predict?’, *Statistical Science*, 25(3), pp. 289–310. doi: 10.2139/ssrn.1351252.

Smith, W. R., Torstensson, M. and Johansson, K. (2011) ‘Perceived Risk and Fear of Crime: Gender Differences in Contextual Sensitivity’, *International Review of Victimology*, 8(2), pp. 159–181. doi: 10.1177/026975800100800204.

SAPS (2016) *Crime situation in South Africa*. Available at: https://www.saps.gov.za.

StatsSA (2014) *Exploration of the extent and circumstances surrounding Burglary/Robbery*. Available at: www.statsa.gov.za/report.

StatsSA (2017) *Crime Statistics Series Volume 4, Victims of Crime Survey*. Available at: http://www.statssa.gov.za/publications/Report 03-40-04/Report 03-40-042016.pdf.

Tobler, W. R. (1970) 'A Computer Movie Simulation Urban Growth in Detroit Region', *Economic Geography*, 46, pp. 234–240. doi: 10.1126/science.11.277.620.

Townsley, M. *et al.* (2015) 'Burglar Target Selection: A Cross-national Comparison', *The Journal of Research in Crime and Delinquency*, 52(1), pp. 3–31. doi: 10.1177/0022427814541447.

Vélez, M. B. (2001) 'The role of public social control in urban neighborhoods : A multi-level analysis of Victimization Risk', *Criminology*, 39(4), pp. 837–864.

Voss, P. R. *et al.* (2006) 'County child poverty rates in the US: A spatial regression approach', *Population Research and Policy Review*, 25(4), pp. 369–391. doi: 10.1007/s11113-006-9007-4.

Waller, L. A., Hill, E. G. and Rudd, R. A. (2006) 'The geography of power : Statistical performance of tests of clusters and clustering in heterogeneous populations', *Statistics in Medicine*, (September 2005), pp. 853–865. doi: 10.1002/sim.2418.

Weir-Smith, G. (2016) 'Changing boundaries: Overcoming modifiable areal unit problems related to unemployment data in South Africa', *South African Journal of Science*, 112(3–4), pp. 1–8. doi: 10.17159/sajs.2016/20150115.

Yan, J. (2007) 'Spatial Stochastic Volatility for Lattice Data', *Journal of Agricultural, Biological and Environmental Statistics*, 12(1), pp. 25–40.

Zhang, H., Suresh, G. and Qiu, Y. (2012) 'Issues in the aggregation and spatial analysis of neighborhood crime', *Annals of GIS*, 18(3), pp. 173–183. doi: 10.1080/19475683.2012.691901.

Zhang, L., Messner, S. F. and Liu, J. (2007) 'A Multilevel analysis of the risk of household Burglary In the City of Tianjin, China', *British Journal of Criminology*, (August), pp. 918–

937. doi: 10.1093/bjc/azm026.