Modelling Internal Migration in South Africa

School of Statistics and Actuarial Science

Submitted in partial fulfilment of the degree of
Master of Science in Mathematical Statistics

BY

Xolani Jozi

Supervisor

C Chimedza

Dedication

I would like to dedicate this project to my parents, Mr Mbulelo and Mrs Lindelwa Vena who always believed in me.

Acknowledgement

The work presented in this study was carried out at the University of Witwatersrand in partial fulfilment of the requirement for the Masters of Science degree in Mathematical Statistics. There are people who have contributed to this work and who deserve my gratitude. First, I would like to thank my supervisor, Charles Chimedza, for his support and encouragement throughout, from the work leading to this thesis to its submission. I am also grateful to members of the School of Statistics and Actuarial Science, for their unwavering assistance in strengthening matters relating to this study. I am indebted my friend Dr Matshawe Tukulula for his assistance in proof reading my work. I do not want to forget my family for their support, more specially, my mother and my father. Finally, I would like to acknowledge my colleagues at Statistics South Africa for their support and for making the data available for this study.

Abstract

The aim of this study was to model internal migration in South Africa using the 2011 Census data. The net-internal migration was modelled in the district municipalities of South Africa using Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR). In this study, the following global and local modelling techniques were used, Gravity, Poisson, Negative Binomial (NB), Gamma, and GWR model (local model). Poisson and NB failed to fit the migration data, while the Gamma model managed to fit the data reasonably well. The GWR model performed better than OLS regression in modelling net-internal migration in district municipalities of South Africa.

The results from these models revealed that there was a strong relationship between internal migration and economic variables, as well as living conditions and demographic variables. The Monte Carlo significance test results showed that the parameters of the white population vary significantly across space.

The results of the study signal that the differences in social and economic disparities in the district municipalities of South Africa are the drivers of internal migration.

Contents

1 Introduction **1**

1.1 Introduction 1

1.2 Motivation of the study 5

1.3 Variables used in the study 6

1.4 Aims 8

1.5 Objectives 8

1.6 Limitations of the Study 8

1.7 Lay out of the project 9

2 Literature Review **10**

2.1 Introduction 10

2.2 Literature Review 10

2.3 Modelling Techniques: Theoretical review 14

2.3.1 Gravity model 15

2.3.2 Poisson model 18

2.3.3 Negative Binomial (NB) model 20

2.3.4 Gamma model 22

2.4 Local Modelling 25

2.4.1 Geographically Weighted Regression (GWR) model . . . 25
2.4.2 Moran's I . . . 29
2.4.3 Monte Carlo significance Test . . . 31
2.5 Modelling Diagnostics . . . 32
2.5.1 Variance Inflation (VIF) . . . 32
2.5.2 Shapiro-Wilk test . . . 33
2.5.3 Jarque-Bera-Test (JB-Test) . . . 34
2.5.4 Durbin-Watson (D-W) . . . 35
2.5.5 Koenker-Breusch Pagan (Koenker-BP) Test . . . 37
2.5.6 Test for significance of linear regression model . . . 38
2.5.7 Reset Test . . . 39
2.5.8 Ord plot . . . 40
2.5.9 Deviance of the generalised linear models . . . 40
2.5.10 Model selection . . . 41
2.6 Cluster Analysis (CA) . . . 43
2.6.1 Cluster Estimation: Silhouette Width . . . 43
2.6.2 The K-means clustering . . . 44
2.7 Summary . . . 45

3 Methodology 47

3.1 Introduction . . . 47
3.2 Data . . . 47
3.3 Description of the data . . . 49
3.4 Modelling . . . 49
3.5 Model Estimation . . . 52
3.6 Model Performance . . . 52

3.7 Summary 53

4 Analysis **55**

4.1 Introduction 55

4.2 Description of the data 55

4.2.1 In-migration, Out-migration and Net-Internal Migration in KZN district municipalities 56

4.2.2 In-migration, Out-migration and Net-Internal Migration in EC district municipalities 57

4.2.3 In-migration, Out-migration and Net-Internal Migration in WC district municipalities 58

4.2.4 In-migration, Out-migration and Net-Internal Migration in FS district municipalities 59

4.2.5 In-migration, Out-migration and Net-Internal Migration in GP district municipalities 59

4.2.6 In-migration, Out-migration and Net-Internal Migration in MP district municipalities 60

4.2.7 In-migration, Out-migration and Net-Internal Migration in LP district municipalities 61

4.2.8 In-migration, Out-migration and Net-Internal Migration in NW district municipalities 61

4.2.9 In-migration, Out-migration and Net-Internal Migration in NC district municipalities 62

4.3 Migration patterns 63

4.3.1 In-migration to Tshwane Metropolitan Municipality (TSH) 63

4.3.2 In-migration to Ekurhuleni Metropolitan Municipality (EKU) 64

4.3.3 In-migration to the City of Cape Town (CPT) 65
4.3.4 Out-migration from O.R Tambo (DC15) 65
4.3.5 Out-migration from Nelson Mandela Metropolitan Municipality (NMA) 66
4.3.6 Out-migration from Capricorn (DC35) 67
4.4 Modelling Results 68
4.4.1 Gravity model 69
4.4.2 Extended Gravity model 70
4.4.3 Results: Reset Test 72
4.4.4 Nonlinear model 73
4.4.5 Comparison of the models that Predict Out-migration . . . 75
4.4.6 Comparison of the models that Predict In-migration 76
4.4.7 Diagnostic plots of the OLS models 76
4.4.8 Normality Assessment 78
4.4.9 Testing for Autocorrelation 78
4.4.10 Outlier Identification for OLS model 79
4.4.11 Poisson model results 81
4.4.12 Overdispersion Test results 82
4.4.13 Negative Binomial (NB) model results 83
4.4.14 Gamma model results 84
4.4.15 Assessment of the models 86
4.4.16 Diagnostic plot of the Gamma models 87
4.4.17 Diagnostic plot: Ord plot 89
4.5 Modelling Net-Internal Migration 89
4.5.1 Results of the OLS model 89

4.5.2 Assessment of the Net-Internal Migration model (OLS) . . 91
4.5.3 Testing the assumptions of the linear model 92
4.5.4 Results for Geographically Weighted Regression (GWR) model 93
4.5.5 Checking multicollinearity 94
4.5.6 Testing for the misclassification of the GWR model 95
4.5.7 Diagnostics statistics for GWR 96
4.5.8 Results of the Monte Carlo Significance Test 98
4.5.9 Local Regression coefficient (pow) 98
4.5.10 Cluster Analysis (CA) results 99
4.5.11 Summary 101

5 Conclusion and Recommendation 103
5.1 Introduction 103
5.2 Summary and discussion 103
5.3 Conclusion 109
5.4 Recommendation 110

References 112

Appendices 123

A Net-Internal Migration by Province 124

B Net-Internal Migration 125

C Descriptive 127

D The R code 129

E Sample results **137**

List of Figures

1.1 Net-Internal Migration by Province 6

4.1 Migration distribution in the district municipalities 56
4.2 In-migration, Out-migration and Net-Internal Migration in KZN district municipalities 57
4.3 In-migration, Out-migration and Net-Internal Migration in EC district municipalities 58
4.4 In-migration, Out-migration and Net-Internal Migration in WC district municipalities 58
4.5 In-migration, Out-migration and Net-Internal Migration in FS district municipalities 59
4.6 In-migration, Out-migration and Net-Internal Migration in GP district municipalities 60
4.7 In-migration, Out-migration and Net-Internal Migration in MP district municipalities 60
4.8 In-migration, Out-migration and Net-Internal Migration in LP district municipalities 61

4.9 In-migration, Out-migration and Net-Internal Migration in NW district municipalities 62
4.10 In-migration, Out-migration and Net-Internal Migration in NC district municipalities 62
4.11 In-migration in Tshwane Metropolitan municipality 63
4.12 In-migration in Ekurhuleni Metropolitan municipality 64
4.13 In-migration in the City of Cape Town Metropolitan municipality . 65
4.14 Out-migration in O.R Tambo District Municipality 66
4.15 Out-migration in Nelson Mandela Metropolitan municipality(NMA) 67
4.16 Out-migration from Capricorn District municipality 68
4.17 Diagnostic: Nonlinear model (In-migration) 77
4.18 Diagnostic: Nonlinear model (Out-migration) 77
4.19 Gamma model (In-migration) 87
4.20 Plot of the deviance residuals against $2\ln(\hat{y}_i)$: In-migration 88
4.21 Gamma model (Out-migration) 88
4.22 Plot of the deviance residuals against $2\ln(\hat{y}_i)$: Out-migration . . . 88
4.23 Ord plot 89
4.24 Diagnostic plot of the Net-Internal Migration model 91
4.25 Investigating possible spatial multicollinearity 94
4.26 Spatial Autocorrelation results 95
4.27 Local R^2 96
4.28 GWR White population (pow) Coefficient 99
4.29 Silhouette width 100
4.30 Cluster plot 101

List of Tables

1.1 Provincial Migration 3

1.2 Variables relating to district municipalities i (origin) and j (destination) 6

1.3 Variables relating to a specific district municipality: the subscript of the variables indicates the district municipality 7

1.4 Variables relating to the percentage of households in a specific district municipality with specific attributes 7

1.5 Variables relating to a rate in a specific district municipality with specific attributes 7

2.1 Analysis of Variance (ANOVA) for Significance of Regression in Multiple Regression 39

4.1 Variables used for the Gravity model 69

4.2 Results of the Gravity model 70

4.3 Variables used for the Extended Gravity model 70

4.4 Results of the Extended Gravity model 72

4.5 Results of the Nonlinear model 75

4.6 Results of the model selection for Out-migration 76
4.7 Results of the model selection for In-migration 76
4.8 Results of the normality test 78
4.9 Results of the Autocorrelation test 79
4.10 Results of the outlier identification: In-migration 80
4.11 Results of the outlier identification: Out-migration 81
4.12 Outliers removed: the normality test 81
4.13 Outliers removed: the autocorrelation test 81
4.14 Results of the Poisson model 82
4.15 Overdispersion test results 83
4.16 Results of the Negative Binomial model 84
4.17 Results of the Gamma model 86
4.18 Diagnostics of the NB and Gamma model 87
4.19 Results of the Net-Internal Migration model (OLS) 90
4.20 Results of the outlier identification 91
4.21 Results of the Net-Internal Migration model assessment 92
4.22 Results of the model assumptions 92
4.23 Testing the significance of autocorrelation 93
4.24 Outlier removed: results of the Net-Internal Migration model assessment 93
4.25 Outlier removed: Testing the significance of autocorrelation . . . 93
4.26 Global Moran's I Summary 95
4.27 Diagnostic results of the GWR 97
4.28 ANOVA results 97
4.29 Results for the spatial variability of coefficients 98

4.30 Cluster means 100

A.1 Net-Internal Migration by Province 124

B.1 Net-Internal Migration in the district municipalities 126

C.1 Descriptive statistics 127

C.2 Descriptive statistics of the log variables 128

Chapter 1

Introduction

1.1 Introduction

Internal migration is generally defined as the movement of people from one place to another within the same country. Internal migration can be broken down into people moving between provinces, districts, towns and villages. This study is based on the 2011 census data. Statistics South Africa (Stats SA) defined internal migration, as the change in a person's usual municipality or municipality of residence during the 10 years period (2001-2011) preceding the collection of census data in 2001.

The questions below (P-10 to P-11d), are migration questions that were included in the 2011 census questionnaire.

P-10: Does $(name)$ usually live in this household for at least four nights a week and has done so for the last six months? OR intends to live in this household for

at least four nights a week for the next six months?

P-10a: In which province does $(name)$ usually live?

P-10b: In which municipality or magisterial district of usual residence does $(name)$ usually live?

P-10c: In which city/town does $(name)$ usually live or what is the nearest city/town

P-11: Has $(name)$ been living in this place since October 2001?

P-11a: When did $(name)$ move to this place?

P-11b: In which province did $(name)$ live before moving to this place?

P-11c: In which municipality or magisterial district did $(name)$ live before moving to this place?

P-11d: In which city/town did $(name)$ live before or what was the nearest city/town?

Questions P-10 to P-11d are used for planning and for measuring internal migration in South Africa. This subject has been debated by researchers all over the world, including those in South Africa, because of its relevance to social and economic development (Pezic, 2009). Internal migration can be explained by variations in wages and employment opportunities that exist between regions or sectors (Mulhern and Watson, 2009).

Vargas-Silva (2011) states that not every characteristic of migration is advantageous for developing countries. Migration may enforce a high cost for developing countries by leaving the country not having the human capital necessary to achieve endless economic growth. This human capital movement may impose a notable

economic responsibility for developing countries as migrants take with them the value of their education, which is often sponsored by a government with few resources.

This study seeks to unpack internal migration using various models in a quest to uncover information that is not really addressed in reports concerning censuses and various surveys. This study will focus on people's movements between district municipalities.

South Africa has nine provinces, which are subdivided into 52 district municipalities. According to the 2011 Census, the population of South Africa stands at 51.8 million and out of this, over 2,19 million people are internal migrants [1] [2] at the provincial level, see Table 1.1 [3].

Table 1.1: Provincial Migration

Province of previous residence	Province of usual residence									In-Migration
	WC	EC	NC	FS	KZN	NW	GP	MP	LP	
WC	5158316	40152	10566	5155	9221	5039	50694	4759	3381	318917
EC	170829	6250135	5081	15542	73831	32341	117964	12001	8877	120940
NC	17577	4077	1054841	8559	5708	11478	16019	4202	1907	55412
FS	12644	8155	7103	2524282	8881	24090	74387	10859	5283	92622
KZN	21857	19178	2437	11481	9812129	8655	184337	28904	4719	174228
NW	6013	3085	17000	9917	3882	3146255	103550	8495	14066	196780
GP	74915	40161	9446	31455	55620	75260	10416258	61269	54145	953024
MP	7256	3390	1932	5032	12511	13091	122578	3723843	25299	169981
LP	7826	2742	1847	5481	4574	26826	283495	39492	5088084	117677
Out-Migration	128967	436466	69527	151402	281568	166008	402271	191089	372283	

[1] In-Migration refers to people moving into one place from another place within a country and Out-Migration is people moving out of one to another place within a country. WC =Western Cape, EC =Eastern Cape, NC =Northern Cape, FS =Free State, KZN =KwaZulu-Natal, NW =North West, GP=Gauteng Province, MP =Mpumalanga and LP=Limpopo.

[2] Source: Statistics South Africa 2011 Census data

[3] Table 1.1 excluded 2396838 cases where the province was Outside South Africa, unspecified and do not know.

The study of internal migration is not a new research area, researchers have addressed this issue in other countries such as those in Europe, North America and China by using a variety of modelling techniques. For instance, Vanderkamp (1968) used time series analysis (from 1947 to 1966) to study the time pattern of migration in Canada. Gale, Hubert, Tobler and Golledge (1983) used ten migration models, such as the Wilson's entropy model, Quadratic programming solution, Anova model as well as variations of push-pull models. The aim of their study was to compare the performance of the models in terms of how well they represent the migration data.

Silvestre (2005) used econometric models to examine the causes and the effects of internal migration in Spain for the period 1877-1930. He concluded that, economic factors are significant in explaining internal migration in Spain. Fan (2005) used the gravity model to investigate internal migration in China.

In South Africa on the other hand, there are very few modelling studies undertaken in order to shed light on internal migration. The only information available on migration is the administrative data which Stats SA, in collaboration with the Department of Home Affairs, collects. This however, relates only to international arrivals and departures.

In this study, the source of data that was used for modelling is the 2011 census, the main reason being the fact that censuses can provide migration totals at lower geographical levels such as municipalities and ward levels. However, the census has one limitation which is the lack of detail regarding why a specific person or a household decided to migrate to a specific district municipality.

Besides the modelling approach, this study attempts to use cluster analysis to characterise internal migrants. This is done by using the K-means clustering algorithm.

1.2 Motivation of the study

There are very few attempts made at utilising models to explore internal migration in South Africa and there is not much being done in applying models to study internal migration. Studies done by Bouare (2000-2001) model internal migration in South Africa at the provincial level. Statistics agencies like Stats SA produce basic descriptive statistics to explore internal migration. The main motivation of this study is to describe and advance the use of mathematical modelling tools in investigating internal migration in South Africa.

The internal migration results can assist the South African government in making informed decisions about the distribution of government resources, such as housing, schools, health resources, job creation and the distribution of police stations in the district municipalities.

Another important motivation of this study is to investigate internal migration at the district level. Figure 1.1 and Appendix A, show the net-internal migration (the difference between in-migration and out-migration) at provincial level. Gauteng and Western Cape Provinces have a positive net internal migration that is increasing between the two censuses. Limpopo, Eastern Cape, Free State, KwaZulu-Natal, Mpumalanga and Northern Cape have a negative net-internal migration. The 2011 census show that Mpumalanga had a less outflow than 2001 census.

North West had an outflow of migrants based on the 2001 census questions, and an inflow based on the 2011 census questions.

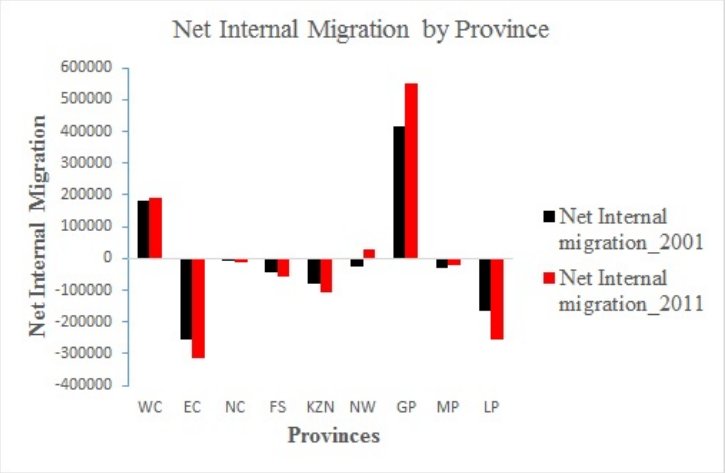

Figure 1.1: Net-Internal Migration by Province

1.3 Variables used in the study

This section defines the variables used in this study.

Table 1.2: Variables relating to district municipalities i (origin) and j (destination)

Variables	Definition
M_{ij}	Migration flow between district i and j
D_{ij}	Distance (travelled by a land transport) between district municipalities i and j

Table 1.3: Variables relating to a specific district municipality: the subscript of the variables indicates the district municipality

Variables	Definition
pop	Population size
poc	Coloured Population
poI	Indian Population
GDP	Gross Domestic Product
CPI	Consumer Price Index
ad	Adult Population
pod	Population density
pob	Black Population
pow	White Population
avh	Average annual household income

There is no data about GDP at the district level. This study used the provincial GDP values. The provincial GDP data is a proxy for districts. Statistics South Africa (2012(a)) provides the mid-points for the household income ranges and the average annual household income in this study is calculated from the mid-points.

Table 1.4: Variables relating to the percentage of households in a specific district municipality with specific attributes

Variables	Definition
ocr	Tenure status: Occupied rent free
ofp	Tenure status: Owned and fully paid off
$Inftrdwell$	Informal or Traditional dwelling
pwa	Access to tap water
$nacstvs$	No access to services
tre	Tenure Status (Rentals)

Table 1.5: Variables relating to a rate in a specific district municipality with specific attributes

Variables	Definition
emp	Employment rate
ump	Unemployment rate
$illitrt$	Illiteracy rate

1.4 Aims

The main aim of this study is to integrate the use of descriptive analysis with modelling approaches in studying internal migration in South Africa. Moreover, the findings of this study are anticipated to make a contribution to internal migration studies in South Africa.

1.5 Objectives

The main objectives of this study are to:

1. Develop a model that explains migration in the district municipalities of South Africa.
2. Investigate the stationarity of the (β) parameters of the Geographically Weighted Regression (GWR) model
3. Profile migrants using selected demographic characteristics and household variables.

1.6 Limitations of the Study

The 2011 census data does not provide detailed information about internal migrants. For instance, it does not give reasons for migrating or the motive for choosing one district as opposed to the other. In light of this limitation, the study assumes that, the variables used in our models can explain internal migration.

Fan (2005) observed that provinces in China, with larger Gross Domestic Product (GDP) per capita had the highest in-migration. In this study we aim to use GDP

to give an indication of economy activity , but it is only available at provincial level, not at district municipality level. Another aim of this study is to use the variable wage but the variable was not measured during the 2011 census and this study used the Average household income at the district level. This variable was derived from annual household income at the district level using the mid point values given by Statistics South Africa, 2012(a) for the income categories.

This study does not take into consideration the issue of temporary migration (it assumes temporary migration is negligible) between the district municipalities in South Africa, therefore migration may be underestimated. The migration data is provided at the Local Municipalities level. As South Africa has 234 local municipalities (it was a 234 * 234 matrix), these were collapsed to districts. The objective of this project was to model migration at the district level. This study had to derive a 52 * 52 matrix, which consists of district municipalities movements, and all that was done manually in Microsoft Excel.

1.7 Lay out of the project

This report is organised into five chapters. The first chapter is an introduction of the internal migration. The second chapter, is the literature review. Chapter three is the Methodology of the analysis. Chapter four is the analysis of the study. Chapter five is the conclusions and the recommendations of the study.

Chapter 2

Literature Review

2.1 Introduction

This chapter is divided into six main sections. Section 2.2 reviews the methods available in literature, section 2.3 and section 2.4 is a review of the theory behind migration models. Section 2.5 reviews the modelling diagnostics. Section 2.6 reviews the cluster analysis. Section 2.7 concludes the chapter.

2.2 Literature Review

Studies done by Kok and Collinson (2006) and Stats SA in South Africa are descriptive in nature with regards to their analysis on internal migration. There are however, very few attempts at using models to investigate internal migration. This is not only an issue in South Africa, but also in other developing countries like China, where it has been observed that relatively little effort has been made to use

the modelling approach to study migration (Fan, 2005). The Gravity model, Poisson, Negative Binomial (NB) and Geographically Weighted Regression (GWR), etc., are popular models in migration literature. In this study, the above listed models will be used to study internal migration in South Africa.

Byrne and Pezic (2004), Pezic (2009), and Islam and Siddiqi (2010) stated that internal migration has an impact on demographic factors, such as age structure, sex ratios and population size, etc. Further more, Islam and Siddiqi (2010) stated that migration changes the distribution of the population both at the place of origin and the place of destination, which can subsequently affect the economy or demography of a country positively or negatively. The regions of departure lose labour force participants while the social and economic infrastructure in regions of arrival may have challenges keeping up with rapidly growing population (Henry, Boyle and Lambin, 2003).

Publishing country internal migration statistical estimates and trends is important, because such results can assist in rationalising and distributing a country's resources. These estimates can also help in explaining and perhaps shed the light on why certain areas lose population through migration, while others are gaining (Congdon, 2010).

The causes of internal migration vary from country to country. Kok and Collinson (2006) stated that the causes of migration are theoretically complex and multilevel in nature, resulting in these being difficult to determine and being harder to generalise. However, some studies believe that migration is caused by differences in socio-economic conditions that can be categorised as either pull or push factors.

The push factors in this case refer to reasons for leaving one place because of difficulties encountered, such as food shortage, war, flood, etc., while the pull factors refer to reasons for moving in to another place because of desirable factors such as a favourable climate, better food supply, better employment opportunities, etc.

Chapter 2 of the South African Bill of Rights 21(3) by The Constitution of Republic of South Africa (1996) provides for freedom of movement and residence by stating that, "Every South African citizen has the right to enter, to remain in and to reside anywhere in the Republic". This bill can affect the budget allocation of other district municipalities, since as more people migrate to and from some district municipalities, the demand for essential services such as housing and water will vary. Henry, Boyle and Lambin (2003) suggested that places with higher migration rates cause problems for spatial planning efforts as there is a need for predicting migration flows.

Vargas-Silva (2013) stated that The Office for Budget Responsibility in UK noted that higher net migration reduces pressure on government debt over time. This observation was based on the fact that incoming migrants were assumed to be more likely to be of working age than the population in general. The author further states that in the short-term migration increases tax receipts.

In South Africa, Bouare (2000-2001) studied the determinants of internal migration, using the Extended Gravity model as a modelling technique based on the 1996 census as a source of data. The author noted that, the relative GDP, relative unemployment, relative number of reported crimes and kinship (the ratio of the greatest size of one dominant population group in province j to the population

size in province i) to determine internal migration in South Africa. In addition, Fan (2005) observed a similar pattern in China, where the GDP and migration stock (number of people living in a different place than where they were born) were the most significant factors in explaining internal migration for the period of 1985-2000.

Czaika and De Haas (2013) said that kinship networks and transnational ties among migrant communities tend to facilitate migration. They further noted that migration networks lower costs for job searches, housing and child care and can reduce vulnerability to exploitation and crime. Kinship networks are particularly strong and effective if their internal composition is characterised by similarities of language, ethnicity and social class. Pedersen et al (2008) and Mayda (2009) (as cited in, Czaika and De Haas (2013, page 4) noted "there is evidence that networks, cultural and historical links have a robust, and strong positive effect on migration."

Juarez (2000) studied the economic determinants of the Spanish interregional migration based on the labour force and concluded that unemployment increases outmigration. Studies done by Czaika and De Haas (2013) also confirmed that economic and labour market factors were major determinants of migration in the UK. However, Maza and Villaverde (2004) studied interregional migration in Spain and their study revealed that unemployment does not affect net migration rates. Vanderkamp (1968) on the other hand observed that unemployment had a significant negative effect on the number of migrants between regions in Canada and he further observed that this relationship is not adequately captured by regional unemployment differentials. Juarez (2000) further comments that the rate of change

of relative wages is a significant factor in explaining migration in Spain. Rogers (1967) observed the same findings in California. Faggian and Royuela (2010), and Czaika and De Haas (2013) noted that when wages and unemployment rates differ among regions, people react to these regional differences by migrating to areas where wages are higher and unemployment rate is lower.

The findings from Bouare (2000-2001), Juarez (2000), Maza and Villaverde (2004) and Vanderkamp (1968) concerning employment are not surprising because migration is affected by a variety of factors.

Henry et al (2003) suggested that it is not only the socio-economic variables that determine the cause of internal migration, but environmental variables are also significant in explaining internal migration. Cebula and Alexander (2006) explored internal migration by involving variables such as quality of life and noted that net-state migration is an increasing function of the warmer temperatures and a decreasing function of the presence of hazardous waste sites.

Other studies have attempted to model migration by using socio-economic and demographic variables. For instance, Islam and Siddiqi (2010) identified age, educational qualification, occupation (before migration), income (before migration), and type of family structure (before migration) as having a significant effect on migration.

2.3 Modelling Techniques: Theoretical review

The type of data encountered in the modelling of internal migration are counts/ net counts, where net counts are the difference between in and out migration counts.

Coxe, West and Aiken (2009) defined a count variable as a variable that takes on discrete values (0, 1, 2, 3, .etc).

Boyle (1995), Bouare (2000-2001), Fan (2005), and Faggian and Royuela (2010) used counts as a response variable in modelling internal migration. Pezic (2009) used rates for estimating migrants moving between regions. Nabi (1992) used net internal migration rates to investigate the dynamics of internal migration in Bangladesh.

2.3.1 Gravity model

The Gravity model is expressed as

$$ln(y_{ij}) = \beta_0 + \beta_1 lnp_i + \beta_2 lnp_j + \beta_3 lnD_{ij} + \epsilon_{ij} \tag{2.1}$$

Where y_{ij} is the total migrants from district municipality i to district municipality j, p_i and p_j are population sizes at origin and destination respectively, D_{ij} is the distance between district municipality *i* and district municipality *j*. The ϵ_{ij} is the error term which is assumed to be a normally distributed random variable.

Stats SA has no data about distance. The distance can be obtained from Google Maps (2014) using the Harvesine formula to calculate the distance between district municipalities. The Harvesine formula as seen from Veness (2002-2015) is expressed as

$$D_{ij} = Rc \tag{2.2}$$

where,

$$c = 2a \tan^2(\sqrt{a}, \sqrt{1-a})$$
$$a = \sin^2(\triangle\phi) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\triangle\lambda}{2}\right)$$

lat is a latitude (in degrees), lon is a longitude (in degrees) ϕ is a latitude (in radians), λ is a longitude (in radians), R is an earth radius (mean radius = 6371km) and the angles need to be converted to radians, where

$$R = 6371 \ \ in \ \ km$$
$$\phi_1 = lat1.toRadians()$$
$$\phi_2 = lat2.toRadians()$$
$$\triangle\phi = (lat2 - lat1).toRadians()$$
$$\triangle\lambda = (lon2 - lon1).toRadians()$$

This is implemented in Google Maps (2014).

The parameters (β's) of the Gravity model are estimated using the Ordinary Least Squares (OLS) method as follows,

$$\hat{\beta} = \left(\mathbf{X^T X}\right)^{-1} \mathbf{X^T Y} \tag{2.3}$$

where the independent observations are the columns of $\boldsymbol{X}$ and the dependent observations are in the single column vector $\boldsymbol{Y}$.

This model is widely used in the studies of internal migration and is applied in economics. Bouare (2000-2001) used a Gravity model to investigate internal mi-

gration in South Africa. Fan (2005) further noted that this Gravity model, which is used in many migration studies, performed reasonably well in fitting the migration flows in China.

Although the Ordinary Least Squares (OLS) regression is popular in internal migration studies. Coxe, et al (2009) stated there are concerns regarding its use in modelling a count response variable. They further said the OLS regression can be used to model a discrete outcome with minimal challenge. Gardner, Mulvey and Shaw (1995) said when the mean of the response is low[1], OLS regression tends to produce wrong results including biased standard errors.

The assumptions of the OLS regression are related to the residuals of a model. These assumptions are conditional normality, homoscedasticity (constant variance) and independence. The assumption of log normality is not acceptable for count data because of the inequality of the variance in the error terms (Flowerdew and Aitikin, 1982; and Congdon, 1992).

The use of the logarithmic transformation affects the estimates produced (Flowerdew and Aitkin, 1982). They further said that the regression produces estimates (β-parameters) of the logarithms of the y_{ij}, not of the y_{ij}'s and the antilogarithms of these estimates are biased of y_{ij}. When heteroskedasticity is observed, estimates produced by using log-linearised models are biased and their distorting the interpretation of the model (Santos and Tenreyro, 2006). Another challenge arises from the use of the log transformation when some of the flows are zero. The log of zero cannot be calculated and in fitting the Gravity model. In this study 0.1 is added to zero flows. According to Santos and Tenreyro (2006) this rule will

[1] A mean that is greater than 10, it is relatively high (Coxe, et al, 2009).

lead to inconsistent estimators of the parameters of interest. O'Hara and Kotze (2010) recommend that a count variable should not be analysed by a simple log-transform, instead models such as Poisson and Negative Binomial models should be used.

2.3.2 Poisson model

An alternative model to the Gravity model is the Poisson model. The Poisson model is preferred in the analysis of modelling count response variables (Flowerdew and Amrhein, 1989; and Coxe, et al, 2009). The Poisson model ensures that the conditional mean of λ_{ij} is non- negative, by taking the exponent of the independent variables.

A statistical model for counts is the Poisson model with probability function

$$P_r = \frac{\lambda_{ij}^{y} e^{-\lambda_{ij}}}{y!}, \quad \lambda_{ij} > 0, \quad y = 0, 1, 2, 3, \cdots \tag{2.4}$$

where λ_{ij}=$e^{{x_{ij}}^T \beta}$ is the mean response variable, λ_{ij} is the flow between district *i* and *j*. The term x_{ij} is a vector of predictor variables associated with the origin and the destination district municipalities, and β is a vector of unknown parameters.

Hilbe (2011, page 33) suggested that the estimation of the parameters (β) be done by taking the first partial derivative of the Poisson log likelihood function. The Poisson log likelihood function is given by

$$L(\beta; y) = \sum_{i=1}^{n} \left\{ y \left({x_{ij}}^T \beta \right) - e^{{x_{ij}}^T \beta} - ln\left(y!\right) \right\} \tag{2.5}$$

Taking the first partial derivative of the Poisson log likelihood function with respect to the parameter β, and equating it to zero, gives;

$$\frac{\partial(L(\beta; y))}{\partial\beta} = \sum_{i=1}^{n} \left(y - e^{{x_{ij}}^T \beta}\right) {x_{ij}}^T = 0 \tag{2.6}$$

Solving equation (2.6) yields a Newton Raphson Maximum likelihood estimate $\hat{\beta}$. Therefore the regression model for the Poisson model is expressed as

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \cdots + \beta_l x_{lij} \tag{2.7}$$

The Poisson regression model is a popular modelling technique in examining migration flow for different countries, Flowerdew and Amrhein (1989) used this model in studying migration flow in Canada focusing on the 1985-1986 period. Boyle (1995) also studied rural in-migration in England and Wales for the period 1980-1981 using Poisson regression.

Besides migration, Poisson regression is used as a modelling technique in other disciplines such as Ecology. Reynolds and Fenster (2008) used the Poisson regression model to predict the number of visits to a patch [2] of plants in a half hour duration. In this model the number of visits was the dependent variable and species was the predictor.

The Poisson distribution has one parameter the mean, which is equal to the variance. In most practical cases this property does not hold, in some instances the variance is greater than the mean, when this occurs, we get a phenomenon known as overdispersion. On the other hand if the mean is greater than the variance,

[2] each patch = one experimental unit

underdispersion occurs. Overdispersion is caused by misspecification of the systematic component of the model. Other issues that can cause overdispersion are the presence of outliers in the observed data set and an inappropriate link function for the Poisson model.

Collett (2003) suggests that the modification or omission of outliers may lead to the residual deviance (defined in equation 2.60) being reduced until the overdispersion disappears. Collett (2003) further suggests that the adequacy of the chosen link function can be checked by studying the Index plot of the standardised deviance or likelihood residuals. Miaou (1994), Simonoff (2003) and Hilbe (2011) presented a statistic that can be used in examining overdispersion for the Poisson regression model. The statistic is known as the dispersion parameter (Pearson Goodness-of-fit (χ^2) divided by degrees of freedom), that is $\frac{\chi^2}{n-p}$, where $n-p$ is the degrees of freedom, n is the number of observations and p represents the number of unknown regression parameters in the Poisson model.

A Poisson model having a value of the dispersion parameter greater than 1 is overdispersed and a model with a value below 1 is underdispersed. When the Poisson regression is well defined and it fits the data reasonably, the Pearson dispersion statistic has a value close to 1. When over dispersion is detected in a Poisson model, a Negative Binomial model must be considered (Hilbe, 2011).

2.3.3 Negative Binomial (NB) model

The probability distribution of the Negative Binomial model (NB) is given by

$$f(y,\lambda,\nu) = \frac{\Gamma(y+\nu)}{y!\Gamma(\nu)}\left(\frac{\nu}{\nu+y}\right)^{\nu}\left(\frac{y}{y+\nu}\right)^{y} \tag{2.8}$$

The means are based on the logarithmic link, $\lambda = e^{x^T\beta}$. Unlike in the Poisson regression model, the NB mean (λ) is not equal to its variance and it is given as, $Var(y) = \lambda + \alpha\lambda^2$, where $\alpha = \frac{1}{\nu}$. As in the Poisson regression model, the parameters β and α of the NB regression model are estimated by taking the partial derivatives of the NB log likelihood with respect to β and α. The NB log likelihood is given by,

$$L(\beta; y, \alpha) = \sum_{i=1}^{n} yln\left(\frac{\alpha e^{x^T\beta}}{1+e^{x^T\beta}}\right) - \frac{1}{\alpha}ln\left(1+e^{x^T\beta}\right) + ln\left[\Gamma\left(y+\frac{1}{\alpha}\right)\right] - ln\left[\Gamma\left(y+\frac{1}{\alpha}\right)\right] - ln\left[\Gamma\left(\frac{1}{\alpha}\right)\right] \tag{2.9}$$

The normal equations are obtained from

$$\frac{\partial L(\beta, y, \alpha)}{\partial \beta} = 0 \tag{2.10}$$

The solutions for the partial derivatives are provided by Hilbe (2011). Fisher scoring can also be used to obtain the estimates. A regression equation of the model with a disturbance term that accounts for the overdispersion is given by

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \cdots + \beta_l x_{lij} \tag{2.11}$$

The NB model is a generalized Poisson model (Cameron and Trivedi, 1986). The NB regression model is known as a parametric model for overdispersion. The assumption for this model is that y has a Poisson distribution with the expected value λ_{ij} conditional on ϵ_{ij} and that $e^{\epsilon_{ij}}$ follows a standard Gamma distribution.

The NB model is like a Poisson model in the sense that it can only take non negative integers. However, it is dissimilar in that it has two parameters, the mean and the alpha (α) term. The alpha term is an indication of the spread of the data around the Poisson mean, when it has values greater than 1. However, if the alpha term is equal to zero then the NB model reverts to the Poisson model.

The NB regression model is not only used to model overdispersion, but has applications in other research areas such Accident Analysis. Miaou (1994) studied the performance of Poisson and NB regression models in investigating the relationship between truck accidents and the design of the road section where the accident occurred. In addition Abdel-Aty and Rwadan (2000) used the NB regression to model the frequency of accident occurrence. The model has applications in psychiatric research. Elhai, Calhoun and Ford (2008) used the NB regression to model the mental health visits count.

The NB regression model has its own drawbacks like any other modelling tool. Lord (2006) observed that the dispersion parameter of NB models can be biased when not enough sample are available for estimating the model. The Poisson and NB models cannot be used to datasets that contained a large number of zeros and heavy tail that forms highly dispersed data (Geedipally, Lord and Dhavala, 2012).

2.3.4 Gamma model

Another model that can be used to model migration is Gamma regression. This section starts by defining the probability function and the parameters of the model. The probability density function of the Gamma distribution is defined as,

$$f(x) = \frac{\alpha^k x^{k-1} e^{-\alpha x}}{\Gamma(k)},\ 0 \leq x \leq \infty,\ \alpha > 0,\ k > 0, \tag{2.12}$$

where the α and k parameters represent, the scale and shape parameter respectively. The Maximum Likelihood method can be used to estimate the parameters, α and k. Γ is known as a Gamma function, mathematically this function is defined by,

$$\Gamma(k) = \int_0^\infty \alpha^k x^{k-1} e^{-\alpha x}\, dx \tag{2.13}$$

The mean (μ) of the Gamma distribution is defined as,

$$\begin{aligned} \mu &= \int_0^\infty x f(x) dx \\ &= \int_0^\infty x \frac{\alpha^k x^{k-1} e^{-\alpha x}}{\Gamma(k)} dx \\ &= \frac{k}{\alpha} \end{aligned} \tag{2.14}$$

To calculate the variance (σ^2), first we need to derive the formula for the second moment μ_2. Burgin (1975) defined the rth moment μ_r of the distribution about the origin as,

$$\begin{aligned} \mu_r &= \int_0^\infty x^r f(x) dx \\ &= \frac{\Gamma(k+r)}{\alpha^r \Gamma(k)} \end{aligned} \tag{2.15}$$

$\Gamma(k)$ for integers is expressed as,

$$\Gamma(k) = 1 \cdot 2 \cdot 3 \cdot 4 \cdots (k-1) = (k-1)! \tag{2.16}$$

Therefore, the expression for the second moment μ_2, (substitute r =2) is,

$$\begin{aligned}\mu_2 &= \frac{\Gamma(k+2)}{\alpha^2\Gamma(k)} \\ &= \frac{(k+)!}{\alpha^2(k-1)!} \\ &= \frac{(k+1)(k)(k-1)(k-2)\cdots 2\cdot 1}{\alpha^2(k-1)(k-2)\cdots 2\cdot 1} \\ &= \frac{(k+1)(k)}{\alpha^2}\end{aligned} \tag{2.17}$$

The variance is equal to,

$$\begin{aligned}Var[x] &= E[x^2] - E[x]^2 \\ &= \frac{k^2+k}{\alpha^2} - \frac{k^2}{\alpha^2} \\ &= \frac{k}{\alpha^2}\end{aligned} \tag{2.18}$$

Similar to Poisson model, the log link function will be used and the regression equation of the Gamma model that is used in this study is expressed as,

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \cdots + \beta_l x_{lij} \tag{2.19}$$

The distribution is parsimonious in parameters and, hence, simple to use. Another advantage of the Gamma distribution, as noted by Singh, Singh, and Kumar (2011), is that the distribution has various shapes of hazard function[3] for different values of the shape parameter. The Gamma regression model is not popular in migration studies, but the model has been applied successfully in financial and

[3]The hazard function is a conditional density, given that the event in question has not yet occurred prior to time t.

the Operational Research area. Milevsky and Posner (1998) used the reciprocal Gamma distribution density to model the Asian pricing options. Burgin (1975) used the Gamma distribution to model inventory control. However, the model has it own disadvantages. Greene (1990) and Singh et al (2011) stated that its distribution function and survival function cannot be presented in a closed forms. Singh et al (2011) further noted that, if the shape parameter is an integer, the hazard function involves the incomplete Gamma function which is difficult to manipulate mathematically. Another draw back of the model, is that, zero values for a response variable is considered unrealistic by this model.

2.4 Local Modelling

This section discuss the Geographically Weighted Regression (GWR), Moran's I and the Monte Carlo significance test.

2.4.1 Geographically Weighted Regression (GWR) model

Geographically Weighted Regression (GWR) is an extension of the traditional regression model. Global models (all the models discussed above) assumed that the parameters estimated are stationary over a geographical space.

The GWR model allows local variations in these parameters to be estimated. The model also allows for an investigation of the way in which an explanatory variable influences changes over a geographical space, rather than simply assuming it has the same influence at all locations as in the global approach (Byrne and Pezic, 2004).

The GWR model is given by

$$y_i = \beta_0 (a_i,\ b_i) + \sum_j \beta_j\ (a_i, b_i) x_{ij} + \epsilon_i \tag{2.20}$$

where y_i represents the value of the internal migration flow, x_{ij} are the explanatory variables, (a_i, b_i) denotes the coordinates of the i-th point in the space and $\beta_j (a_i, b_i)$ is the local coefficient for the j explanatory variable at location *i*. The estimation of the parameters for the GWR model is similar to that of Weighted Least Squares (WLS). The exception is that the weights are conditioned on the location (a_i, b_i) relative to the other observations in the data set, and hence any change of location. The expression for the GWR model estimator takes the form

$$\hat{\beta} (a_i, b_i) = \left(X^T\ w (a_i, b_i)\ X\right)^{-1} X^T w (a_i, b_i)\, Y_i \tag{2.21}$$

where $w (a_i b_i)$ represent a square matrix of weights relative to the position of $(a_i b_i)$, $X^T w (a_i b_i)\, X$ is the geographically weighted variance-covariance and Y_i is the vector of the response variables. The off diagonal elements of the square matrix $w (a_i b_i)$ are equal to zero.

$$w (a_i b_i) = \begin{pmatrix} w_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_{in} \end{pmatrix} \tag{2.22}$$

where w_{in} represent the weighting at point n on the model calibration around point i. These weights vary with i which distinguishes the GWR model from traditional WLS model where the scaling matrix is constant.

A global model such as OLS regression is the same as a local model in which each data point has a unit weight so that there is no spatial variation in the estimated parameters (Brunsdon, Fotheringham and Charlton, 1996; and Fotheringham, Brunsdon and Charlton, 2000). This suggests that, when the diagonal elements of $w(a_i, b_i)$ are equal to one in equation (2.22) then equation (2.21) is equivalent to an OLS. The parameters of the GWR model change over space. In this case w_{ij} are defined as continuous functions of the d_{ij}'s, the distances between i and j, so that,

$$w_{ij} = e^{\frac{-{d_{ij}}^2}{h^2}} \tag{2.23}$$

where h is known as the bandwidth. If i and j coincide, the scaling of the observation will be unity. The weighting or scaling of other data will decrease according to a normal curve as the distance between i and j increases. An important step in fitting the GWR model, is the selection of h which controls the rate at which the scaling decays.

A large h in the GWR results tends to be informative, approaching the OLS results as h gets closer to infinity. If h is too small, the GWR parameter estimates will increase depending on data points in close proximity to i and hence will have increased variance. The Cross validation (CV) method assists in choosing the optimum bandwidth, and is defined as,

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i})^2 \tag{2.24}$$

where n is the number of observations and $\hat{y}_{\neq i}$ is the fitted value of y_i with the observations for data point *i* omitted from the calibration process. More discussion

on cross validation methodology can be sourced from Brunsdon et al (1996).

Alternatively, the bandwidth (h) may be chosen by minimising the Bias Corrected Akaike Information Criteria (AIC_c) score, defined as,

$$AIC_c = 2nln(\hat{\sigma}) + nln(2\pi) + n\left\{\frac{n + tr(M)}{n - 2 - tr(M)}\right\} \tag{2.25}$$

where n is the number of observations, $\hat{\sigma}$ is the estimated standard deviation of the error term, and $tr(M)$ is the trace of the hat matrix M, which is a square matrix.

The spatial weighting function can be implemented equally at each calibration point. The problem with that is, in some regions, the local regressions fits might be based on relatively few data points, if the data is sparse. To correct this, another scaling function is included into GWR, it is known as a spatially adaptive weighting function and is defined by

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2, if \;\; d_{ij} < h_i \\ 0, \, otherwise \end{cases} \tag{2.26}$$

The spatially adaptive weighting function excludes points outside radius d, but tappers the scaling of points inside the radius, such that the weighting w_{ij} is a continuous and once differentiable function for all points less than a distance d from the centre point.

The GWR model has applications in migration studies, for example Nakaya (2001) applied it to study migration flows in Japan during the latter half of 1980. The model is gaining popularity in other areas of study, these include the social sci-

ences and spatial economic analysis. Cahill and Mulligan (2007) used the GWR model to study local crime patterns in Portland and Oregon. Kalogirou and Hatzichristos (2007) presented a spatial modelling framework for income estimation in Athens by using the GWR model as a modelling technique for their study.

2.4.2 Moran's I

Moran's I is a global statistic because it estimates the overall degree of spatial autocorrelation for a data set. The possibility of spatial differences suggests that the level of autocorrelation may significantly vary across the geospace. Local spatial autocorrelation statistics give estimates that are expanded to the level of the spatial analysis units, allowing assessment of the relationship across space.

Moran's I behaves like a Pearson correlation coefficient. Its value is generally between -1 and 1, but can sometimes exceed -1 or 1 (Fortin and Legendre, 1989). Positive values indicate positive autocorrelation and vice versa, but if no spatial autocorrelation detected then this suggest that the spatial arrangement of the residuals is random. Moran's I is calculated as follows:

$$I = \frac{n \sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i \sum_j w_{ij} \sum_i (z_i - \bar{z})^2} \tag{2.27}$$

where I is the Moran coefficient for the distance class d, and n is the number of spatial units that are labelled by i and j, z is the variable of interest, $\bar{z}$ is the average of z, i and j vary from 1 to n, w'_{ij}s take the value 1 when the pair of location (i, j) pertains to distance d and 0. W is the sum of the w'_{ij}s. The Moran's I statistic can be interpreted by the evaluation of the standard normal deviate that

is computed as

$$Z = \frac{[I - E(I)]}{\sigma(I)} \tag{2.28}$$

where $E(I)$ is the expected I, and is defined as,

$$E(I) = \frac{-1}{(n-1)} \tag{2.29}$$

and $\sigma(I)$ is the standard deviation of I, and is defined by,

$$\sigma(I) = \sqrt{\frac{nv_4 - v_3 v_5}{(n-1)(n-2)(n-3)(\sum_i \sum_j w_{ij})^2}} \tag{2.30}$$

Where, v_1, v_2, v_3, v_4 and v_5, are defined as follows,

$$v_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \tag{2.31}$$

$$v_2 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji} \right)^2 \tag{2.32}$$

$$v_3 = \frac{n^{-1} \sum_i (z_i - \bar{z})^4}{(n^{-1} \sum_i (z_i - \bar{z})^2)^2} \tag{2.33}$$

$$v_4 = (n^2 - 3n + 3)v_1 - nv_2 + 3 \left(\sum_i \sum_j w_{ij} \right)^2 \tag{2.34}$$

$$v_5 = v_1 - 2nv_1 + 6 \left(\sum_i \sum_j w_{ij} \right)^2 \tag{2.35}$$

For the hypothesis testing, Moran's I values are transformed to Z-scores, and

values that are greater than $Z_{0.025}$ (1.96) or smaller than $-Z_{0.025}$ (-1.96) specify the presence of spatial autocorrelation at the 5% level of significance.

2.4.3 Monte Carlo significance Test

This section presents a method that is used to assess the spatial variability of the beta (β) parameters of the GWR. The method is in the form of hypothesis testing as discussed by Brunsdon, Fotheringham and Charlton (1998). The hypothesis test is formulated as follows,

$$H_0 : \beta_{ij} = \beta_j \ \forall i, \ vs \ H_a : \beta_{ij} \text{ not all the same } \forall i \tag{2.36}$$

The statistic that is used to calculate the variability of β_{ij} as i varies for a fixed j

$$v_j = \sum_i \frac{(\beta_{ij} - \beta_{.j})^2}{N} \tag{2.37}$$

where $\beta_{.j}$ denotes averaging over j. Brunsdon et al (1998) noted that the smaller the v_j the greater the evidence that the coefficient matching v_j is fixed. Individual variables, presented in the hypothesis in equation (2.36) can be investigated if the null hypothesis of v_j were known. From the GWR modelling frame work this is not so, however the Monte Carlo method offers an alternative approach. A randomisation test needs to be carried out, under the null hypothesis which assumes that the beta parameters (β_{ij}) do not vary with i for variable j. Brunsdon et al (1998, page 436) state that "if the GWR model were to be calibrated with the locations of the data points randomly assigned to the predictor and response variables, then there should be little difference in the patterns of beta parameters,

if the beta parameters are fixed over space, then, the spatial location should not greatly affect their calibration. Using Monte Carlo tests it should be possible to compare the distribution of the v_j under the randomisation hypothesis". The procedure, for a given j, is as follows.

1. Make a record of v_j for the accurately located observed data points.
2. Randomise the locations of the observations.
3. Redo step 2, P-1 times, recording v_j at each step.
4. Measure the rank of v_j for the correctly specified case, R.
5. The $p - value$ for the randomisation hypothesis is R/P.

2.5 Modelling Diagnostics

This section concentrates on the diagnostic measures that will be used to assess the assumptions, performance and the strength of the individual models.

2.5.1 Variance Inflation (VIF)

The variance inflation (VIF) is a diagnostic measure that is used in a linear regression to check for multicollinearity among explanatory variables. Multicollinearity suggests that there is a near-linear dependence among the independent variables. The explanatory variables are the columns of the X matrix. A linear dependence would result in a singular X^TX, or alternatively, the inverse of the matrix X^TX will not exist. Montgomery, Peck and Vining (2006) further explained that, the presence of multicollinearity can dramatically impact the ability to estimate re-

gression coefficients. The VIF statistics that will be used to investigate the presence of multicollinearity, is defined as follows.

$$VIF_l = \frac{1}{1 - R_l^2} \tag{2.38}$$

Where R_l^2 is the coefficient of determination obtained from the explanatory regression of independent variable l on the remaining independent variables. In this study, an explanatory variable with a VIF greater than 6, indicates a multicollinearity and meaning such variables are dropped. Alternatively, the variance inflation factor can be used as a measure of variable redundancy, and the value of the VIF can help to decide which variables need to be removed from the model (Rosenshein, Scott, and Pratt, 2011).

2.5.2 Shapiro-Wilk test

Shapiro and Wilk (1965) developed a test known as the Shapiro-Wilk test which is used to check the assumptions of the linear regression model. This statistical test is used to investigate if a sample came from a normally distributed population. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.39}$$

where, $y_{(i)}$ is the ith order statistic, the $ith-$ smallest number in the sample;

$$\bar{y} = (y_1 + \cdots + y_n)/n \tag{2.40}$$

is the sample mean; the constants a_i are given by

$$(a_i, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1})^{\frac{1}{2}}} \tag{2.41}$$

where

$$m = (m_1, \cdots, m_n)^T \tag{2.42}$$

and m_1,$\cdots$,m_n are the mean values of the order statistics of the response and identically distributed random variables which are sampled from the standard Gaussian distribution, and V is the covariance matrix of those order statistics. The default hypothesis may be rejected if W is below a predetermined threshold. The default hypothesis in this case is that the population is Gaussian. If the probability of obtaining a value as extreme as W is smaller than a specified α level, the default hypothesis is rejected. This means that there is evidence that the data tested are not from a Gaussian population.

2.5.3 Jarque-Bera-Test (JB-Test)

As in the Shapiro-Wilk test, the Jarque-Bera Test (JB-Test) is used to test the hypothesis that the distribution of the residuals are normally distributed (Thadewald and Buning, 2004). The test depends on the measures of skewness (S) and kurtosis (K). The JB test is defined by

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{2.43}$$

where the sample skewness

$$S = \frac{\hat{\mu_3}}{(\hat{\mu_2})^{\frac{3}{2}}} \tag{2.44}$$

is an estimator of

$$b_1 = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} \tag{2.45}$$

and the sample kurtosis

$$K = \frac{\hat{\mu_4}}{(\mu_2)^2} \tag{2.46}$$

an estimator of

$$k_2 = \frac{\mu_4}{(\mu_2)^2} \tag{2.47}$$

μ_2 and μ_3 are the theoretical second and third central moments, respectively, with its estimates

$$\hat{\mu_j} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^j,\ j = 2.3, 4... \tag{2.48}$$

Bowman and Shenton (1975) stated that the JB test is asymptotically chi-squared distributed with two degrees of freedom because it is just the sum of squares of two asymptotically independent standard normal distribution. This means H_0 has to be rejected at level α if $JB \geq \chi^2_\alpha(2)$.

2.5.4 Durbin-Watson (D-W)

The Durbin-Watson (D-W) is a test statistic that is used to check if there is autocorrelation in the residuals of a regression model. It is important to test for autocorrelation, because the presence of autocorrelation leads to wrong standard errors for regression coefficients.

The D-W test is based on the assumption that errors in a regression model are generated by first-order autoregressive process observed at equally spaced time periods, defined by,

$$\epsilon_t = \rho\epsilon_{t-1} + a_t \tag{2.49}$$

where ϵ_t is the error term in the model at time period t, a_t is a normally distributed independent random variable with zero mean and constant variance, and ρ is a parameter that defines the relationship between the model errors ϵ_t and ϵ_{t-1}. In the case of the OLS model, testing the presence of autocorrelation is important, because if it is ignored the model will predict estimates with biased standard errors.

The D-W statistic is defined as,

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{2.50}$$

where e_t, $t = 1, 2, \cdots, n$ are the residuals from an OLS regression of y_t on x_t. For uncorrelated errors the value of the D-W statistic should be close 2. We present the set of rules that need to be followed, when deciding whether the assumption of uncorrelated errors is violated or not as suggested by Montgomery et al (2006).

The formal test of positive first order serial correlation is as follows:

$$H_0 : \rho = 0, \text{ there is no autocorrelation}$$
$$H_a : \rho > 0, \text{ there is positive autocorrelation}$$

If the test statistic $d < d_{L_\alpha}$ reject H_0, while if $d > d_{U_\alpha}$ do not reject H_0, but if

$d_{L_\alpha} < d < d_{U_\alpha}$ the test is inconclusive.

Also the statistic d can be used to test the presence of the negative autocorrelation as well. To test for the significance of a negative autocorrelation, the test statistic (4-d) is compared to both the d_{L_α} (lower) and d_{U_α} (upper) critical values. The hypothesis is presented as,

$$H_0 : \rho = 0, \text{ there is no autocorrelation}$$
$$H_a : \rho < 0, \text{ there is a negative autocorrelation}$$

If (4-d) $< d_{L_\alpha}$, there is a sufficient evidence to support H_a. However, if $(4 - d)$ $< d_{U_\alpha}$, there is evidence to suggest that there is no negative autocorrelation, and if 4-$d_{L_\alpha} < d < 4 - d_{U_\alpha}$, then, the test is inconclusive

2.5.5 Koenker-Breusch Pagan (Koenker-BP) Test

The Koenker (1981) and Breusch Pagan (BP) (1979) derived the same test statistic that is used to test for Heteroscedasticity in the linear regression model. Unlike the BP test, Koenker (1981) relaxes the assumption that the error terms are normally distributed and the test is less sensitive to non-normality of data and small sample sizes. Also, if the Koenker-BP $p - value$ is small and statistically significant, that simply suggests that the relationship varies across the study area and the parameter estimates are nonstationary (Rosenshein et al, 2011).

The Koenker-BP statistic also known as Koenker's Studentised Breusch Pagan statistic is distributed as,

$$BP^{Koenker} \sim \chi^2_m \tag{2.51}$$

i.e., $BP^{Koenker}$ has a Chi-square distribution with m degrees of freedom.

2.5.6 Test for significance of linear regression model

The test for significance of regression is a procedure to check if there is a linear relationship between the response or dependent variable (y) and any of the explanatory variables. This procedure assesses the overall fit of the linear model. The appropriate hypothesis is,

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

(None of the variables are useful in predicting the response)

$$H_a\text{: } \beta_j \neq 0 \quad \text{for at least one } j$$

The rejection of the null hypothesis (H_0) suggests that at least one of the regressors, $x_1, x_2, \cdots, x_k$ contributes significantly to the model. The total sum of squares SS_T is partitioned into a sum of squares due to regression, SS_R, and a residual sum of squares, SS_{Res}. Then, the total sum of squares SS_T is defined as,

$$SS_T = SS_R + SS_{Res} \tag{2.52}$$

$$SS_R = \hat{\beta}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{2.53}$$

$$SS_{Res} = y^T y - \hat{\beta}^T X^T y \tag{2.54}$$

$$SS_T = y^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{2.55}$$

If the null hypothesis is true, then $\frac{SS_R}{\sigma^2}$ follows a χ_k^2 distribution with the same number of degrees of freedom as number of regressor variables in the model, and $\frac{SS_{Res}}{\sigma^2}$ follows χ_{n-k-1}^2, where SS_{Res} and SS_R are independent. The statistic that

is used to test the overall significance of the linear model is defined as,

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} \sim F_{k,n-k-1} \tag{2.56}$$

Table 2.1: Analysis of Variance (ANOVA) for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of freedon	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Where n is the number of the observations, y_i is the observed values of the response variable, $\bar{y}$ is the, mean of y, and $\hat{y}_i$ is the estimated value of y.

2.5.7 Reset Test

In testing the structure of the model we present the artificial model which was defined as,

$$ln(y_{ij}) \quad = \quad \beta_0 + \beta_1 ln(x_1) + \beta_2 ln(x_2) + \cdots + \delta_1 ln(x_n)^2 + \eta_{ij} \tag{2.57}$$

Where η_{ij} is the random error, and the parameter δ_1 is estimated by least squares. However, the aim is to investigate the significance of the artificial model by testing a hypothesis,

$$H_0 : \delta_1 = 0$$
$$H_a : \delta_1 \neq 0$$

If the $p - value$ from the output of the Reset test, is less than 5% (0.05), we then reject H_0, and conclude that the test has detected misspecification on the model.

2.5.8 Ord plot

Ord (1967) developed a diagnostic plot that can be used to assist in identifying a discrete model for a count response variable. The author defined a class of discrete distributions by the difference equation,

$$\triangle f_{l-1} = f_l - f_{l-1} = \frac{(a-l)f_{l-1}}{l(b_1 + b_2 l)} \tag{2.58}$$

Where $l \geq 1$ and a, b_1, b_2 are parameters. When b_2 =0, we have,

$$u_l = \frac{l f_l}{f_{l-1}} = \frac{[a + l(b_1 - 1)]}{b_1} = c_0 + c_1 l \tag{2.59}$$

Ord (1967) showed that a linear relationship of the form above holds for these distributions; Poisson, NB and logarithmic series. The slope c_1 is zero for the Poisson, negative for the binomial, positive for the NB and logarithmic series distributions. The intercept c_0 is positive for the following distributions, Poisson, binomial and NB, and negative for the logarithmic series.

2.5.9 Deviance of the generalised linear models

An alternative global test to F statistics is the deviance, this statistic is more relevant to count models, such as, Poisson, NB and Gamma model. This section defines the deviance statistic that is used to test the overall fit of generalised linear

models. The deviance is defined as,

$$D = 2\phi(L(y : y) - L(\hat{\mu} : y)) \tag{2.60}$$

Where $L(\hat{\mu} : y)$ is the log-likelihood function expressed as the function of the predicted mean values $\hat{\mu}$ given the response variables, and $L(y : y)$ is the log-likelihood function computed by replacing $\hat{\mu}$ with y.

The deviance for Poisson, NB and Gamma model, are defined as,

$$D_{Poisson} = 2\sum\left(y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right) \tag{2.61}$$

$$D_{NB} = 2\sum\left(y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i + \alpha^{-1}) ln\left(\frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}}\right)\right) \tag{2.62}$$

$$D_{Gamma} = 2\sum\left(-ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) \tag{2.63}$$

Smaller values of the deviance indicates that the model fits the data better.

2.5.10 Model selection

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. The AIC is used in the identification of the best model in a list of the competing models. However, the AIC does not provide any information about the quality of the models, for instance, if the list of the competing models fits a particular data set poorly, then the AIC fails to give any warning about that.

The AIC is based on the maximum likelihood function. According to Mutua

(1994) the AIC produces fairly good results for $n \geq 30$

AIC value is defined as follows,

$$AIC = 2g - 2ln(L) \tag{2.64}$$

where g is the number of parameters in the model, and L is the maximised value of the log likelihood function for the model. In competing models, the model with the smaller AIC value is preferable. Hence, the AIC incorporates a penalty that increases with the number of estimated parameters.

The Corrected Akaike Information Criterion (AIC_c) is the adjustment of the AIC for known sample sizes:

$$AIC_c = AIC + \frac{2g(g+1)}{n-g-1} \tag{2.65}$$

Where n is the sample size. Thus, AIC_c is an AIC with a penalty term. When n is small or g is large. Hurvich and Tsai (1989) strongly recommend using AIC_c, rather than using AIC. AIC_c tends to AIC for large n. Hurvich and Tsai (1989) further state that for linear regression, AIC_c is unbiased, but this fact is based on the assumption that the candidate family of competing models includes a true model.

If the AIC_c values for two models differ by more than 3, this suggest that the model with the lower AIC_c is the better one.

2.6 Cluster Analysis (CA)

Cluster Analysis (CA) is a procedure that is used to divide the data into homogeneous groups. The homogeneous groups are known as clusters. The observations in each cluster are similar and distinct to those in other clusters. Similarity is measured by estimating the distance (Euclidean, Average Linkage method, etc.) between the pairs of clusters. However, clusters with small distances between each other are similar and clusters with larger distances are distinct. Segmentation is a popular application of the cluster analysis for example, Lee, Lee, Bernhard and Yoon (2006) used CA to segment the casino gambling market in Korea.

The important step in cluster analysis is to identify variables that are needed for clustering. A challenging issue is to determine the number of clusters that are needed from the data set. Matignon (2007) suggested that this can be achieved by constructing scatter plots and analysing various clustering statistics to determine the number of clusters.

After identifying certain variables for clustering, the next step is to select the clustering method that will divide the data into homogeneous groups. The most popular clustering methods are hierarchical methods and partitioning methods. In this study the k-means clustering method is used.

2.6.1 Cluster Estimation: Silhouette Width

The mean of each observation's Silhouette value is called a Silhouette Width. The silhouette width can be used to determine how many clusters a data set has. The method can also be used as a diagnostic measure. The silhouette value measures

the extent of one in the cluster allocation of a specific observation. The Silhouette values near 1 indicates that the observations are well clustered, while value near -1 suggest that the observations are poorly clustered. For data point i, the silhouette is defined as

$$S(i) = \frac{b_i - a_i}{max(b_i,\, a_i)} \tag{2.66}$$

where a_i is the mean of all observations in the same cluster as data point i, and b_i is the mean of all observation in the nearest neighbouring cluster data point i.

$$b_i = \min_{Ck \in \mathcal{C}\ C(i)} \sum_{j \in Ck} \frac{dist(i,j)}{n(Ck)} \tag{2.67}$$

where $C(i)$ is the cluster containing observations i, $dist(i,j)$ is the distance (e.g Euclidean, Manhattan) between observations i and j, and $n(C)$ is the cardinality of cluster C. The Silhouette Width is between -1 and 1 and should be high (Rousseeuw, 2008).

2.6.2 The K-means clustering

The $K-means$ is a clustering method that falls under the category of an unsupervised learning algorithm. This clustering technique attempts to identify relatively K- different groups based on deliberately selected variables such as demographic and non demographic variables $K - means$ is applied when one is aware of the number of clusters that need to be formed in the data.

The following steps are used in the $K - means$ algorithm,

1. Enter the entries that need to be clustered and input K for total number of clusters to be formed

2. Randomly select K entries to be the original cluster centers.

3. Allocate each entry to the cluster with the closest mean.

4. Determine the new average of each cluster.

5. Redo step 3

6. Stop when the convergence condition is reached.

Convergence condition suggest that no object form a new cluster. Another condition, that is frequently used, is to minimise the squared error E of all the objects in the data set:

$$E = \sum_{i=1}^{k} \sum_{\varepsilon \in C_i} |o - \mu_i|^2 \tag{2.68}$$

Where o is the entry that belongs to cluster C_i, μ_i is the average of the cluster C_i, and K is the number of clusters (Mitsa, 2010).

2.7 Summary

The gravity model is used by many authors in migration studies and the model has a linear structure. The parameters of this model are derived using OLS. When using the Gravity model, certain conditions need to be verified, such as normality, constant variation of the error terms and independence. This study uses using Shapiro-Wilk and JB-Test to assess the normality assumption.

In this study positive numbers are modelled and it was discovered that the assumption of log normality is not acceptable for a count response variable. Another challenge arises from the use of the log transformation when some of the flows are zero. The log of zero cannot be calculated and in fitting the Gravity model. It is

advisable that a count variable should not be analysed by a simple log-transform, instead models such as Poisson and Negative Binomial models should be used.

The Poisson model is a better alternative model for count response variables. This model has one parameter (mean) and this model assumes that the mean is equal to its variance. However, when the assumptions of the Poisson model are violated, this will lead to biased the standard errors.

When overdispersion occurs the NB in most cases is used to model count variables alternatively, the Gamma model is used. Both NB and Gamma model are known as generalised Poisson models. The NB model is only used to account for overdispersion. This study will consider the use of diagnostic plots such as, the Ord plot, to investigate if either the Poisson or NB model are suitable for the modelling the count response variable.

The NB, Gamma, Poisson and the gravity model assume that the parameters estimated are stationary over a geographical space. The difference between these models and GWR, is that the GWR model allows local variations in these parameters to be estimated.

In this study cluster analysis is applied with the silhouette width used to estimate the number of clusters from the data set. The K-means method is used to study the profile of the data.

Chapter 3

Methodology

3.1 Introduction

This chapter discusses the data set that will be used in the study, including when the data was collected (time frames) and the methods that were used to collect the data. The methods used to described and model the data are also described in detail.

3.2 Data

The study was based on the 2011 census data, collected by Stats SA. The 2011 census counted all the people in the country and collected information about their ages, education levels, employment, housing conditions and migration status among others.

South Africa conducted a defacto population census, in which individuals were

counted at the place where they spent the census night, (on the 9/10 October 2011). The data collection was undertaken by field staff in excess of 160000 over 3 weeks, using face-to-face interviews. The main objective of the census was to count everyone, but during the counting process some people were missed (undercount). The undercount for the census 2011 was 14.6 percent for persons and 14.3 percent for households (Statistics South Africa, 2012(b)).

After the data collection phase, the next phase was data processing. The objective of data processing was to accurately process census questionnaires (15,821,302 questionnaires) in order to establish a clean, accurate, consistent and reliable data set. Data processing involved the following stages: process of storage of boxes, data capturing, editing, tabulation and analysis. The information received from questionnaires collected during data collection was converted into data represented by numbers or characters.

Census data are characterised by numerous errors ranging from content to data processing errors. In order to detect and minimise some of the errors, the automated error detection and correction method was used based on a predefined set of editing rules (specifications). The aim of editing the data was to make the processed data complete and internally consistent, while keeping the number of changes to a minimum. For the 2011 Census, the editing system used a combination of both logical imputation approaches and hot decks imputation when inconsistencies were found in the census data (Statistics South Africa, 2012(a)).

The variables used in this study are categorised as follows, demography and economic variables, see Section 1.3. From the 2011 census, the migration questions were available at local municipality level, and in this study we derive the migra-

tion totals for district municipalities from 234 local municipalities in South Africa. Statistics South Africa (2012(c)) was used as a data source to extract the GDP values. Google Maps (2014) was used to capture the distance between district municipalities.

The objective of this study was to investigate internal migration at the district level. This implies 52*(52-1) = 2652 different flows or observations from the migration matrix.

3.3 Description of the data

In this study, the data will be described by using the following description, bar graphs, tables and maps.

3.4 Modelling

One of the key objectives of this study was to develop a model that explains migration in the district municipalities of South Africa. According to Fan (2005) the Gravity model does not take into consideration the effect of uneven regional or district disparities such as economic development. In order to correct this problem, as already suggested by the author, additional explanatory variables will be included into equation (2.1).

The data set consists of 2652 migration flows meaning that the number of observations are more than the number of parameters. All zero observation in this data will be increased by 0.1 but only for the Gravity model, to avoid the $\log(0)$ which does not exist, and the R code is given in Appendix D. Therefore, the Gravity

model that will be used to estimate parameters for the variables in Section 1.3, and the model is expressed as follows.

$$\begin{aligned} ln(M_{ij}) = \beta_0 + \beta_1 ln(pop_i) + \beta_2 ln(pop_j) + \beta_3 ln(pod_i) + \cdots \\ + \beta_4 ln(pod_j) + \cdots + \beta_{40} ln(illitrt_j) + \epsilon_{ij} \end{aligned} \tag{3.1}$$

The explanatory variables are from the district of origin and destination, as in the gravity model the beta parameters β's will be estimated by OLS. In section 2.3.1 the assumption of the OLS are discussed, in order to verify whether those assumption are met, the following will be done, (i) investigate the distribution of the error terms of this model by running diagnostics plots such as the box plot of the error terms in R, alternatively, (ii) check whether the residuals are normally distributed or not, by performing the Shapiro Wilk Test. The test will be done in R using the package $lmtest$ which was developed by Achim and Torsten (2002).

Section 2.3.1 captured the disadvantage of the gravity model in modelling a count response variable. Section 2.3.2 highlight various reasons why, the Poisson model is most suitable for modelling a count variable. Using equation (2.7) the regression model for the Poisson model is expressed as

$$\begin{aligned} ln(M_{ij}) = \beta_0 + \beta_1 (pop_i) + \beta_2 (pop_j) + \beta_3 (pod_i) + \cdots \\ + \beta_4 (pod_j) + \cdots + \beta_{40} (illitrt_j) \end{aligned} \tag{3.2}$$

From Section 2.3, $E(M_{ij})$ was defined as the expected value of the migration flow between districts. The Poisson model assumes that its variance is equal to the mean. This assumption will be verified by the $dispersiontest()$ function

in R from AER package, the package was developed by Christian and Achim (2008). The function tests the hypothesis that the response variable is overdispersed against the hypothesis that the response variable is not overdispersed.

If it is discovered that the response variable is overdispersed, then as stated in Section 2.3.3 the NB and Gamma model will be used. If the response variable (M_{ij}) is overdispersed. A NB regression model in equation (2.11) that accounts for overdispersion only is used.

$$\begin{aligned} ln(M_{ij}) = \beta_0 + \beta_1(pop_i) + \beta_2(pop_j) + \beta_3(pod_i) + \cdots \\ + \beta_4(pod_j) + \cdots + \beta_{40}(illitrt_j) \end{aligned} \tag{3.3}$$

The GWR will be used to model net-internal migration in the district municipalities of South Africa, but before using the GWR model, the OLS model is fitted first.

$$M_{ij} = \beta_0 + \beta_1 pod_k + \beta_2 tre_k + \beta_3 pow_k + \cdots + \epsilon_{ij} \tag{3.4}$$

After estimating the OLS model,we then apply the Koenker-BP Test to the residuals, if the test is statistically significant, this means that the relationships from the beta parameters(β's) vary across space .

The GWR model, from equation (2.20) is expressed as follows

$$M_{ij} = \beta_0(a_i, b_i) + \beta_1 pod_k(a_i, b_i) + \beta_2 tre_k(a_i, b_i) + \beta_3 pow_k(a_i, b_i) + \cdots + \epsilon_{ij} \tag{3.5}$$

To investigate the significance of the spatial variability of the beta coefficients from GWR model, the study used the $montecarlo.gwr()$ function in R from the

package GWmodel. The package was developed by Lu, Harris, Charlton, Brunsdon, Nakaya, and Gollini (2014). Before applying the test in this study, the data set will be converted to a spatial data frame using the function $SpatialPointsDataFrame()$ from the sp package (Pebesma, Bivand, 2005; and Roger, Edzer and Virgilio, 2013).

3.5 Model Estimation

A general description of migration will be given from the tables. Data analysis for the study will be carried out in R and ArcGIS. In this study the Poisson and Gamma regression are fitted in R by using the $glm()$ function while $glm.nb()$ is used to fit NB. These functions are from the $MASS$ package that was developed by Venables and Ripley (2002). The gravity model is estimated in R, by using the lm function (from the $lmtest$ package).

The GWR will be estimated in using ArcGIS software that was developed by Esri (2013). The study used the pam function (sillhouette method) from the cluster package to estimate the number of clusters from a data set. The package was developed by Maechler, Rousseeuw, Struyf, Hubert and Hornik (2014). The $k-means$ algorithm is performed in R. This is done using the clustering function $kmeans()$. The function is within the $stats$ package that was developed by R Core Team (2014).

3.6 Model Performance

The global validation of linear model assumptions function abbreviated as $gvlma()$ will be used to asses the assumptions of the linear model, this function was devel-

oped by Pena and State (2006). To examine the significance of the linear model we will be using the F-statistic (F-Test: is used to test the overall significance of the linear model). The z statistic or t-test will be used to test the significance of each beta parameter in the model. The variance inflation factor (VIF) is used to investigate multicollinearity between the predictors or independent variables. For GWR, if the condition number is 30 or above, suggest there is multicollinearity. Furthermore, the study investigates the presence of the outliers. The R function $outlierTest()$ from the $outliers$ package by Lukasz (2011) will be used.

For the Poisson, NB and Gamma model, a chi-square test or a deviance is used to test the significance of the models. In order to check the notion of underdispersion or overdispersion in the Poisson model, the Pearson Goodness-of-fit test is performed. Underdispersion or overdispersion for the NB and Gamma model are assessed using the dispersion parameter.

The models are compared by examining the Akaike Information Criterion (AIC) and Corrected Akaike Information Criterion (AIC_c). Furthermore the $AIC()$ from the $stats$ package and $ACc()$ function from $AICcmodavg$ package by Marc (2014), and these functions are used to calculate the AIC and AIC_c of the models.

3.7 Summary

This chapter explained the data used for this study, then the methodology that was used to conduct the 2011 census, etc. The actual models, such as, Gravity, Poisson, Negative Binomial (NB), Gamma and Geographically Weighted Regression (GWR) were also discussed in chapter 2, and the aim of this chapter is to demon-

strate the application of the models using the 2011 census data. The methods used to assess the models are described.

Chapter 4

Analysis

4.1 Introduction

This chapter presents the results of the analysis of the methods that were discussed in chapter 1 and chapter 2, i.e, Gravity, Poisson, NB, and GWR models. Furthermore, in this chapter the codes for the district municipalities are used instead of the full names for easier presentation in the graphs (for example DC10 is the Cacadu district municipality). The full names are defined in the Appendix B.

4.2 Description of the data

In this section the data is described by using bar graphs and maps. The distribution of the migration numbers was left skewed in the district municipalities, as observed in Figure 4.1. The minimum and maximum value of net-internal migration variable is -104906 and 232837. The mean of this variable is zero and it is larger than the median (-422.5). This indicates this variable is positively skewed,

from Appendix C, in Table C.1 the skewness is 1.38, and the kurtosis is 4.14 and this value means that the distribution of the net-internal migration is nonnormal (The kurtosis is 3 for a normal distribution). The CPI (Consumer Price Index) has a smallest standard deviation, and its value is 0.64 and this suggest that the observations of this variable are less spread out.

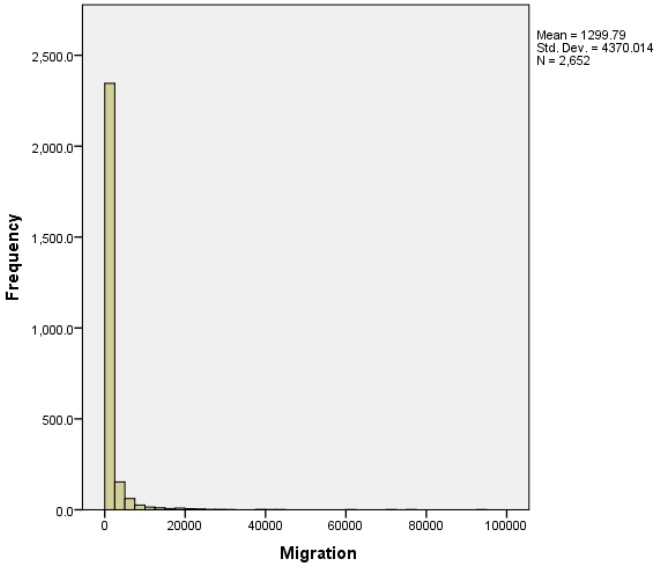

Figure 4.1: Migration distribution in the district municipalities

4.2.1 In-migration, Out-migration and Net-Internal Migration in KZN district municipalities

Figure 4.2 shows that the following district municipalities (DC) in KZN, DC25 (Amajuba), DC28 (Uthungulu), DC29 (iLembe), and DC22 (Umgungundlovu) have positive net-internal migration, which is explained by the high number of in-migration in these district municipalities. DC22 showing highest positive net-internal migration (more people are moving into this district municipality than

those going out) while DC27 (Umkhanyakude) has the highest negative net-internal migration in the KZN Province. The ETH (Ethekwini) metropolitan (metro) has a lot of movement in and out of the province, however, it has negative net-internal migration, this means more people are moving out of this metro.

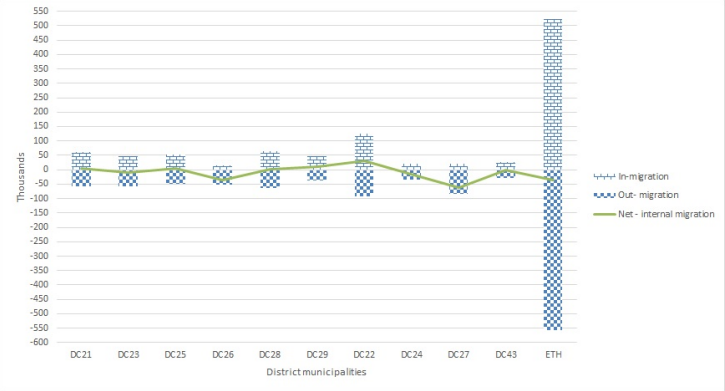

Figure 4.2: In-migration, Out-migration and Net-Internal Migration in KZN district municipalities

4.2.2 In-migration, Out-migration and Net-Internal Migration in EC district municipalities

Figure 4.3 indicates that the majority of the district municipalities, including two metros, BUF (Buffalo City) and NMA (Nelson Mandela Metropolitan) have negative net-internal migration, this indicates that more people are migrating out of the Eastern Cape, and DC 10 (Cacadu) is the only district with the positive net-internal migration.

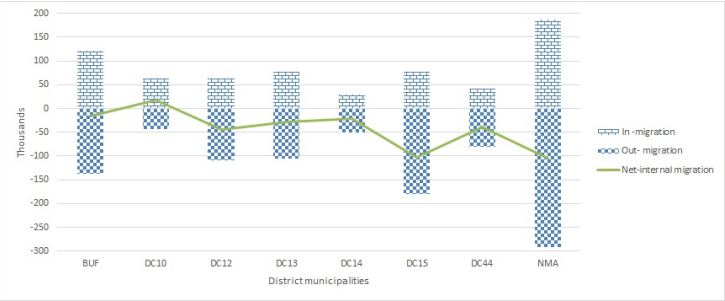

Figure 4.3: In-migration, Out-migration and Net-Internal Migration in EC district municipalities

4.2.3 In-migration, Out-migration and Net-Internal Migration in WC district municipalities

Figure 4.4 indicates that all the district municipalities, including the metro (CPT) having the higher positive net-internal migration, this suggest that the Western Cape province is a popular destination for internal migrants in South Africa.

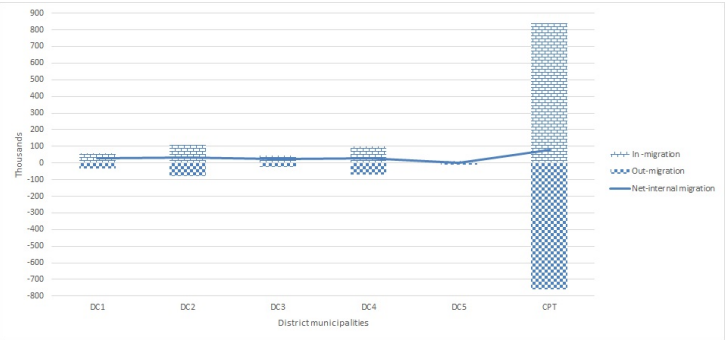

Figure 4.4: In-migration, Out-migration and Net-Internal Migration in WC district municipalities

4.2.4 In-migration, Out-migration and Net-Internal Migration in FS district municipalities

Figure 4.5 indicates that, DC20 (Fezile Dabi), and MAN (Mangaung Metropolitan) have the positive net-internal migration, and MAN has the highest net-internal migration in the Free State, DC18 (Lejweleputswa) has the highest out-migration, DC18 seems to repeling citizens.

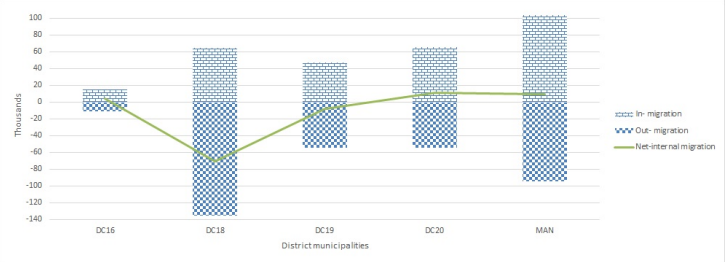

Figure 4.5: In-migration, Out-migration and Net-Internal Migration in FS district municipalities

4.2.5 In-migration, Out-migration and Net-Internal Migration in GP district municipalities

Figure 4.6 suggest that, DC42 (Sedibeng) is the only district municipality with negative net-internal migration in Gauteng. While the TSH (Tshwane Metropolitan Municipality) has the largest net-internal migration in South Africa, this makes TSH a popular destination to internal migrants. Johannesburg (JHB) almost has as many people moving in, as are moving out. While EKU (Ekurhuleni) has a significant number of internal migrants moving in than are moving out.

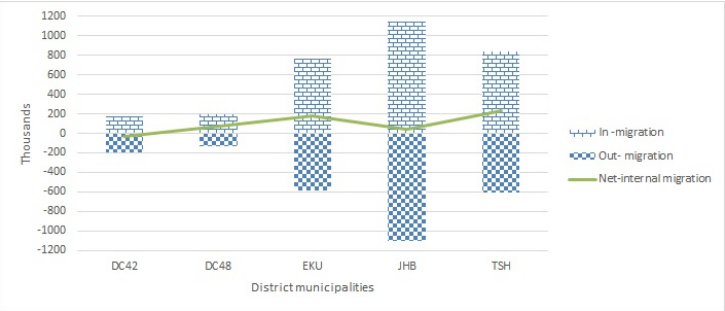

Figure 4.6: In-migration, Out-migration and Net-Internal Migration in GP district municipalities

4.2.6 In-migration, Out-migration and Net-Internal Migration in MP district municipalities

Figure 4.7 shows that, DC32 (Ehlanzeni) is the only DC in Mpumalanga province with a negative net-internal migration. While DC30 (Gert Sibande) and DC31 (Nkangala) are attracting more in- migrants than those leaving.

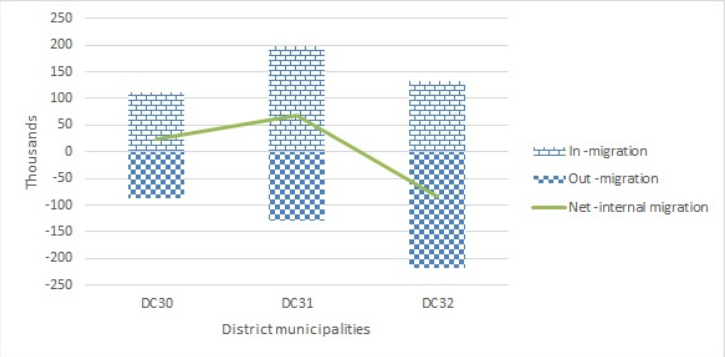

Figure 4.7: In-migration, Out-migration and Net-Internal Migration in MP district municipalities

4.2.7 In-migration, Out-migration and Net-Internal Migration in LP district municipalities

Figure 4.8 reveals that, DC36 (Waterberg) is the only DC in Limpopo province with a positive net-internal migration. While DC35 (Capricorn) has the highest negative net-internal migration.

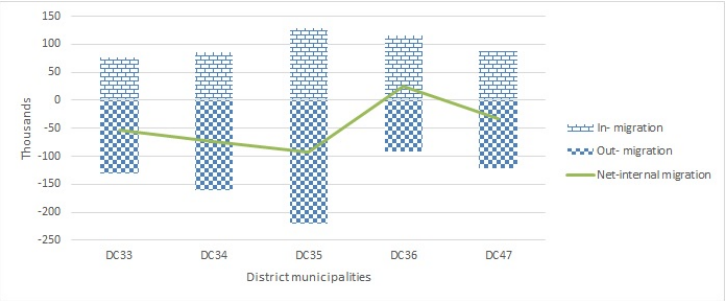

Figure 4.8: In-migration, Out-migration and Net-Internal Migration in LP district municipalities

4.2.8 In-migration, Out-migration and Net-Internal Migration in NW district municipalities

Figure 4.9 shows that only two DC's in North West province with the negative net-internal migration, namely DC38 (Ngaka Modiri Molema) and DC39 (Dr Ruth Segomotsi Mompati). DC37 (Bojanala) and DC40 (Dr Kenneth Kaunda) have the positive net-internal migration, people are migrating into these districts in North West.

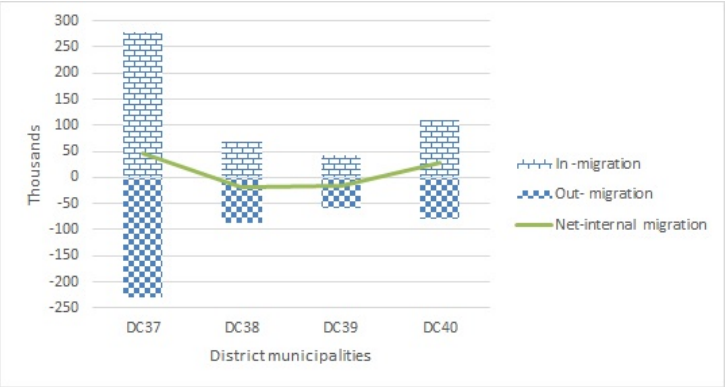

Figure 4.9: In-migration, Out-migration and Net-Internal Migration in NW district municipalities

4.2.9 In-migration, Out-migration and Net-Internal Migration in NC district municipalities

Figure 4.10 Shows DC8 (Siyanda) being the only district municipality in Northern Cape province with the positive net-internal migration, and the rest have the negative net-internal migration, this implies that, the out migration is higher in those district municipalities.

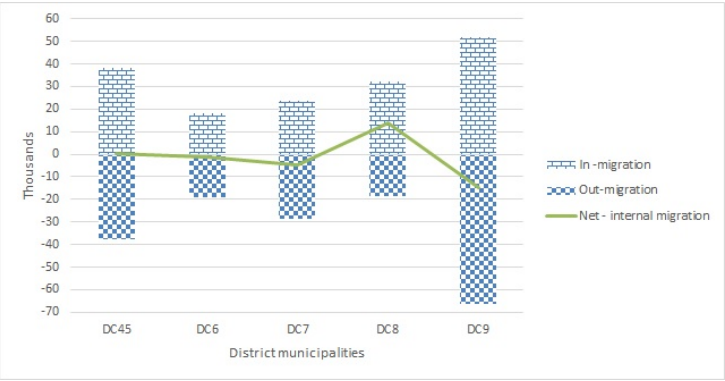

Figure 4.10: In-migration, Out-migration and Net-Internal Migration in NC district municipalities

4.3 Migration patterns

In this section migration patterns are presented and a review of the district with highest net-internal migration, and those with the lowest net-internal migration. During the 2011 census, the following had the highest net-internal migration TSH (Tshwane), EKU (Ekurhuleni) and CPT (City Of Cape Town metro). The DC15 (O.R Tambo), NMA (Nelson Mandela Bay Metropolitan) and DC35 Capricorn, had the lowest net-internal migration.

4.3.1 In-migration to Tshwane Metropolitan Municipality (TSH)

During the 2011 Census, the bulk of the migrants that were in Tshwane, were coming from Limpopo district, DC35 and Ekurhuleni Metro (Figure 4.11).

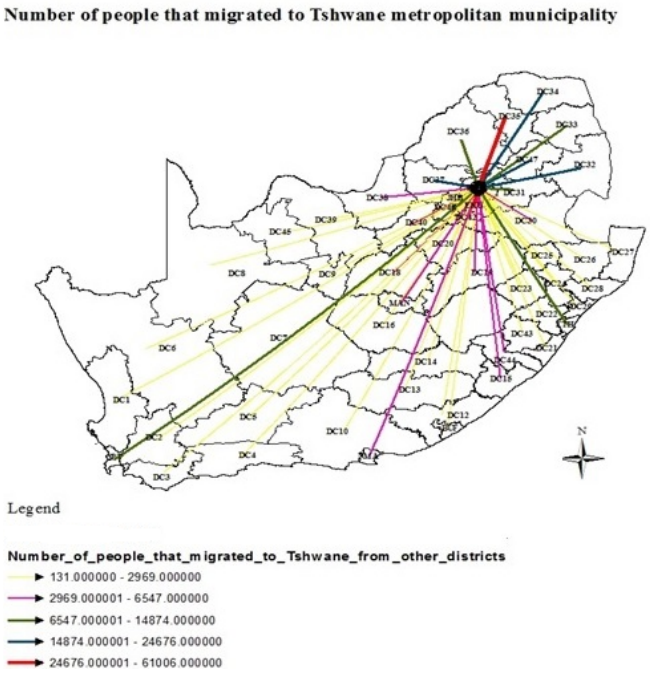

Figure 4.11: In-migration in Tshwane Metropolitan municipality

4.3.2 In-migration to Ekurhuleni Metropolitan Municipality (EKU)

The bulk of the migrants to Ekurhuleni, were coming from JHB, DC35, ETH, NMA and CPT metro (Figure 4.12).

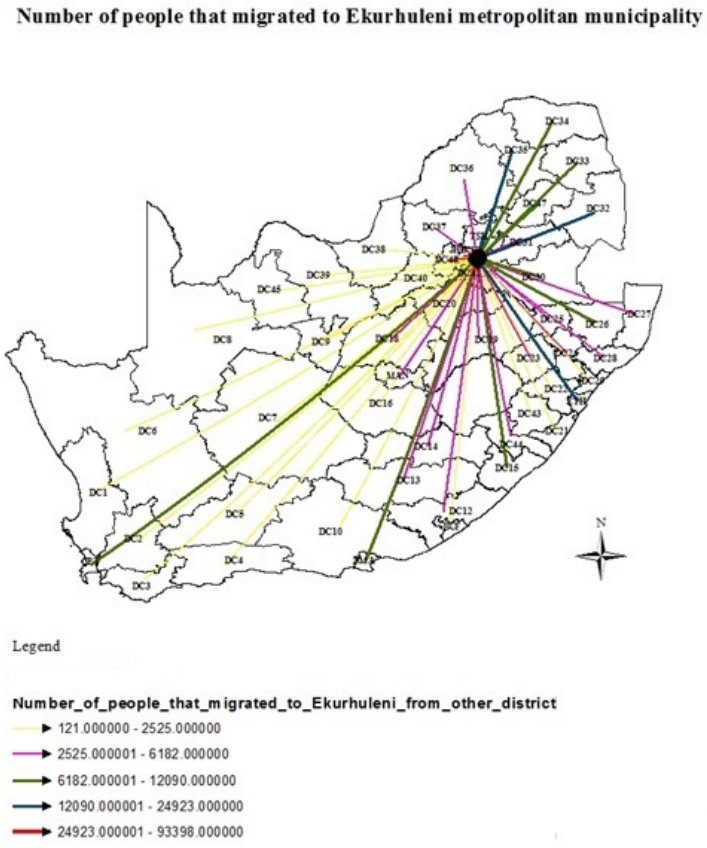

Figure 4.12: In-migration in Ekurhuleni Metropolitan municipality

4.3.3 In-migration to the City of Cape Town (CPT)

Figure 4.13 Shows that, most of the migrants to the City of Cape Town, were coming from JHB metro and Eastern Cape districts, namely, DC12 (Amathole), DC15 and NMA metro.

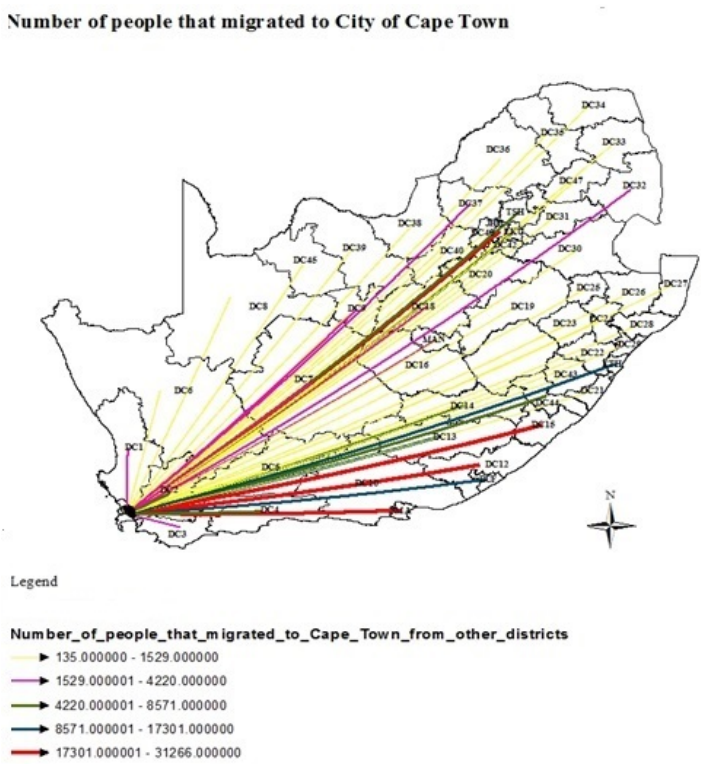

Figure 4.13: In-migration in the City of Cape Town Metropolitan municipality

4.3.4 Out-migration from O.R Tambo (DC15)

Figure 4.14 reveals that the migrants from O.R. Tambo moved to CPT, ETH, DC37, TSH and JHB. However, district municipalities in these provinces, such as Northern Cape, Limpopo and Mpumalanga were not attracting many migrants from DC15.

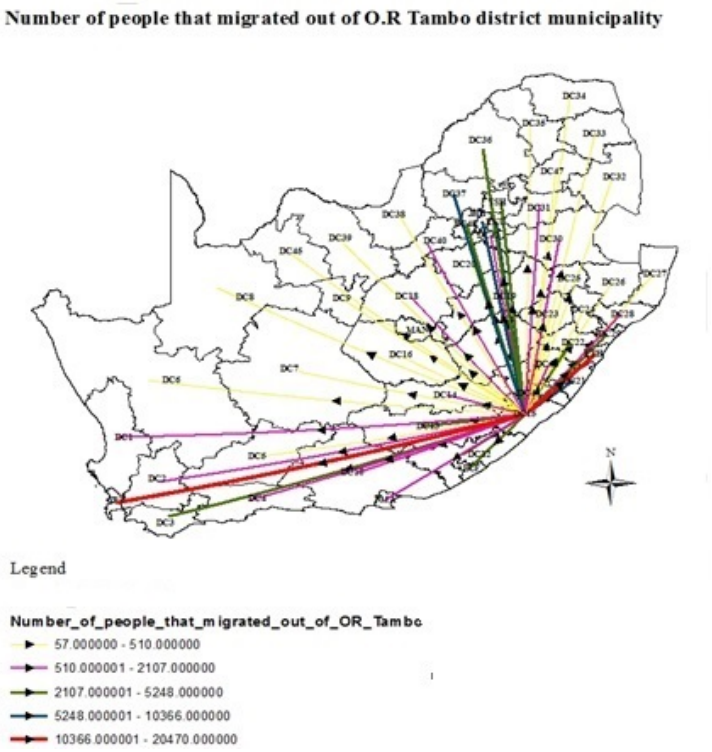

Figure 4.14: Out-migration in O.R Tambo District Municipality

4.3.5 Out-migration from Nelson Mandela Metropolitan Municipality (NMA)

Figure 4.15 shows that the bulk of the migrants from NMA , moved to CPT , BUF, DC10 (Cacadu) , JHB and ETH . The Northern Cape, Limpopo and Mpumalanga district municipalities were not attracting many migrants from NMA and DC15 as shown in Figure 4.14.

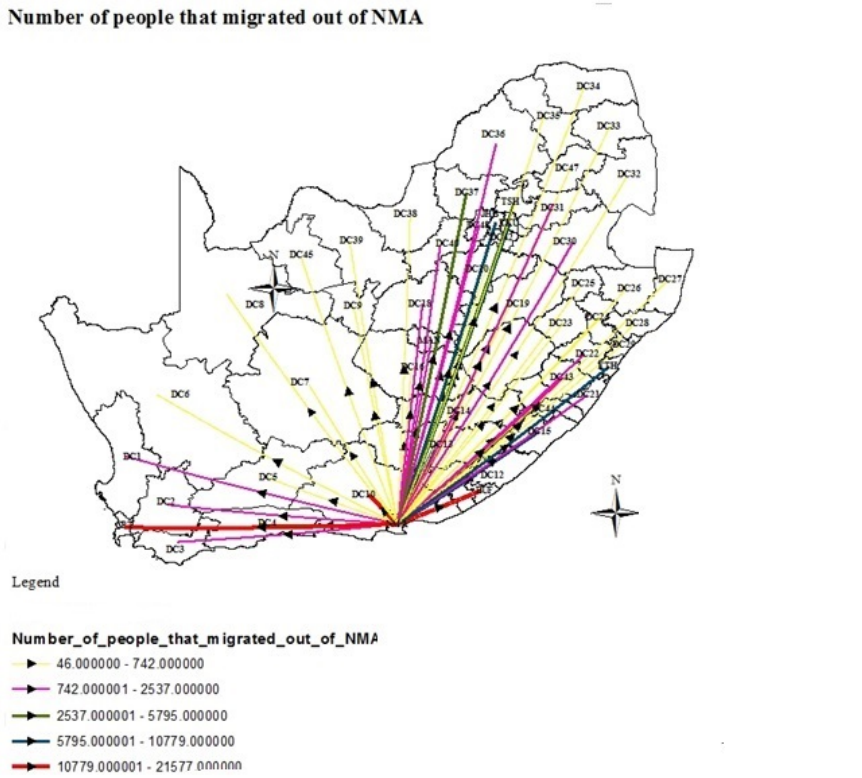

Figure 4.15: Out-migration in Nelson Mandela Metropolitan municipality(NMA)

4.3.6 Out-migration from Capricorn (DC35)

Figure 4.16 indicates that large numbers of the migrants from Capricorn district municipality migrated to TSH, EKU, JHB, DC47 (Greater Sekhukhune), DC36 and DC37. This indicates that the migrants from Capricorn municipality, migrate, between three provinces, Gauteng, Limpopo and North West. Another observation from the map was that, the migrants from Capricorn, they turn to migrate to near proximity district municipalities.

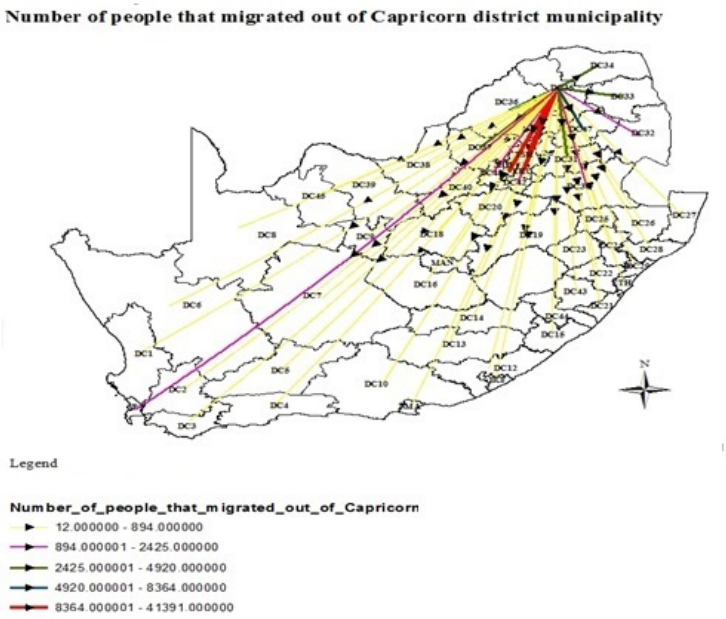

Figure 4.16: Out-migration from Capricorn District municipality

4.4 Modelling Results

This section presents the results of the models, as described in chapter 2. The following models, Gravity, Extended Gravity and Nonlinear Gravity, Poisson, NB, Gamma and GWR model. This section present the results, as well as diagnostic plots for the models to test of the model assumptions. The model performance will also be investigated.

The definition of the variables that are used for a Gravity model are shown in Table 4.1.

Table 4.1: Variables used for the Gravity model

Variables	Definition
M_{ij}	Migration flow between district i and j
D_{ij}	Distance between district municipalities i and j
pop_i	Population size at the origin i
pop_j	Population size at the destination j

4.4.1 Gravity model

This section presents the results of the Gravity model as defined in equation (2.1).

The population at the district of origin (pop_i) and the population at the destination (pop_j) coefficients are positive, this means an increase in either of the populations results in an increase in, both the in-or out-migration. While an increase in the distance (D_{ij}) between district municipalities decrease the in-or out-migration. The distance coefficient in both in- and out-migration is negative, this tells us that a 1 unit increase in the distance between the two district municipalities, decreases migration by $e^{-1.43}$ in Table 4.2 which is approximate to 0.24. These results suggest a similar pattern to that observed by Millington (2000). According to Henry et al (2003) and Fan (2005) this relationship agrees with theory, or this is the expected relationship. The Gravity models that predict, in- and out-migration were found significant, $F(3, 2648) = 1202$, $p - value = 2.2e^{-16}$, and $F(3,2648) = 1216$, $p - value = 2.2e^{-16}$, respectively. The Gravity model explains close to 60% of the observed variation in migration, and there is no evidence of multicollinearity. All the VIF's for the independent variables are less than 6. The predictors are significant at the 5% level.

It was also observed that the beta parameters for the populations were less than 1. According to Flowerdew and Amrhein (1989) when this happens both in-and

out-migration become less likely for larger cities.

Table 4.2: Results of the Gravity model

Independent variables	In-migration				Out-migration			
	Estimate	SE	pvalue	VIF	Estimate	SE	p-value	VIF
intercept	-9.00	0.67	$< 2e^{-16}$		-9.05	0.66	$< 2e^{-16}$	
$\ln(D_{ij})$	-1.43	0.04	$< 2e^{-16}$	1.03	-1.43	0.04	$< 2e^{-16}$	1.03
$\ln(pop_i)$	0.85	0.03	$< 2e^{-16}$	1.02	0.91	0.03	$< 2e^{-16}$	1.02
$\ln(pop_j)$	0.91	0.03	$< 2e^{-16}$	1.02	0.85	0.03	$< 2e^{-16}$	1.02

4.4.2 Extended Gravity model

This section presents, the variables used in Table 4.3 and the results of the Extended Gravity model.

Table 4.3: Variables used for the Extended Gravity model

Variables	Definition
D_{ij}	Distance between district municipalities i and j
ocr	Tenure status:Occupied rent free)
ofp	Tenure status :Owned and fully paid off
poc	Coloured Population
poI	Indian Population
emp	Employment rate
GDP	Gross Domestic Product
CPI	Consumer Price Index
ad	Adult Population
pwa	Access to tap water
$Inftrdwell$	Informal or Traditional dwelling

At 5% level, in-migration in the districts was found to be linearly related to 21 significant variables (pull factors). The following variables, $\ln(obp_j)$, $\ln(poc_i)$, $\ln(poc_j)$, $\ln(emp_i)$, $\ln(CPI_j)$, $\ln(ad_i)$, $\ln(ad_j)$ and $\ln(Infrdwell_i)$ were positive related with in-migration. This indicates that an increase in $\ln(CPI_j)$ increase the in-migration by $e^{1.34}$. The out-migration was found to be related to 20 significant factors (push factors), and seven of these factors, $\ln(obp_i)$, $\ln(poc_i)$, $\ln(poc_j)$, $\ln(emp_j)$, $\ln(CPI_i)$, $\ln(ad_i)$ and $\ln(ad_j)$ were positively related to out-migration.

Furthermore, the following demographic variables, $\ln(poc_i)$, $\ln(poc_j)$, $\ln(ad_i)$ and $\ln(ad_j)$ shows the positive relationship with both in- and out-migration in Table 4.4.

The Extended Gravity models that predict, in- and out-migration were found significant, $F(21,\ 2630)$ = 483.8, $p - value$ = $2.2e^{-16}$, and $F(20,2631)$ = 512.9, $p-value$ = $2.2e^{-16}$, respectively. The adjusted R^2 (R^2_{adj}) of the Extended Gravity model for both in- and out-migration was 79.3% and 79.4%, respectively, and it was higher compared to the R^2 of the simple Gravity model.

The results suggests that the coefficient of the distance variable was negative, this showed that migration was less likely as distance between two district municipalities increases. The coefficient of the log of the employment rate at the origin suggest that, the employment rate was positively related with in-migration, and employment rate at the destination was positively related with out-migration. Surprisingly the log of the GDP at the destination and origin were negative related with in-and out-migration. However, this observation contradicts with the findings of (Fan, 2005), who noted that, if migrants moved from the less developed to more developed provinces, then the beta coefficient of the log of the GDP_i (GDP at the origin) is expected to be negative, and the beta coefficient of the log of the GDP_j (destination) is expected to be positive. This could be the effect of averaging out since the province GDP was used as a proxy.

Table 4.4: Results of the Extended Gravity model

Independent variables	In-migration			Out-migration		
	Estimate	SE	p-value	Estimate	SE	p-value
intercept	27.91	3.16	$< 2e^{-16}$	27.89	3.15	$< 2e^{-16}$
$\ln(D_{ij})$	-1.71	0.03	$< 2e^{-16}$	-1.69	0.03	$< 2e^{-16}$
$\ln(obp_i)$	-0.51	0.08	$1.85e^{-09}$	0.45	0.08	$4.03e^{-09}$
$\ln(obp_j)$	0.41	0.08	$1.11e^{-07}$	-0.48	0.08	$1.52e^{-08}$
$\ln(ocr_i)$	-0.77	0.11	$3.24e^{-11}$			
$\ln(ofp_i)$	-1.71	0.17	$< 2e^{-16}$	-0.61	0.13	$4.53e^{-06}$
$\ln(ofp_j)$	-0.63	0.13	$2.67e^{-06}$	-1.75	0.16	$< 2e^{-16}$
$\ln(poc_i)$	0.19	0.02	$< 2e^{-16}$	0.17	0.02	$< 2e^{-16}$
$\ln(poc_j)$	0.18	0.02	$< 2e^{-16}$	0.18	0.02	$< 2e^{-16}$
$\ln(poI_i)$	-0.06	0.02	$8.44e^{-03}$	-0.06	0.02	$6.98e^{-03}$
$\ln(poI_j)$	-0.07	0.02	$1.08e^{-03}$	-0.05	0.02	0.01
$\ln(emp_i)$	0.89	0.07	$< 2e^{-16}$	-0.30	0.07	$1.51e^{-05}$
$\ln(emp_j)$	-0.31	0.07	0.84	$1.87e^{-05}$	0.07	$< 2e^{-16}$
$\ln(GDP_i)$	-0.30	0.05	$4.54e^{-11}$	-0.31	0.05	$4.28e^{-11}$
$\ln(GDP_j)$	-0.30	0.05	$5.37e^{-10}$	-0.28	0.04	$3.42e^{-10}$
$\ln(CPI_i)$	-1.97	0.29	$1.71e^{-11}$	1.28	0.29	$1.23e^{-05}$
$\ln(CPI_j)$	1.34	0.29	$5.08e^{-06}$	-1.70	0.27	$4.34e^{-10}$
$\ln(ad_i)$	0.84	0.04	$< 2e^{-16}$	1.18	0.04	$< 2e^{-16}$
$\ln(ad_j)$	1.20	0.04	$< 2e^{-16}$	0.85	0.03	$< 2e^{-16}$
$\ln(pwa_j)$	-1.92	0.18	$< 2e^{-16}$			
$\ln(Inftrdwell_i)$	0.08	0.04	0.029	-0.29	0.04	$2.93e^{-14}$
$\ln(Inftrdwell_j)$	-0.27	0.04	$1.12e^{-11}$			
$\ln(pwa_i)$				-1.95	0.18	$< 2e^{-16}$
$\ln(ocr_j)$				-0.79	0.11	$1.73e^{-12}$
F	493.80			512.90		
R^2	0.793			0.794		

4.4.3 Results: Reset Test

In Table 4.4 the sign of the GDP parameters is incorrect from the results of the Extended Gravity model. This can be attributed to the fact that GDP is a proxy and the provincial figure are distorting the statistics. In this regard, the study presents the Reset Test results to investigate the model adequacy of the Extended Gravity regression. The Reset Test results, in Appendix E from the R, function $resettest$ was used, the results suggest that the parameter δ_1 was significant at 5%, because the $p-value < 2.2e^{-16}$ and it was less than the level of significance, 5%, thus it was concluded that the Extended Gravity model was misspecified.

4.4.4 Nonlinear model

The significance of the beta parameters for quadratic terms in Table 4.5 suggests a nonlinear relationship. Quadratic terms $[\ln D_{ij}]_c^2$, $[\ln(poIc_i)]_c^2$ and $[\ln(adc_j)]_c^2$ have negative beta values and the rest of the quadratic terms have significant positive values. Ganzach (1997) said a significant negative beta parameter of a square predictor in a regression model indicates a concave and a significantly positive value suggests a convex relationship. From the results, the distance effect is -0.19$[\ln D_{ij}]_c^2$ -1.87$[\ln(D_{ij})]_c$. Let d = $[\ln(D_{ij})]_c$, so that the equation is -0.19d^2 -1.87d, and has a maximum at,

$$\begin{aligned} \frac{\partial ln(M_{in})}{\partial d} &= 0 \\ -0.38d - 1.87 &= 0 \\ d &= -4.92 \end{aligned} \tag{4.1}$$

Note that d =$[\ln(D_{ij})]_c$, and $[\ln(D_{ij})]_c$ = -4.92. The variable $[\ln(D_{ij})]_c$ is centered, and therefore $[\ln(D_{ij})]_c$ is equal to $\ln(D_{ij})$-$\overline{\ln(D_{ij})}$, and the mean ($\overline{\ln(D_{ij})}$) is given as 6.54 in Table C.2. So the distance effect increases in-migration as one travels to around $e^{-4.92+6.54}$ = 5.05km, which is approximately equal to 5km but then it is expected to decrease in-migration (M_{in}) as one continues to move further. However, this value is outside the range of the observed values, and the lowest distance value is 44km from Appendix C in Table C.1. This supports the fact that as the distance increase in-migration decreases. The signs of the main effects in Table 4.5 compare to those in Table 4.4 did not change. Except the sign for the log of the variable Indian/Asian population at the origin ($\ln(poI_i)$), the

sign change in Table 4.5, and the variable $[\ln(Inftrdwell_j)]_c$ was not significant. The model was found significant, F(24, 2627) = 434 and $p - value$ = 2.2e^{-16}. The R^2 of the nonlinear model was 0.7969 close to 0.8 (80%) suggesting it is a good fit. The main effect of the quadratic term ($[\ln(poI_j)]_c^2$), namely, $[\ln(poI_j)]_c$ in Table 4.5 changes sign as compared to Table 4.4 but it is not significant. The VIF's were less than 6 this suggests that there was no serious presence of multi-collinearity among the predictors. The model that predicts out-migration (M_{out}) in Table 4.5 also had significant quadratic terms. Furthermore, the coefficient of the linear term for the log of employment rate at the destination ($[\ln(emp_j)]_c$) was positively related to out-migration and its quadratic term was also positive. From this expression, 1.07$[\ln(emp_j)]_c^2$ + 1.39$[\ln(emp_j)]_c$, Let $[\ln(emp_j)]_c$ =w, so that 1.07w^2 + 1.39w. This means that the out-migration has minimum at,

$$\begin{aligned} \frac{\partial \ln(M_{out})}{\partial w} &= 0 \\ 2.14w + 1.39 &= 0 \\ w &= -0.65 \end{aligned} \tag{4.2}$$

substitute, $[\ln(emp_j)]_c$ =w = -0.65, then emp_j =$e^{-0.65+3.45}$ =16.44. This indicates that the employment rate at the destination decreases out-migration at 16.44%, and after that value in-migration increase. The model was found significant, F(24, 2627) = 445 and $p - value$ = 2.2e^{-16}. The R^2 of the nonlinear model was 0.8008 (80.08%) suggesting its a good fit.

Table 4.5: Results of the Nonlinear model

Independent variables	In-migration				Out-migration			
	Estimate	SE	p-value	VIF	Estimate	SE	p-value	VIF
intercept	4.91	0.05	$< 2e^{-16}$		4.97	0.05	$< 2e^{-16}$	
$[\ln(D_{ij})]_c$	-1.87	0.04	$< 2e^{-16}$	1.73	-1.81	0.04	$< 2e^{-16}$	1.78
$[\ln(obp_i)]_c$	-0.31	0.08	$5.88e^{-05}$	3.60	0.47	0.08	$6.57e^{-10}$	3.45
$[\ln(obp_j)]_c$	0.41	0.08	$1.29e^{-07}$	3.52				
$[\ln(ocr_i)]_c$	-0.19	0.09	0.04	1.80				
$[\ln(ocr_j)]_c$					-0.37	0.11	$6.33e^{-04}$	2.45
$[\ln(ofp_j)]_c$					-0.78	0.17	$7.61e^{-06}$	5.16
$[\ln(poc_i)]_c$	0.2	0.02	$< 2e^{-16}$	3.45	0.14	0.02	$< 2e^{-16}$	2.91
$[\ln(poc_j)]_c$	0.15	0.02	$< 2e^{-16}$	3.01	0.14	0.02	$< 2e^{-16}$	3.04
$[\ln(poI_i)]_c$	0.11	0.03	$1.80e^{-05}$	5.59	-0.16	0.03	$1.21e^{-10}$	5.54
$[\ln(poI_j)]_c$	-0.17	0.03	$2.25e^{-11}$	5.71	$12.25e^{-04}$	0.02	0.96	5.31
$[\ln(emp_i)]_c$	1.58	0.08	$< 2e^{-16}$	3.26	-0.53	0.07	$6.18e^{-15}$	2.36
$[\ln(emp_j)]_c$	-0.52	0.07	$1.58e^{-13}$	2.50	1.39	0.10	$< 2e^{-16}$	5.50
$[\ln(GDP_i)]_c$	-0.15	0.05	$1.24e^{-03}$	3.67	-0.36	0.04	$< 2e^{-16}$	2.38
$[\ln(GDP_j)]_c$	-0.35	0.04	$< 2e^{-16}$	2.38	-0.30	0.04	$2.47e^{-11}$	3.63
$[\ln(CPI_i)]_c$	-1.15	0.27	$2.00e^{-05}$	2.51				
$[\ln(CPI_j)]_c$					-1.34	0.27	$8.21e^{-07}$	2.59
$[\ln(ad_i)]_c$	0.73	0.04	$< 2e^{-16}$	3.05	1.31	0.04	$< 2e^{-16}$	3.02
$[\ln(ad_j)]_c$	1.31	0.04	$< 2e^{-16}$	3.06	0.84	0.03	$< 2e^{-16}$	2.68
$[\ln(Inftrdwell_i)]_c$					-0.03	0.04	0.36	1.93
$[\ln(Inftrdwell_j)]_c$	$-24.17e^{-04}$	0.04	0.94	1.99				
$[\ln(D_{ij})]_c^2$	-0.19	0.04	$1.33e^{-07}$	1.57	-0.18	0.03	$2.18e^{-07}$	1.58
$[\ln(poc_i)]_c^2$					0.06	$60.50e^{-04}$	$< 2e^{-16}$	1.61
$[\ln(poc_j)]_c^2$	0.06	$61.03e^{-04}$	$< 2e^{-16}$	1.60				
$[\ln(poI_i)]_c^2$	-0.03	$58.27e^{-04}$	$2.47e^{-08}$	1.90	0.05	$76.88e^{-04}$	$4.96e^{-11}$	3.38
$[\ln(poI_j)]_c^2$	0.05	$77.90e^{-04}$	$5.562e^{-11}$	3.39	-0.03	$55.56e^{-04}$	$2.65e^{-06}$	1.75
$[\ln(emp_i)]_c^2$	1.64	0.19	$< 2e^{-16}$	2.46				
$[\ln(emp_j)]_c^2$					1.07	0.22	$1.29e^{-06}$	3.51
$[\ln(GDP_i)]_c^2$					0.39	0.03	$< 2e^{-16}$	1.85
$[\ln(GDP_j]_c^2$	0.38	0.03	$< 2e^{-16}$	1.83				
$[\ln(ad_i)]_c^2$					-0.34	0.04	$< 2e^{-16}$	5.11
$[\ln(ad_j)]_c^2$	-0.33	0.04	$< 2e^{-16}$	5.00				
$[\ln(Inftrdwell_i)]_c^2$					0.38	0.04	$< 2e^{-16}$	2.88
$[\ln(Inftrdwell_j)]_c^2$	0.36	0.04	$< 2e^{-16}$	2.88				
F	434				445			
R^2	0.7969				0.8006			

4.4.5 Comparison of the models that Predict Out-migration

This section compares the models that were fitted for out-migration. The purpose of this comparison was to identify the best performing model. The models are Gravity, Extended Gravity and Nonlinear models, respectively.

In Table 4.6, list the values of AIC and R^2, and suggests that the Non linear model

was the best performing among the fitted models, as it had a lower AIC and the highest R^2.

Table 4.6: Results of the model selection for Out-migration

Models	AIC	R^2
Gravity model	8637.973	0.579
Extended gravity model	6756.633	0.794
Nonlinear model	6673.679	0.801

4.4.6 Comparison of the models that Predict In-migration

This section compares the models that were fitted for the prediction of in-migration in the district municipalities of South Africa. The purpose of this section was to identify the best model that describes in-migration.

According to the results in Table 4.7 the Nonlinear model was the best performing among the model. Because it had a lower AIC and the highest R^2. Although this difference is only slight.

Table 4.7: Results of the model selection for In-migration

Models	AIC	R^2
Gravity model	8664.183	0.576
Extended gravity model	6785.451	0.793
Nonlinear model	6734.491	0.797

4.4.7 Diagnostic plots of the OLS models

This section presents the diagnostic plots of the models, plots of the residuals against the fitted values, normal probability, residuals against the fitted values and residuals against leverage. The plots, shown in Figure 4.17 and Figure 4.18, suggests that the residuals were distributed around zero and there was no pattern

between the residuals and fitted values. Montgomery et al (2006) indicated that if the residuals showed no patterns then the model may be adequate. On this basis, this study concluded that there were no model defects. However, there were few non influential outliers and the normality probability plot indicates that there could be slight normality problems, because some observations were deviating away from the straight line especially in the out-migration model.

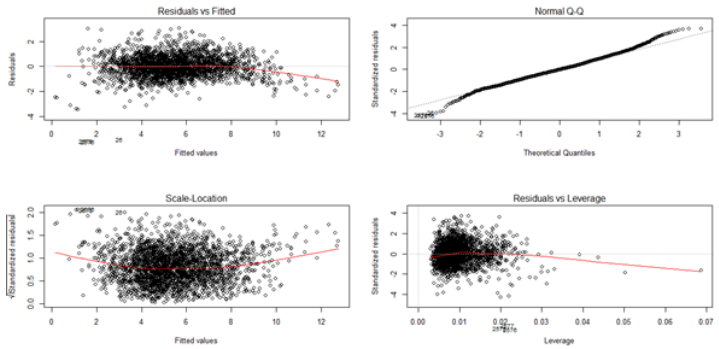

Figure 4.17: Diagnostic: Nonlinear model (In-migration)

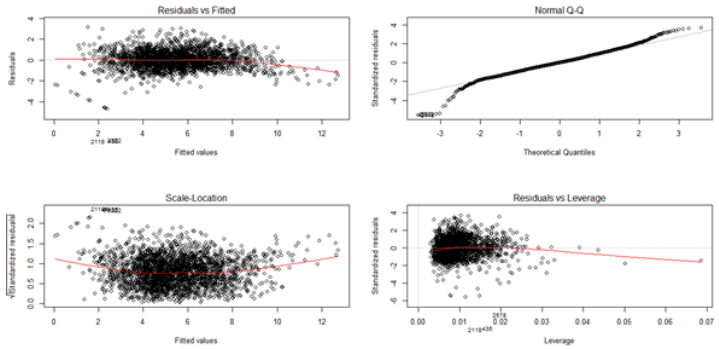

Figure 4.18: Diagnostic: Nonlinear model (Out-migration)

4.4.8 Normality Assessment

From the model selection results, it was concluded that Nonlinear model was the best performing model. This section test the normality assumption in nonlinear model.

H_0: the residuals of Nonlinear model are normally distributed.

H_a: the residuals of Nonlinear model are not normally distributed.

The JB and Shapiro-Wilk normality Test, at 5 % level of significance, the results in Table 4.8 suggests that the residuals of the models were not normal and this indicates that the models were biased and the predictions from these models are not reliable.

Table 4.8: Results of the normality test

Models	Shapiro-Wilk normality Test		Jarque Bera Test	
	W	$p-value$	X^2	$p-value$
$Nonlinear_{in-migration}$	0.9823	$< 2.2e^{-16}$	468.8199	$< 2.2e^{-16}$
$Nonlinear_{out-migration}$	0.9814	$< 2.2e^{-16}$	502.0778	$< 2.2e^{-16}$

4.4.9 Testing for Autocorrelation

To prove that the error terms of the models were uncorrelated or independent, the Durbin-Watson (D-W) test statistic was used. The D-W statistic using the $durbinWatsonTest()$ function in R. The following hypothesis test;

H_0: $\rho = 0$ or the error terms of the Nonlinear model are independently normally distributed.

H_a: $\rho \neq 0$ or the error terms of the Nonlinear model follows the first-order autoregressive process.

Table 4.9, shows that at 5%, ρ parameter was different from zero, because $d = 1.448194 \leq d_L$= 1.554. From these results, the null hypothesis was rejected and conclude that the error terms of the model ($Nonlinear_{In-migration}$) are positively autocorrelated, the same was observed with the $Nonlinear_{Out-migration}$ model.

Table 4.9: Results of the Autocorrelation test

Models	lag	$Autocorrelation$	$D-WStatistic$	$p-value$
$Nonlinear_{In-migration}$	1	0.2758861	1.448194	0
$Nonlinear_{Out-migration}$	1	0.4243746	1.15095	0

4.4.10 Outlier Identification for OLS model

This section presents the results obtained from the Bonferonni method, which was used to identify the presence of outliers from the Nonlinear models (for both In- and Out-migration).

The results suggest that there were outliers in the model in Table 4.10 and Table 4.11. In the case of the Nonlinear model that predicts in-migration, the following observations, were significant outliers, 1066, 1367, 1371, 1374 and 1386, and for out-migration, these observation were significant outliers, 435, 1427, 2118, 2322, 2367 and 2475. The observations 1367, 1371 and 1374, represents the in-migrants from Zululand in Namakwa, John Taolo Gaentswe and Overberg district municipality in Table 4.10. Also the data point 1066, represent in-migrants from Uthukela that were reported in West Coast district municipality, and the observation 1386 represents the in-migrants from iLembe that were in Xhariep district municipality .

In Table 4.11 the unsual points or outliers correspond to 2118, 2322, and 2475

suggest that they were 22, 2, and 30 out-migrants from Zululand that moved to Namakwa, John Taolo Gaentswe and Overberg district municipality. There were 145 out-migrants from Uthukela that moved to West Coast district municipality. From iLembe they were 22 out-migrants that were reported in Xhariep district municipality. They were no out-migrants from Central karoo to iLembe district municipality. The result suggest that there is no chance that the population in Central Karoo migrates to iLembe district municipality. This is understood because the distance between iLembe and Central karoo district municipalities is around 1298 km. From the Nonlinear model results, this study noted that if the distance between district municipalities is above 5.05km then in-migration is expected to be less likely between those districts.

This study further investigated the effect of the outliers by removing the outliers and refitting the models. It was noted that the change in values of the diagnostics statistics such as, R^2 and AIC for these Nonlinear models that predicts in-migration and out-migration, the R^2 changed from 0.7968 to 0.8016, and 0.8008 to 0.806, respectively and, the AIC values decreased to 6580.254 and 6493.873 when the outliers were removed. After the outliers were removed the results in Table 4.12 and Table 4.13 indicates that there is positive autocorrelation and the residuals for the Nonlinear models failed the normality test.

Table 4.10: Results of the outlier identification: In-migration

Outlier observation	rstudent	unadjusted p-value	Bonferonni p
1386	-5.797154	$7.5511e^{-9}$	$2.0026e^{-05}$
1066	-5.549333	$3.1545e^{-08}$	$8.3656e^{-05}$
1367	-5.371076	$8.5131e^{-08}$	$2.2577e^{-04}$
1371	-5.334528	$1.0396e^{-07}$	$2.7571e^{-04}$
1374	-4.492200	$7.3539e^{-06}$	$1.9503e^{-02}$

Table 4.11: Results of the outlier identification: Out-migration

Outlier observation	rstudent	unadjusted p-value	Bonferonni p
2118	-5.614844	$2.1738e^{-08}$	$5.7649e^{-05}$
435	-5.510930	$3.9166e^{-08}$	$1.0387e^{-04}$
2322	-5.473942	$4.8179e^{-08}$	$1.2777e^{-04}$
2367	-5.435374	$5.9711e^{-08}$	$1.5835e^{-04}$
1427	-4.683809	$2.9593e^{-06}$	$7.8480e^{-03}$
2475	-4.581270	$4.8374e^{-06}$	$1.2829e^{-02}$

Table 4.12: Outliers removed: the normality test

Models	Shapiro-Wilk normality Test		Jarque Bera Test	
	W	$p-value$	X^2	$p-value$
$Nonlinear_{In-migration}$	0.9923	$1.255e^{-10}$	100.9858	$< 2.2e^{-16}$
$Nonlinear_{Out-migration}$	0.9928	$3.264e^{-10}$	93.0686	$< 2.2e^{-16}$

Table 4.13: Outliers removed: the autocorrelation test

Models	lag	$Autocorrelation$	$D-WStatistic$	$p-value$
$Nonlinear_{In-migration}$	1	0.2795935	1.440786	0
$Nonlinear_{Out-migration}$	1	0.457688	1.0843	0

4.4.11 Poisson model results

The Poisson model was fitted to the data and the results displayed in Table 4.14. The deviance $p-value$ is almost zero for both models this shows that model failed to fit the data.

Table 4.14: Results of the Poisson model

Independent variables	In- migration			Out-migration		
	Estimate	SE	p-value	Estimate	SE	p-value
$intercept$	6.33	$2.34e^{-02}$	$< 2e^{-16}$	7.43	$2.25e^{-02}$	$< 2e^{-16}$
D_{ij}	$-2.49e^{-03}$	$1.91e^{-06}$	$< 2e^{-16}$	$-2.40e^{-03}$	$1.90e^{-06}$	$< 2e^{-16}$
obp_i	$4.52e^{-03}$	$3.53e^{-04}$	$< 2e^{-16}$	0.04	$2.96e^{-04}$	$< 2e^{-16}$
obp_j	0.03	$2.95e^{-04}$	$< 2e^{-16}$	$6.39e^{-03}$	$3.36e^{-04}$	$< 2e^{-16}$
ocr_i	$-2.78e^{-02}$	$2.62e^{-04}$	$< 2e^{-16}$			
ofpi	$-1.80e^{-02}$	$1.63e^{-04}$	$< 2e^{-16}$	$-8.55e^{-03}$	$1.23e^{-04}$	$< 2e^{-16}$
ofp_j	$-2.96e^{-03}$	$1.23e^{-04}$	$< 2e^{-16}$	$-2.22e^{-02}$	$1.61e^{-04}$	$< 2e^{-16}$
poc_i	$1.69e^{-07}$	$3.23e^{-09}$	$< 2e^{-16}$	$-3.69e^{-07}$	$3.54e^{-09}$	$< 2e^{-16}$
poc_j	$-2.65e^{-07}$	$3.51e^{-9}$	$< 2e^{-16}$	$1.48e^{-7}$	$3.20e^{-9}$	$< 2e^{-16}$
poI_i	$-1.69e^{-06}$	$7.14e^{-09}$	$< 2e^{-16}$	$-8.81e^{-07}$	$6.53e^{-09}$	$< 2e^{-16}$
poI_j	$-7.97e^{-07}$	$6.54e^{-09}$	$< 2e^{-16}$	$-1.60e^{-06}$	$7.17e^{-09}$	$< 2e^{-16}$
emp_i	$2.39e^{-02}$	$1.32e^{-04}$	$< 2e^{-16}$	$-3.77e^{-02}$	$9.05e^{-05}$	$< 2e^{-16}$
emp_j	$-3.29e^{-02}$	$9.56e^{-05}$	$< 2e^{-16}$	$2.40e^{-02}$	$1.17e^{-04}$	$< 2e^{-16}$
GDP_i	$-2.47e^{-13}$	$4.36e^{-15}$	$< 2e^{-16}$	$-3.05e^{-13}$	$4.73e^{-15}$	$< 2e^{-16}$
GDP_j	$-2.13e^{-13}$	$4.64e^{-15}$	$< 2e^{-16}$	$-2.65e^{-13}$	$4.32e^{-15}$	$< 2e^{-16}$
CPI_i	$1.46e^{-01}$	$9.79e^{-04}$	$< 2e^{-16}$	$2.59e^{-01}$	$1.14e^{-03}$	$< 2e^{-16}$
CPI_j	$2.81e^{-01}$	$1.12e^{-03}$	$< 2e^{-16}$	$1.18e^{-01}$	$9.77e^{-04}$	$< 2e^{-16}$
ad_i	$1.36e^{-05}$	$2.21e^{-08}$	$< 2e^{-16}$	$2.01e^{-05}$	$2.37e^{-08}$	$< 2e^{-16}$
ad_j	$1.96e^{-05}$	$2.35e^{-08}$	$< 2e^{-16}$	$1.28e^{-05}$	$2.23e^{-08}$	$< 2e^{-16}$
pwa_j	$-1.05e^{-02}$	$9.38e^{-05}$	$< 2e^{-16}$			
$Inftrdwell_i$	$6.23e^{-03}$	$7.62e^{-05}$	$< 2e^{-16}$	$-1.54e^{-02}$	$7.90e^{-05}$	$< 2e^{-16}$
$Inftrdwell_j$	$-1.23e^{-02}$	$7.87e^{-05}$	$< 2e^{-16}$			
ocr_j				$-3.13e^{-02}$	$2.48e^{-04}$	$< 2e^{-16}$
pwa_i				$-1.09e^{-02}$	$9.28e^{-05}$	$< 2e^{-16}$
degrees of freedom	2630			2631		
Deviance	2550636			2233367		
Goodness-of-fit			0			0

4.4.12 Overdispersion Test results

This section investigates whether the Poisson model was either over or underdispersed, in a form of hypothesis,

H_0: dispersion parameter of the Poisson model is not greater than 1.

H_a: dispersion parameter of the Poisson model is greater than 1.

Table 4.15 indicates that at 5% level of significance, the dispersion parameters were greater than 1. The dispersion parameters for the models were 1216.875 and 1586.846. This observation further highlights that the Poisson models were over dispersed, that means that the standard errors of the models were not reliable.

From the goodness of fit, it was observed that the Poisson model failed to fit the data at 5% level of significant.

Table 4.15: Overdispersion test results

Model	$Z-value$	$p-value$	Dispersion estimates	AIC
$Poisson_{Out-migration}$	5.4564	$2.429e^{-08}$	1216.875	2252676
$Poisson_{In-migration}$	4.3932	$5.586e^{-06}$	1586.846	2569951

4.4.13 Negative Binomial (NB) model results

Since the Poisson model was overdispersed, the NB model was used to account for overdispersion, in predicting migration. The results show that the $p-value$ of the deviance statistic for in-and out-migration in Table 4.16 is almost zero. This shows the failure of the NB model in fitting the migration data.

Table 4.16: Results of the Negative Binomial model

Independent variables	In-migration			Out-migration		
	Estimate	SE	p-value	Estimate	SE	p-value
(Intercept)	10.3	$7.03e^{-01}$	$< 2e^{-16}$	10.5	$6.75e^{-01}$	$< 2e^{-16}$
D_{ij}	$-2.49e^{-03}$	$5.08e^{-05}$	$< 2e^{-16}$	$-2.43e^{-03}$	$4.96e^{-05}$	$< 2e^{-16}$
$obpi$	$-4.5e^{-02}$	$1.17e^{-02}$	$8.42e^{-05}$	$6.06e^{-02}$	$9.45e^{-03}$	$1.43e^{-10}$
obp_j	$5.84e^{-02}$	$9.66e^{-02}$	$1.47e^{-09}$	$-3.66e^{-02}$	$1.10e^{-02}$	$8.84e^{-04}$
ocr_i	$-4.51e^{-02}$	$6.29e^{-03}$	$7.40e^{-13}$			
ofp_i	$-4.28e^{-02}$	$5.02e^{-03}$	$< 2e^{-16}$	$-1.65e^{-03}$	$3.74e^{-03}$	0.66
ofp_j	$4.23e^{-03}$	$3.84e^{-03}$	0.27	$-4.00e^{-02}$	$4.83e^{-03}$	$< 2e^{-16}$
poc_i	$2.16e^{-07}$	$1.30e^{-07}$	0.01	$3.22e^{-07}$	$1.26e^{-07}$	0.01
poc_j	$3.42e^{-07}$	$1.29e^{-07}$	$7.94e^{-03}$	$1.18e^{-07}$	$1.26e^{-07}$	0.35
poI_i	$-2.27e^{-06}$	$3.02e^{-07}$	$5.97e^{-14}$	$-1.18e^{-06}$	$2.94e^{-07}$	$6.28e^{-05}$
poI_j	$-1.20e^{-06}$	$3.02e^{-07}$	$6.94e^{-05}$	$-2.13e^{-06}$	$2.95e^{-07}$	$5.18e^{-13}$
emp_i	$3.60e^{-02}$	$3.12e^{-03}$	$< 2e^{-16}$	$-6.73e^{-03}$	$2.53e^{-03}$	$7.74e^{-03}$
emp_j	$-5.55e^{-04}$	$2.67e^{-03}$	0.84	$3.70e^{-03}$	$2.60e^{-03}$	$< 2e^{-16}$
GDP_i	$-7.86e^{-13}$	$1.37e^{-13}$	$9.24e^{-09}$	$-5.59e^{-13}$	$1.36e^{-13}$	$4.00e^{-05}$
GDP_j	$-5.30e^{-13}$	$1.39e^{-13}$	$1.32e^{-04}$	$-6.97e^{-13}$	$1.27e^{-13}$	$4.15e^{-08}$
CPI_i	$-2.38e^{-01}$	$3.65e^{-02}$	$6.95e^{-11}$	$1.49e^{-01}$	$3.58e^{-02}$	$3.04e^{-05}$
CPI_j	$1.64e^{-01}$	$3.67e^{-02}$	$8.45e^{-06}$	$-2.29e^{-01}$	$3.48e^{-02}$	$4.36e^{-11}$
ad_i	$1.54e^{-05}$	$7.21e^{-07}$	$< 2e^{-16}$	$1.70e^{-05}$	$7.11e^{-07}$	$< 2e^{-16}$
ad_j	$1.69e^{-05}$	$7.31e^{-07}$	$< 2e^{-16}$	$1.48e^{-05}$	$7.06e^{-07}$	$< 2e^{-16}$
pwa_j	$-1.93e^{-02}$	$3.02e^{-03}$	$1.69e^{-10}$			
$Inftrdwell_i$	$5.16e^{-03}$	$1.81e^{-03}$	$4.35e^{-03}$	$-9.48e^{-03}$	$2.35e^{-03}$	$5.44e^{-05}$
$Inftrdwell_j$	$-6.59e^{-03}$	$2.45e^{-03}$	$7.18e^{-03}$			
ocr_j				$-4.27e^{-02}$	$5.81e^{-03}$	$1.99e^{-13}$
pwa_i				$-1.77e^{-02}$	$2.89e^{-03}$	$9.31e^{-10}$
degrees of freedom	2630			2631		
Deviance	3023.50			3010.80		
Theta	1.13	$2.81e^{-02}$		1.18	$2.97e^{-02}$	
Goodness-of-fit			$1.08e^{-07}$			$2.68e^{-07}$
AIC	36804			36692		

4.4.14 Gamma model results

From Figure 4.1 it is clear the distribution of the migration in the district municipalities is left skewed, in this section Gamma regression is used to model the data. The results from this model are presented in Table 4.18.

From the results, the coefficients of the variables, distance (D_{ij}), GDP and Indian/Asian population (poI) indicate that these were negatively related with migration. At 5% level of significance, there were only three variables that were not significant in predicting in-migration, those variables were, ofp_j, poc_i and

emp_j. Similarly, for out-migration only two variables were not significant, ofp_i and poc_j in Table 4.17. The variable distance had a negative coefficient -0.0025, meaning that for each one unit increase in distance, the expected log count of the in-migration decreased by 0.0025 holding all other predictors constant. This was similar to the results of out-migration, the distance variable had the same coefficient.

A value of 1 for the scale parameter indicates that the Gamma model is equivalent to the exponential distribution. The estimated value of the scale parameter for the Gamma model that predict in and out-migration was 1.13 and 1.19. The 95% confidence interval for the scale parameter of the models was (1.08, 1.18) and (1.13, 1.25), which does not contain 1. The hypothesis of an exponential distribution for the migration data is rejected at the 5% level. The NB and Gamma model showed the same patterns in terms of the parameter signs and the standard errors from these models are the same.

Table 4.17: Results of the Gamma model

Independent variables	In-migration			Out-migration		
	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	10.2	$7.00e^{-01}$	$< 2e^{-16}$	10.4	$6.73e^{-01}$	$< 2e^{-16}$
D_{ij}	$-2.47e^{-03}$	$5.05e^{-05}$	$< 2e^{-16}$	$-2.40e^{-03}$	$4.94e^{-05}$	$< 2e^{-16}$
obp_i	$-4.37e^{-02}$	$1.16e^{-02}$	$1.64e^{-04}$	$6.16e^{-02}$	$9.46e^{-03}$	$7.52e^{-11}$
obp_j	$5.93e^{-02}$	$9.65e^{-03}$	$-3.45e^{-02}$	$8.03e^{-10}$	$1.10e^{-02}$	$1.68e^{-03}$
ocr_i	$-4.37e^{-02}$	$6.26e^{-03}$	$2.98e^{-12}$			
ofp_i	$-4.18e^{-02}$	$5.00e^{-03}$	$< 2e^{-16}$	$-8.97e^{-04}$	$3.73e^{-03}$	0.81
ofp_j	$5.01e^{-03}$	$3.83e^{-03}$	0.19	$-3.90e^{-02}$	$4.82e^{-03}$	$5.62e^{-16}$
poc_i	$2.11e^{-07}$	$1.30e^{-07}$	0.10	$3.13e^{-07}$	$1.25e^{-07}$	0.01
poc_j	$3.35e^{-07}$	$1.28e^{-07}$	$9.12e^{-03}$	$1.13e^{-07}$	$1.26e^{-07}$	0.37
poI_i	$-2.24e^{-06}$	$3.01e^{-07}$	$1.05e^{-13}$	$-1.15e^{-06}$	$2.94e^{-07}$	$9.00e^{-05}$
poI_j	$-1.17e^{-06}$	$3.01e^{-07}$	$1.01e^{-04}$	$-2.10e^{-06}$	$2.95e^{-07}$	$1.04e^{-12}$
emp_i	$3.58e^{-02}$	$3.11e^{-03}$	$< 2e^{-16}$	$-6.20e^{-03}$	$2.52e^{-03}$	0.01
emp_j	$-3.15e^{-05}$	$2.66e^{-03}$	0.99	$3.66e^{-02}$	$2.59e^{-03}$	$< 2e^{-16}$
GDP_i	$-7.61e^{-13}$	$1.36e^{-13}$	$2.42e^{-08}$	$-5.50e^{-13}$	$1.36e^{-13}$	$4.93e^{-05}$
GDP_j	$-5.19e^{-13}$	$1.38e^{-13}$	$1.70e^{-04}$	$-6.69e^{-13}$	$1.27e^{-13}$	$1.32e^{-07}$
CPI_i	$-2.40e^{-01}$	$3.63e^{-02}$	$3.51e^{-11}$	$1.46e^{-01}$	$3.57e^{-02}$	$4.49e^{-05}$
CPI_j	$1.61e^{-01}$	$3.66e^{-02}$	$1.11e^{-05}$	$-2.31e^{-01}$	$3.47e^{-02}$	$2.62e^{-11}$
ad_i	$1.53e^{-05}$	$7.19e^{-07}$	$< 2e^{-16}$	$1.68e^{-05}$	$7.10e^{-07}$	$< 2e^{-16}$
ad_j	$1.67e^{-05}$	$7.30e^{-07}$	$< 2e^{-16}$	$1.47e^{-05}$	$7.05e^{-07}$	$< 2e^{-16}$
pwa_j	$-1.93e^{-02}$	$3.01e^{-03}$	$1.54e^{-10}$			
$Inftrdwell_i$	$5.25e^{-03}$	$1.80e^{-03}$	$3.59e^{-03}$	$-9.58e^{-03}$	$2.35e^{-03}$	$4.44e^{-05}$
$Inftrdwell_j$	$-6.74e^{-03}$	$2.45e^{-03}$	$5.91e^{-03}$			
ocr_j				$-4.13e^{-02}$	$5.79e^{-03}$	$8.96e^{-13}$
pwa_i				$-1.76e^{-02}$	$2.88e^{-03}$	$9.92e^{-10}$
degrees of freedom	2619			2620		
Deviance	2655.70			2523.00		
Scale	1.13	$2.77e^{-02}$		1.19	$2.92e^{-02}$	
Goodness-of-fit (GOF)			0.30			0.91
AIC	36804			36637		

4.4.15 Assessment of the models

This section compares the performance of these NB and Gamma models. The results in Table 4.18 show that the Gamma model has lower AIC value compared to the NB model, suggesting that the Gamma model performs better than the NB model in predicting migration at the district municipalities of South Africa. The $p - value$ of the deviance suggests that the Gamma model fits the data reasonably well compared to the NB model. The dispersion parameters for a Gamma models that predicts in-migration and out-migration was $\frac{1}{1.13}$ or 0.88, and $\frac{1}{1.19}$ or 0.84, respectively. The dispersion parameters were almost to one.

Table 4.18: Diagnostics of the NB and Gamma model

Model	AIC	Goodness-of-fit	Dispersion estimates	Theta
$NB_{Out-migration}$	36692	$2.68205e^{-07}$		1.18
$NB_{In-migration}$	36859	$1.078438e^{-07}$		1.13
$Gamma_{Out-migration}$	36637	0.91	0.84	
$Gamma_{In-migration}$	36804	0.30	0.88	

4.4.16 Diagnostic plot of the Gamma models

Figure 4.19 and 4.21 indicates that the residuals plot against the fitted values, these plots shows that there is no pattern or trend formed by the residual and this suggests that there is no violation of the independence in the model. Also these plots indicates that there is no need to transform the response variable. The plot of the deviance residuals against the transform values ($2 \ln \hat{y}_i$) in Figure 4.20 and 4.22 are centred around zero and the variance is constant. The plots also show the presence of the outliers in the model. In addition the Figure 4.20 and Figure 4.22 suggests that the Gamma distribution with a log link is adequate in modelling the migration data.

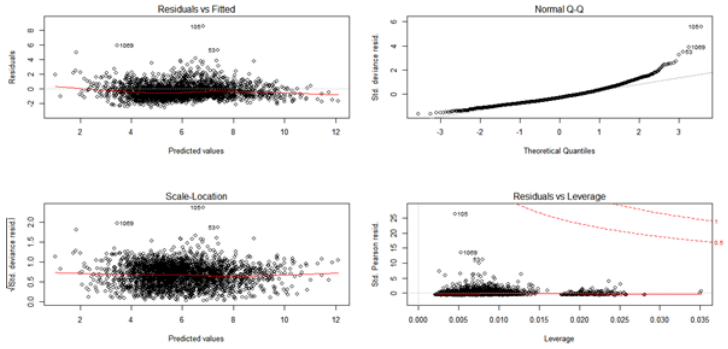

Figure 4.19: Gamma model (In-migration)

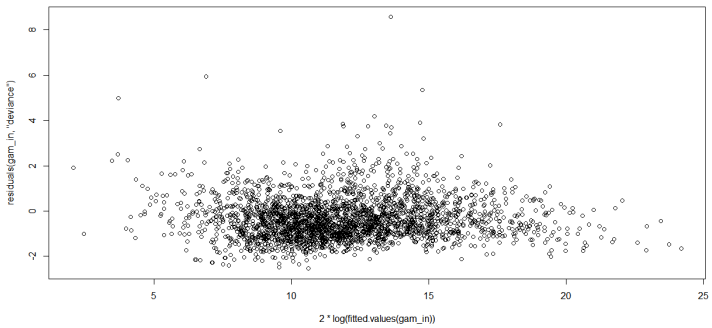

Figure 4.20: Plot of the deviance residuals against $2\ln(\hat{y}_i)$: In-migration

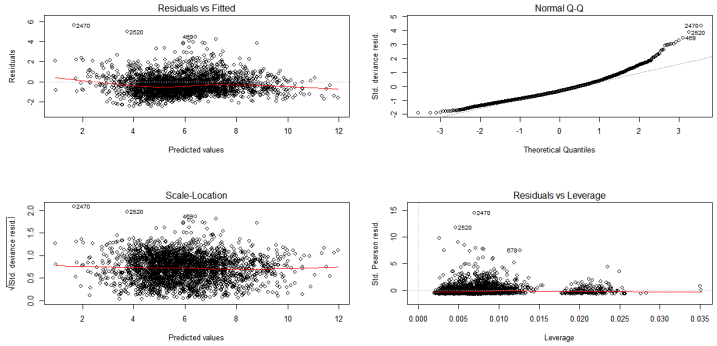

Figure 4.21: Gamma model (Out-migration)

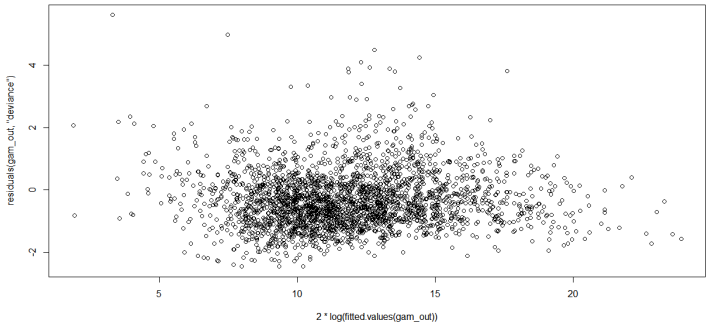

Figure 4.22: Plot of the deviance residuals against $2\ln(\hat{y}_i)$: Out-migration

4.4.17 Diagnostic plot: Ord plot

In this section the results of the Ord plot discussed in Chapter 2 are presented. From the Ord plot, Figure 4.23 the slope of the straight line was positive (1.95) and the intercept was negative (-20.985). These results suggest that the migration counts follows a logarithmic series distribution since the slope of the thicker line is positive and the intercept is negative (see section 2.5.8).

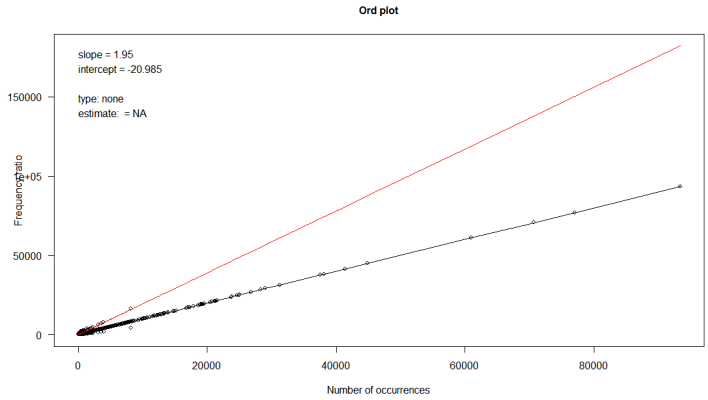

Figure 4.23: Ord plot

4.5 Modelling Net-Internal Migration

This section focus on the results of the OLS and GWR models that were used to study the relationship between net-internal migration in the 52 district municipalities of South Africa.

4.5.1 Results of the OLS model

Net-Internal Migration was modelled using OLS, the beta coefficients and their standard errors are captured. Only the significant predictors (at 5%) are shown in

Table 4.19. Multicollinearity was not detected as all the VIF's were less than 6. The overall fit (F test), suggests that the model was significant with $p - value <$ 0.05. The R^2 of the model was found to be 71%, this suggests that the model describe 71% of the observed variation in the net-internal migration. The variables, $nacstvs$ (Percentage of the households with no access to services, such as formal dwelling, sanitation, tap water inside dwelling, electricity for lighting and refuse removal), pow (White population) and, tre (Percentage of the households that are renting) are positively related to net-internal migration. The coefficient of the pod (Population density) variable was -63,82, this means that for each unit increase in population density (pod) at district municipality, the net-internal migration decreases by 63,82. The relationship of net-internal migration with $nacstvs$ suggests a link between net-internal migration in the district municipalities and percentages of the households with no access to services.

Table 4.19: Results of the Net-Internal Migration model (OLS)

Independent variables	Estimate	SE	p-values	VIF
intercept	$-1.46e^{05}$	$2.99e^{04}$	$1.22e^{-05}$	
pod	-63.8	18.3	$1.07e^{-03}$	4.34
$nacstvs$	$1.18e^{03}$	$3.95e^{02}$	$4.56e^{-03}$	3.53
pob	$-3.11e^{-02}$	$1.36e^{-02}$	0.03	4.36
pow	$4.48e^{-01}$	$6.72e^{-02}$	$2.91e^{-08}$	4.79
tre	$5.13e^{03}$	$1.07e^{03}$	$1.66e^{-05}$	3.68
$F_{statistic}$	26.25			
$p - value$	$1.958e^{-12}$			
AIC_c	1235.682			
R^2	0.71			

The error terms of the model shown in Figure 4.24 show no pattern with the fitted values, but some of the data points deviates away from the normality probability line, this indicates that some of the error terms were not normally distributed.

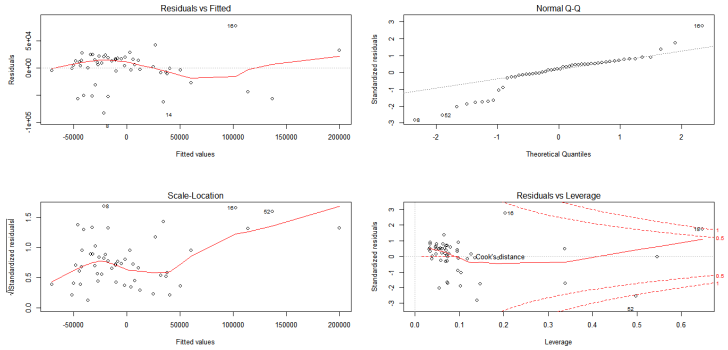

Figure 4.24: Diagnostic plot of the Net-Internal Migration model

In Table 4.20 the Bonferonni method, detected a significant outlier corresponding to observation number 8 from the data set and this observation was the net-internal migration equal to -104906 in Nelson Mandela Metropolitan Municipality (NMA).

Table 4.20: Results of the outlier identification

Outlier observation	rstudent	unadjusted p-value	Bonferonni p
8	-3.087078	0.0034551	0.17967

4.5.2 Assessment of the Net-Internal Migration model (OLS)

The JB and Koenker's Studentised Breusch-Pagan Test were performed and the results of the tests are shown in Table 4.21.

The BP test was not significant at the 5% level, telling us that there was no evidence to conclude that the beta parameters of the model were nonstationary. The JB Test was just significant at 5% meaning that the regression model may be biased. The results from this model cannot be trusted, as suggested by Montgomery et al (2006).

Table 4.21: Results of the Net-Internal Migration model assessment

$Model$	Studentized Breusch-Pagan test		Jarque Bera Test	
	BP	$p-value$	X^2	$p-value$
$Model_{Net-Internal\ Migration}$	10.77	0.05613	6.3527	0.04174

4.5.3 Testing the assumptions of the linear model

The global validation of the linear model assumptions, $gvlma$ in R was used on the OLS model. It tested the fit, the shape of the distribution of the residuals, (skewness and kurtosis), the linearity and the homoscedasticity. The result presented in Table 4.22 were at 5% level of significance. The general statistic indicates that the linear model, almost does not fit the data. The $gvlma$ suggests that the residuals of the model are significantly skewed. However, the kurtosis of the model does not differ from the normal distribution kurtosis. The linearity assumption of the model was accepted based on the link function. Also, there was no evidence to suggest that homoscedasticity was violated.

Table 4.22: Results of the model assumptions

	Value	p-value	Decision
Global Stat	9.4964	0.04982	NOT satisfied!
Skewness	4.8061	0.02836	Assumptions NOT satisfied!
Kurtosis	1.5466	0.21364	Assumptions acceptable.
Link Function	0.1309	0.71754	Assumptions acceptable.
Heteroscedasticity	3.0129	0.08261	Assumptions acceptable.

The D-W test was used to test for independence of the error terms. The results in Table 4.23, show a D-W value of 2.277278 or d =2.277278 and that the autocorrelation is equal to -0.1784677 (ρ =-0.1784677). To check the significance of the negative autocorrelation at 5%, the formal the test is expressed as,

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

Based on the $D - W$ test in Table 4.23, this study fails to reject H_0 and conclude that the errors are not correlated.

Table 4.23: Testing the significance of autocorrelation

Models	lag	$Autocorrelation$	$D - WStatistic$	$p - value$
$Model_{net-internal\,migration}$	1	-0.1784677	2.277278	0.47

Table 4.24 and 4.25 shows the results without the outlier. The BP test was significant at 5%, this means that there was evidence to suggest that the beta parameters of the model vary across the space, and the JB Test was significant. The D-W test indicates that the error terms are not correlated.

Table 4.24: Outlier removed: results of the Net-Internal Migration model assessment

$Model$	Studentised Breusch-Pagan test		Jarque Bera Test	
	BP	$p - value$	X^2	$p - value$
$Model_{net-internal\,migration}$	14.1025	0.01497	6.8861	0.03197

Table 4.25: Outlier removed: Testing the significance of autocorrelation

Models	lag	$Autocorrelation$	$D - WStatistic$	$p - value$
$Model_{net-internal\,migration}$	1	-0.1828431	2.231612	0.534

4.5.4 Results for Geographically Weighted Regression (GWR) model

This section presents the results of the GWR that was used to model net-internal migration in district municipalities. Since the The Koenker-BP or Koenker's Studentised Breusch-Pagan test suggested that there was evidence to suggest that the

beta parameters were nonstationary, since the $p - value$ was less than to 5% after an outlier was removed. The GWR model was fitted.

4.5.5 Checking multicollinearity

In linear models the results become biased when two or more variables indicates the presence of multicollinearity. The GWR is a local model, it builds local regressions for each feature in a data set. When the observations for a particular explanatory variable cluster spatially, that signals the presence of multicollinearity. The condition number was used to check for the presence of multicollinearity. Figure 4.25 indicates that in all district municipalities and metros of South Africa, there was no evidence to suggest that there was multicollinearity, because the results suggest that the condition number was above zero and less than 30 in all district municipalities and metros.

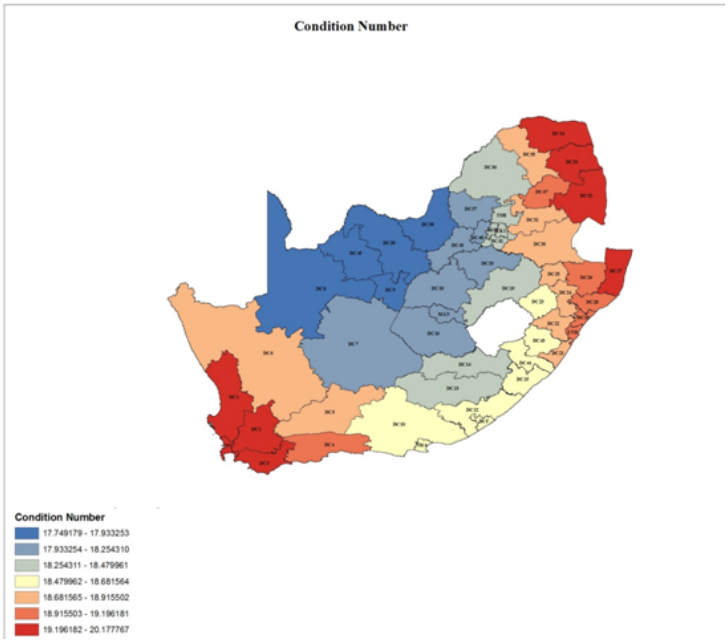

Figure 4.25: Investigating possible spatial multicollinearity

4.5.6 Testing for the misclassification of the GWR model

The spatial distribution of the residuals of the GWR was investigated using Moran's I, and the results of the test are shown in the Spatial results in Figure 4.26. The Moran's I, $z-score$ and $p-value$ in Table 4.26. The $z-score$ of the Moran's I was -1.09, this means that the pattern of the residuals does not appear to be significantly different from a random process, and the z-score was not statistically significant, $p-value$ =0.335980 $>$ 0.05 (5%), this means that the GWR model was well specified.

Table 4.26: Global Moran's I Summary

$Moran's\ Index$	-0.094436
$Expected\ Index$	-0.019608
$Variance$	0.004714
$z-score$	-1.089908
$p-value$	0.275754

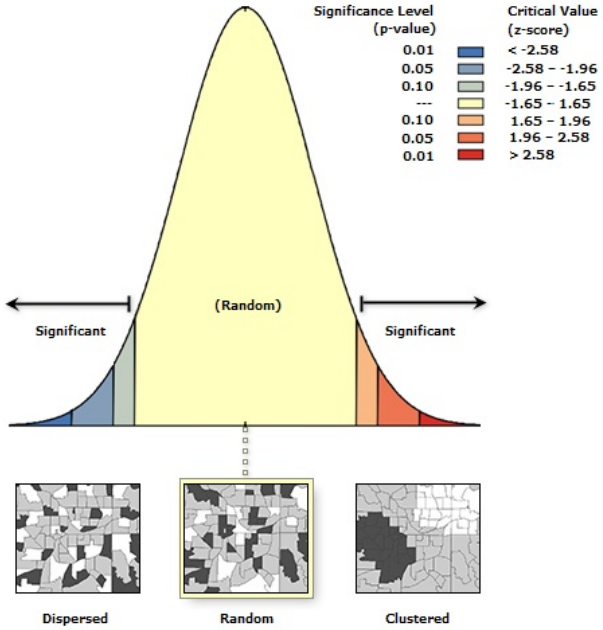

Figure 4.26: Spatial Autocorrelation results

4.5.7 Diagnostics statistics for GWR

Figure 4.27, indicates that there is not much variation in the values of the local is R^2 (R^2_{adj}) statistics. The R^2 were lower in the Western Cape district municipalities in the ranges of 66% and 68%, this indicates that the model predicts poorly in those districts. The R^2 were higher in the district municipalities of Limpopo, the local R^2 values were above 80%, this indicates that the GWR model predicts well in those district municipalities.

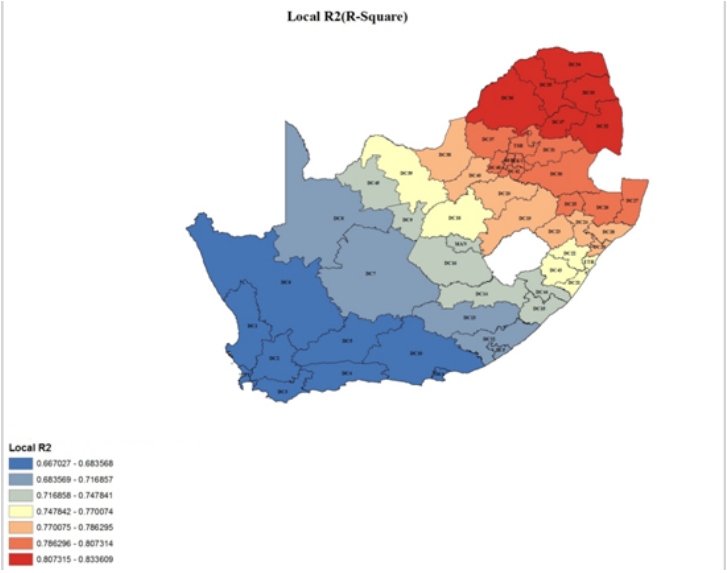

Figure 4.27: Local R^2

Table 4.27 shows the values of the diagnostic measures, Bandwidth, Residual Squares, Effective number, Sigma [1], AIC_c, R^2. The AIC_c value of the GWR was

[1]Effective Number reflects a trade off between the variance of the fitted values and the bias in the coefficient estimates and is said to be related to the bandwidth.

Sigma this value is calculated as the square of the normalised residual sum of squares, where the sum of squares is divided by the effective degree of freedom of residuals.

1228.86, and it lower than 1235.682 the AIC_c value of the OLS model, and R^2 of the GWR model is 0.76 and it was larger than 0.71, the R^2 of the OLS model. That means the GWR model is a better model than the OLS.

Table 4.27: Diagnostic results of the GWR

Statistic	Values
Bandwidth	8.59
Residual Squares	35478709047.40
Effective Number	9.86
Sigma	29016.79
AIC_c	1228.86
R^2	0.76

The Analysis of Variance $(ANOVA)$ results in Table 4.28 compares the global model to the GWR model. The ANOVA tests the null hypothesis that the GWR model represents no improvement over a global model. This study observed a reduction in the residual sum of squares when the GWR was used. It can be seen from the F test the $p - value < 5\%$. According to Brundson, Fotheringham and Charlton (1999) the results in Table 4.28 suggests that the GWR model is a significant improvement on the global model for predicting net-internal migration in the district municipalities of South Africa. [2]

Table 4.28: ANOVA results

Source	SS	DF	MS	F	p-value
OLS Residuals	46295000000	6			
GWR Improvement	10816290952.6	3.86	2802147915.18		
GWR Residuals	35478709047.40	42.14	841924751.96	3.33	0.02

[2]The degrees of freedom (df) 3.86 and 42.14 in Table 4.28 are not expected to be integers. Brundson et al (1999) said the F- distribution is well explained by any non-negative *df* parameter and they also refer to those two quantities as effective degrees of freedom (df).

4.5.8 Results of the Monte Carlo Significance Test

The results of the randomisation tests on each β coefficients is given in Table 4.29, at the 5% level of significance. In most cases standard errors generated by the OLS model exceeds $\sqrt{v}_j$, only the beta parameter for the pow variable seems to be non stationary. This suggests that the coefficient for the population size of the white population (pow) vary in space. The Monte Carlo test should be equal or less than 5% for a parameter to exhibit a significant spatial variation, and the results from the test indicates that the beta parameter for population size for the white population (pow) was the only variable with a parameter that vary significantly in space.

These are useful results from the test because in the case of mapping the local estimates this study focuses only on one variable (pow) for which the local estimates are significantly non-stationary.

Table 4.29: Results for the spatial variability of coefficients

$Variables$	$\sqrt{v_j}$	SE	Monte Carlo significance test ($p - value$)
intercept	7533.110	29850	0.73
pod	6.900	18.28	0.53
$nacstvs$	104.663	394.7	0.66
pob	0.012	0.01356	0.05
pow	0.080	0.06719	0.01
tre	374.625	1067	0.54

4.5.9 Local Regression coefficient (pow)

Figure 4.28 indicates that, through out the intervals the coefficient of the white population was positive, and this shows a positive relationship between the white population and the net-internal migration in the district municipalities of South Africa. The beta parameters are larger in the district municipalities of Limpopo,

this means that a unit increase in the population size of the white population increase the net-internal migration.

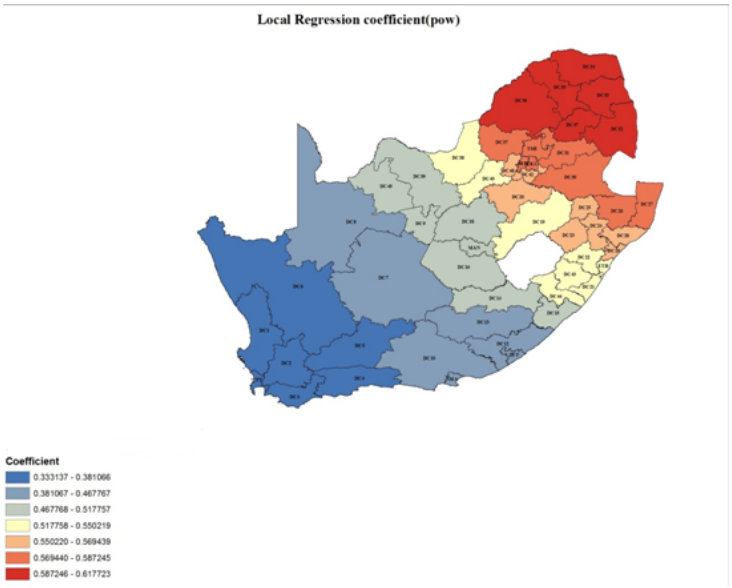

Figure 4.28: GWR White population (pow) Coefficient

4.5.10 Cluster Analysis (CA) results

In this section cluster analysis results are presented. The data set consists of the explanatory variables in Table 4.19 which were standardised before estimating the number of clusters. The average silhouette width was used to estimate the number of clusters formed by the data set.

The silhouette width method estimated 2 clusters from the data set (excluding the net-internal migration). Figure 4.29 and Figure 4.30, show that, cluster 1 was the largest cluster, with 47 observations, and cluster 2 with 5 observations. The average silhouette width was 0.66. According to Watts, Toth, Murphy and Lovas

(2001) this value means that a reasonable structure has been identified, see Figure 4.29. The k-means method was used to form the clusters. The results from the cluster means in Table 4.30 indicates that cluster 2 formed by EKU, JHB, TSH, ETH and CPT, and the names of the district municipalities that form clusters are shown in a R code in Appendix E. This cluster has positive net-internal migration. Furthermore cluster 2 has the highest, Gross Domestic Product, Average household income, population density, percentage of households that were renting dwellings, and the lowest illiteracy rate. This cluster represents developed metropolitan municipalities. Cluster 1 mostly consists of district municipalities and this cluster has negative net-internal migration. This cluster has the highest, unemployment rate, illiteracy rate and percentage of the households with no access to services. Cluster 1 represents underdeveloped district municipalities.

Table 4.30: Cluster means

Cluster	pod	$nacstvs$	pob	pow	tre	$Net-InternalMigration$	ump	avh	GDP	$illitrt$
1	-0.28	0.11	-0.25	-0.3	-0.18	-10562.09	32.51	72120.62	296325425532	25.27
2	2.67	-1.05	2.38	2.77	1.69	99283.60	26.42	153274.40	780400000000	10.26

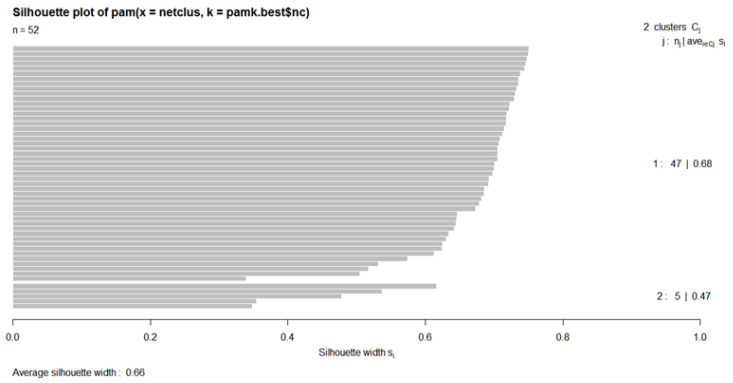

Figure 4.29: Silhouette width

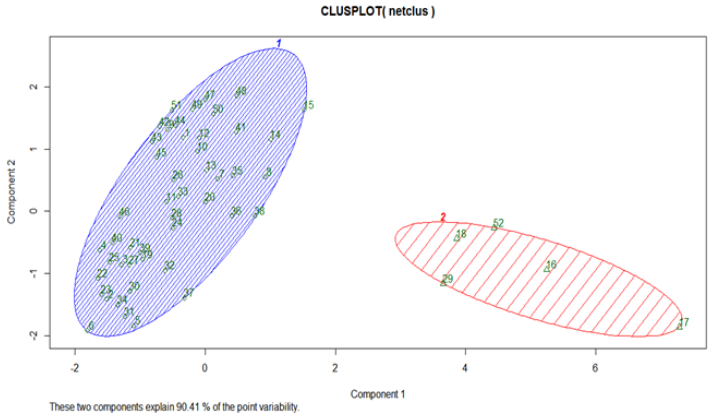

Figure 4.30: Cluster plot

4.5.11 Summary

The analysis indicates that Tshwane, Ekurhuleni and City of Cape Town have the highest net-internal migration. While, Nelson Mandela Metropolitan, O.R Tambo, and Capricorn district municipalities, they have the highest negative net-internal migration.

This chapter presented the results of the following models, OLS, Poisson, NB, Gamma and GWR models. The results from the models, suggest that the economic, demographics, tenure status were significant push and pull factors in the district municipalities of South Africa. The Poisson and NB model failed to fit the data while Gamma model fitted the migration data. The OLS model failed to satisfy some of the assumptions of the linear model and that suggest that the predictions of this model may not be reliable, especially in the metropolitan municipalities. The GWR model performed well in predicting the net-internal migration in the district municipalities of South Africa compared to the OLS model.

The results from the GWR model indicates that only the parameter for the white population vary in space. The results of the GWR suggest that we cannot assume that the beta parameter is stationary for the white population size when modelling the net-internal migration at the district municipalities of South Africa.

The results from the cluster analysis revealed that there are two clusters estimated by silhoutte width method. The k-means is used to form clusters. The results suggest that cluster 1 was the largest cluster with 47 observations and cluster 2 with 5 observations. Cluster 1 has negative net-internal migration. Cluster 2 has positive net-internal migration and this cluster that is largely represented by metros (Ekurhuleni, Johannesburg, Tshwane, EThekwini and City of Cape Town) is more developed than cluster 1.

Chapter 5

Conclusion and Recommendation

5.1 Introduction

This study demonstrates the use of the statistical models to model migration data using 2011 Census data, which was conducted by Statistics South Africa (Stats SA). The models that were used for this study are the Gravity, Extended Gravity, Poisson, Negative Binomial and Geographically Weighted regression models. The Gravity, Extended Gravity, Poisson, and NB are known as the global models. The GWR model is known as a local model. This chapter presents the summary and discussion, recommendation and conclusion of the study.

5.2 Summary and discussion

These models are widely used in many studies as already indicated in the literature review. This study suggests that push and pull factors, as described by eco-

nomic, demographics, living conditions variables were important in explaining in-and out-migration in the district municipalities of South Africa. Furthermore the significance of these variables shows the economic and social disparities in the district municipalities. These observation agrees with the findings of other researchers as discussed in the literature review.

The reset test confirmed that the linear relationship of the Extended Gravity regression that was used by Bouare (2000-2001), and Fan (2005) to model migration in the district municipalities was inadequate. It was observed that nonlinear terms such as, the quadratic terms were found significant.

Metros like Tshwane, Ekurhuleni, and City of Cape Town had the highest net-internal migration. The results from the cluster analysis in Table 4.30 show that these metros are from the developed provinces with the highest GDP values, lowest illiteracy rate and excellent employment opportunities in South Africa. This observation seems to support the claim that, migrants, migrate to places where there are high levels of employment rates or low levels of unemployment rates as well as higher GDP values (Congdon, 1992; Faggian and Royuela, 2010).

It was observed that many migrants from Capricorn are migrating to Tshwane. According to 2011 Census, close to (84.9%) 85% of the population speak Sepedi as their first language, and close (19.9%) 20% of the population in Tshwane speak Sepedi as their first language, this makes Sepedi a popular official language in Tshwane. This observation could explains why the migrants from Capricorn chose to migrate to Tshwane. Flowerdew and Amrhein (1989), motivates the necessity of including the language variable in a model, by saying language could be linked with cultural reasons or social networks, and migration was more likely

to take places where the same language is spoken. Large social networks in an area contribute to higher migration rates. The findings of this research show that language is an important variable, in deciding which area to migrate to. This can be seen as the rural to urban migration, because, the population from Capricorn largely resides in rural areas, its municipalities are mostly rural municipalities, while Tshwane is 100 % urban. The inequalities between these municipalities were high, for instance, from 2011 Census, the illiteracy rate, in Capricorn and Tshwane, was 20.8% and 10.1%, respectively it is more than twice the illiteracy rate of Tshwane. The average annual household incomes, in Tshwane and Capricorn, were R182867.00 and R69233, respectively. These results, suggest there are inequalities, between rural and urban district municipalities. Although these variables, illiteracy rate, average annual household income and other variables were not significant in explaining in-and out-migration. Faggian and Royuela (2010) noted that, variables with non-significant coefficients do not mean that they play no role in migration, they explained that, the territorial distribution of variables might be relatively homogeneous over territory and this observation reduces the significance of these variables as factors for moving decisions. Furthermore, in this study only the provincial GDP and CPI (for some district municipalities) were used. Since this information was not available at the district level.

From paragraph (4) above, it was observed that, the migrants from Capricorn, O.R Tambo district municipality, and Nelson Mandela Metropolitan, tend to migrate to areas close to their places of origin see, Figure 4.16, 4.14 and 4.15. However migrants proceeding long distances generally target large commercial and industrial centers (Ravenstein,1885), Johannesburg migrants join the list of the municipalities with the highest migrants in the City of Cape Town. It is clear from the map,

distance wise, none of these municipalities are close to each other, this could be the influence of the employment (emp) opportunities which act as pull and push factors. The variable distance plays a very critical role in predicting movements within the district municipalities of South Africa, this variable was significant in all the models. Also the variable distance is associated with cost of movement.

The distance, between the population in the origin (pop_i) and population in the destination have the expected signs in predicting in and out migration from the results of the Gravity model, similar findings were found by (Fan, 2005). The Nonlinear model did well in terms of explaining the variances accounted for by both in-and out-migration in the district municipalities of South Africa, this was confirmed by the higher R^2 ($>$75%), the model were the best among the other model that were observed from the lower AIC. The results from the Nonlinear models suggests that the distance effect increases in-migration as one moves to around 5.05 km, then after it is expected to decrease. Furthermore, the employment rate effect at the destination decreases out-migration at 16.44% and after that in-migration increases.

Although the Nonlinear models did well in explain the variation of the response variable, the models failed to satisfy the assumptions of a linear model and also the models were found to be biased, by the Jarque-Bera Test (JB-Test), this suggests that the residuals of the OLS models were not normally distributed.

This study showed that there were outliers in the best performing models (model with quadratic terms) that were fitted for in- and out-migration, these outliers affected diagnostic measures, R^2 and AIC, i.e, after removing the outliers, R^2 increased and AIC decreased. There was an outlier from the OLS model that was

used to model net-internal migration in the district municipalities of South Africa.

The outlier (-104906) came from Nelson Mandela Metropolitan, this indicates that NMA lost 104906 through out-migration. This was unusual, because NMA is coastal area and coastal areas such as Port Elizabeth Waterfronts and NMB South coast are suitable for tourism. Community services, trade and manufacturing sectors are the sectors that create the most employment opportunities in NMA (Spies, 2013). However, from 2001 Census to 2011 Census, in NMA unemployment rate increased from 28.2% to 36.6%, an increased in unemployment rate could be the one of the factors that push the population out of this metro.

The application of the OLS in modelling the count response variable is discouraged by many researchers as indicated in the literature review, count model like, Poisson, NB and Gamma model are preferable. The study used the Poisson, NB and Gamma model. The results from the Poisson model indicated that the explanatory variables were highly significant, this study proceeded to test the assumptions of the model. From the dispersion test, the standard errors of the Poisson model cannot be trusted, because the model was overdispersed and the it failed to fit the data. Since the Poisson model was overdispersed, the NB model was used. The NB model accounts for dispersion and the Poisson model failed to capture, but in this study NB model did not fit the data.

Figure 4.1, showed that the distribution of the migration numbers was left skewed, this suggested, the use of the Gamma distribution to model the data. The results of the goodness-of-fit, suggest that the model was significant at 5% level.

The net-internal migration was explored in two steps. In the first part the OLS

model (global model) was used, and the second part we used the GWR model also known as a local model. The study used predictors, such percentage of the households that are renting (tre), white population (pow), population density (pod), black population (pob), and no access to services ($nacstvs$) were used in these models. These predictors were all significant at 5% with no presence of multicollinearity (all the VIF' s for the predictors were less than 6). The OLS model was significant, and with the good R^2 (71.4%). However the model failed to satisfy the assumptions of the linear model and also the model was proven to be biased by the JB-Test. The Koenker-BP Test was not significant at 5%, this implied there was no evidence to model the data using GWR, but this study proceeded in using the GWR model.

The diagnostics of the GWR model such as AIC_c and R^2 suggest that GWR model performed better than the OLS in predicting net-internal migration in the district municipalities of South Africa. The Moran's I test on the residuals of the GWR model, showed a random spatial pattern and these results suggested that the GWR model is well specified. The GWR model predict poor in the district municipalities of Western Cape, because lower $R^2 \approx 67\%$ was observed. However the model predict well in the district municipalities of Limpopo, with higher $R^2(> 80\%)$.

The study tested the stationarity of each of the beta parameters in a GWR by using a method describe by (Brunsdon, et al, 1998). The results from the test at 5% level suggested that only the beta parameters for white population (pow) is non stationary over space. This was surprising because the Koenker-BP Test was not significant at 5%, the test suggest that there was no evidence of non-stationarity

of the estimated parameters or the relationship does not vary in the geographical space, this could be the weakness of the global test, in a sense that this test does not identify the individual parameters that are non-stationary or stationary.

5.3 Conclusion

The results of this study confirmed that is not advisable to use the gravity approach in modelling the migration flow (count response). Researchers are encourage to use models like Poisson, NB, Gamma, etc. These models are designed to model count response variables. In cases where there is overdispersion, the Poisson model may not be the best alternative model, while the NB that accounts for overdispersion is the recommended model. The goodness-of-fit results however, show that the NB failed to model migration. It is suspected that this model fails to account for metros which tend to have greater migration figures. The Gamma model on the other hand passed the goodness-of-fit and it explains migration well.

The OLS and GWR were also used to model the net internal migration at the district level. The results shows that the GWR is superior to OLS in predicting the net-internal migration at the district level. The Monte Carlo significance tests show that the parameter estimates of the size of the white population is not the same across the district municipalities of South Africa. We conclude that the relationship between net internal migration and size of the white population is different across the district municipalities of South Africa.

The results from the models, suggest that, the tenure status, such as households that owned but not yet paid off (obp), owned and fully paid off (ofp), occupied rent free (ocr) and were renting dwellings (tre), economic variables such as Gross

Domestic Product (GDP), and Consumer Price Index (CPI), living conditions such as access to tap water (pwa), and no access to services, and the demographics, race, population density (pod), and adult population (ad), were important driving forces of internal migration in South Africa. These findings suggest that economics, demographics and living conditions are the major drivers for individuals to migrate from one district municipality to another. From a policy perspective, we believe that the results of this study hint a useful information on what factors influence internal migration and ultimately the population distribution of individuals in the district municipalities of South Africa.

5.4 Recommendation

This study, recommends that researchers should properly access the fit of models. They need to check for the model assumptions and for transparency the results must be reported. They are many publicised models that are fitted with higher R^2, but nothing is said about whether those models are biased or not. The results of the OLS (Gravity model), cannot be trusted because the model assumptions were violated, such as, normality. It is important to use distributions that were designed to model discrete observations.

This study revealed that, push and pull factors such as, employment, GDP, Access to water, tenure status, type of dwelling (Informal or traditional dwelling), distance between the districts, demographic variables were found to be related in explaining internal migration in South Africa.

The push and pull factors in this study hint the presence of inequalities between the district municipalities of South Africa, and these inequalities need to be studied

and outlined. Future studies should investigate such variables as proximity to malls, access to newer public transport like Rea Vaya and Gautrain etc. Instead of using proxy variables future studies should use the correct municipality level predictors to reveal hidden dynamics these variables, and those variables may play a huge role in migration studies.

This study models migration at the district level and, Stats SA have the migration figures (Census 2011) up to Local municipalities, it is recommended that future researchers explore migration patterns and modelling at this level. We suggest the use of the offset variable in count models. The results from the Ord plot in Figure 4.23 recommend the use of the logarithmic series to model the migration data. Also, it is important to test whether the explanatory variables vary over space, if that is the case, then it is recommended that models like GWR, Geographically Weighted Poisson (GWP), etc, that account for variation in space should be investigated with other predictors.

Reference

Abdel-Aty, M. A. and Rwadan, A. E., (2000). Modelling traffic accident occurrence and involvement. *Accident Analysis and Prevention* 32, pp. 633-642.

Achim, Z . and Torsten, H., (2002). *Diagnostic Checking in Regression Relationships*. R News 2(3), 7-10. URL http://CRAN.R-project.org/doc/Rnews/.

Bouare, O., (2000-2001). Determinants of Internal Migration in South Africa. *SA Journal of Demography*, 8(1), pp. 23-28.

Bowman, K.O. and Shenton, L.R., (1975). Omnibus Test Contours for Departures from Normality based $\sqrt{b_1}$ and b_2. *Biometrika*, Vol. 62, pp. 243-250

Boyle, P., (1995). Rural in-migration in England and Wales 1980-1981. *Journal of Rural Studies*, 11(1), pp. 65-78.

Breusch, T.S. and Pagan. A.R., (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*. Vol. 47, No. 5, pp.1287-1294.

Brunsdon, C., Fotheringham, AS. and Charlton, M., (1996). Geographically Weighted Regression for Exploring Spatial Nonstationery. *Geographical*

Analysis, 28(4), pp. 281-298.

Brunsdon, C., Fotheringham, AS. and Charlton, ME., (1998). Geographically weighted regression - modelling spatial non-stationarity. *Journal of the Royal Statistical Society*, Series D-The Statistician 47(3):431-443.

Brunsdon, C., Fotheringham, AS. and Charlton, ME., (1999). SOME NOTES ON PARAMETRIC SIGNIFICANCE TESTS FOR GEOGRAPHICALLY WEIGHTED REGRESSION. *JOURNAL OF REGIONAL SCIENCE*, VOL. 39, NO. 3, pp. 497-524.

Burgin, T.A., (1975). The Gamma Distribution and Inventory Control. *Operational Research Quarterly* (1970-1977), Vol. 26, No. 3, Part 1, pp. 507-525.

Byrne, G. and Pezic, A., (2004). *Modelling internal Migration drivers with Geographically Weighted Regression*. Canberra, Australia, Australian Population Association.

Cahill, M. and Mulligan, G., (2007). Using Geographically Weighted Regression to Explore Local Crime Patterns. *Social Science Computer Review*, 25(2), pp. 174-193.

Cameron, A. C. and Trivedi, P. K., (1986). Econometric Models based on Count data:Comparisons and Applications of some estimators and tests. *Journal of Applied Econometrics*, Volume 1, pp. 29-53.

Cebula, R. J. and Alexander, G. M., (2006). Determinants of Net-Interstate Migration, 2000-2004. *The Journal of Regional Analysis and Policy*, 36(2), pp. 116-123.

Christian, K. and Achim, Z., (2008). *Applied Econometrics with R*. New

York: Springer-Verlag. ISBN 978-0-387-77316-2. URL http://CRAN.R-project.org/package=AER.

Collett, D., (2003). *Modelling binary data*. 2nd ed. Chapman & Hall/CRC Text in Statistical Science Series.

Congdon, P., (1992). Aspects of General Linear Modelling of Migration. *Journal of The Royal Statistical Society*, Series D (The Statistician), 41(2), pp. 133-153.

Congdon, P., (2010). Random-effects models for migration attractivity and relentivity: a Bayesian methodology. *Journal of the Royal Statistical Society*, 173(4), pp. 755-774.

Coxe, S., West, S. G. and Aiken, L. S., (2009). The Analysis of Count data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment*, 91(2), pp. 121-136.

Czaika, M. and H. de Haas ., (2013). Determinants of Migration to the UK. *Migration Observatory briefing*, COMPAS, University of Oxford, UK, January 2013.

Elhai, J. D., Calhoun, P. S. and Ford, J. D., (2008). Statistical procedures for analysing mental health services data. *Psychiatry research* 160, pp. 129-136.

Esri., (2013). *ArcGIS for Desktop: Release 10.2*. CA: Environmental Systems Research Institute.

Faggian, A. and Royuela, V., (2010). Migration Flows and Quality of Life in a Metropolitan Area: The case of Barcelona-Spain. *Application Research*

Qualities Life, Volume 5, pp. 241-259.

Fan, C. C., (2005). Modelling Interprovincial Migration in China, 1985-2000. *Eurasian Geography and Economics*, 46(3), pp. 165-184.

Flowerdew, R. and Aitikin, M., (1982). A method of fitting the Gravity Model based on the Poisson distribution. *Journal of Regional Science*, 22(2), pp. 191-202.

Flowerdew, R. and Amrhein, C., (1989). Poisson Regression Models of Canadian Census Division Migration flows. *Papers of The Regional Science Association*, Volume 67, pp. 89-102.

Fortin, M.T. and Legendre, F., (1989). Spatial Pattern and Ecological Analysis. *Vegetatio*, Vol. 80, No. 2, pp. 107-138.

Fotheringham, A., Brunsdon, C. & Charlton, M., (2000). *Quantitative Geography*. SAGE Publications.

Gale, N., Hubert, L. J., Tobler, W. R. and Golledge, R.G., (1983). Combinatorial Procedures for The Analysis of Alternative Models: An Example from Interregional Migration. *Papers of The Regional Science Association*, Volume 53, pp. 105-115.

Ganzach, Y., (1997). Misleading Interaction and Curvilinear Terms. *Psychological Methods*, Volume 2, No. 3, pp. 235-247.

Gardner, W., Mulvey, E. P. and Shaw, E.C., (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and Negative Binomial models. *Psychological Bulletin*, 118, 392-404.

Geedipally, S.R., Lord, D. and Dhavala, S.S., (2012). The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis and Prevention* 45. pp. 258-265.

Google Maps., (2014), [online], available: https://www.google.co.za/maps/dir///@-25.9549148,27.8224344,10z, [13 February 2014].

Greene, W.H., (1990). A gamma-distributed Stoichastic Frontier Model. *Journal of Economics* 46. pp. 141-163. North-Holland.

Henry, S., Boyle, P. and Lambin, E. F., (2003). Modelling Inter-provincial migration in Burkina Faso, West Africa: The role of socio-demographic and environmental factors. *Applied Geography* 23, pp. 115-136.

Hilbe, J. M., (2011). *Negative binomial regression*. 2nd ed. University Press, Cambridge.

Hurvich, C.M. and Tsai, C., (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 2, pp. 297-307.

Islam, R. and Siddiqi, N. A., (2010). Determinants and Modelling of Male Migrants in Bangladesh. *Current Research Journal of Economic Theory*, 2(3), pp. 123-130.

Juarez, J. P., (2000). Analysis of Interregional Labor Migration in Spain Using Gross Flow. *Journal of Regional Science*, 40(2), pp. 377-399.

Kalogirou, S. and Hatzichristos, T., (2007). A Spatial Modelling Framework for Income Estimation. *Spatial Economic Analysis*, 2(3), pp. 297-316.

Koenker, R., (1981). A note on studentising a test for heteroscedasticity. *Journal*

of Economics 17, 107-112.

Kok, P. and Collinson, M., (2006). *Migration an Urbanisation in South Africa*. Report no. 03-04-02, Pretoria: Statistics South Africa.

Lee, C. K., Lee, Y. K., Bernhard, B. J. and Yoon, Y. S., (2006). Segmenting casino gamblers by motivation: A cluster analysis of Korean gamblers. *Tourism Management* 27, pp. 856-866.

Lord, D., (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention* 38, 751-766.

Lu, B., Harris, P., Charlton, M., Brunsdon, C., Nakaya, T. and Gollini, I., (2014). *GWmodel:Geographically weighted models*. R package version 1.2-3. http://CRAN. R-project.org/package=GWmodel.

Lukasz, K., (2011). *outliers: Tests for outliers*. R package version 0.14. http://CRAN. R-project.org/package=outliers.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K., (2014). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.15.2.

Marc J. M., (2014). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. R package version 2.00. http://CRAN.R-project.org/ package=AICcmodavg.

Matignon, R., (2007). *Data Mining Using SAS Enterprise Miner*. John Wiley and Sons, inc.

Maza, A. and Villaverde, J., (2004). Interregional in Spain: A Semiparametric Analysis. *The Review of Regional Studies*, 34(2), pp. 156-171.

Miaou, S.P., (1994). The relationship between Truck Accidents And Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention*, 26(4), pp. 471-482.

Milevsky, M.A. and Posner, S.E., (1998). Asian Options, the Sum of Lognormals, and the Reciprocal Gamma Distribution. *The Journal of Financial and Qualitative Analysis*. Vol. 33, No. 3, pp. 409-422.

Millington, J., (2000). Migration and Age: The Effect of Age on Sensitivity to Migration Stimuli. *Regional Studies*, 34:6, pp. 521-533.

Mitsa, T., (2010). *Temporal Data Mining*. Chapman and Hall/CRC Data Mining and Knowledge discovery series.

Montgomery, D.C, Peck, E.A. and Vining, G.G., (2006). *Introduction to linear regression analysis*. 4th Edition. Wiley and Sons, Inc., Hoboken, New Jersey.

Mulhern, A. and Watson, J.G., (2009). Spanish Internal Migration: Is there Anything New to Say?, *Spatial Economic Analysis*, 4:1, 103-120.

Mutua, F.R., (1994). The use of the Akaike Information Criterion in the identification of an optimum flood frequency model. *Hydrological Sciences, Journal- des Sciences Hydrologiques*, 39,3.

Nabi, N., (1992). Dynamics of Internal Migration in Bangladesh. *Canadian Studies in Population*, 19(1), pp. 81-98.

Nakaya, T., (2001). Local spatial interaction modelling based on the geographically weighted regression approach. *GeoJournal* 53:, pp. 347-358.

O'Hara, R. and Kotze, D., (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, Volume 1, pp. 118-122.

Ord, J.K., (1967). Methods for a class of discrete distributions. *Journal of the Royal Statistical Society*. Series A (General), Vol. 130. No. 2. pp. 232-238.

Pebesma, E.J. and Bivand R.S., (2005). *Classes and methods for spatial data in R*. R News 5 (2), http://cran.r-project.org/doc/Rnews/. http://CRAN.R-project.org/package=sp.

Pena, EA. and State, EH., (2006). Global validation of linear model assumptions(gvlma). *J.Amer. Statist. Assoc*. 101 (473):341-354.

Pezic, A., (2009). Modelling Regional Migration. *Bulletin of the Australian Mathematical Society*, 80(2), pp. 173-176.

R Core Team., (2014). *R: A language and environment for statistical computing* . R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ravenstein, E.G., (1885). The Laws of Migration. *Journal of the Statistical Society of London*, Vol. 48, pp. 167-235.

Reynolds, R. J. and Fenster, C. B., (2008). Point and Interval Estimation of Pollinator Importance: A Study Using Pollination Data of Silence caroliniana. *Oecologia*, 156(2), pp. 325-332.

Roger, S. B, Edzer, P. and Virgilio, G.R., (2013). *Applied spatial data anal-*

ysis with R, Second edition. Springer, NY. http://www.asdar-book.org/. http://CRAN.R-project.org/package=sp.

Rogers, A., (1967). A regression Analysis of Interregional Migration in California. *The Review of Economics and Statistics*, 49(2), pp. 262-267.

Rosenshein, L., Scott., L . and Pratt, M., (2011). *Finding a Meaningful Model*. ArcUser Winter 2011, Esri.

Rousseeuw, P. J., (1987). Silhouttes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. North-Holland.

Santos, S. and Tenreyro, S., (2006). The Log of Gravity. *The Review of Economics and Statistics*, 88(4), pp. 641-658.

Shapiro, S.S. and Wilk. M.B., (1965). An analysis of variance test of normality (Complete Samples). *Biometrika*. Vol. 52. No. 3 and 4, pp . 591-611.

Silvestre, J., (2005). Internal migration in Spain, 1877-1930. *European Review of Economic History* 9, pp. 233-265.

Simonoff, J. S., (2003). *Analyzing Categorical data*. Springer Texts in Statistics, New York: Springer.

Singh, S.K, Singh, U. and Kumar, D., (2011). *Bayesian Estimation Of The Exponentiated Gamma Parameter and Reliability Function under Asymmetric Loss Function*. REVSTAT-

Spies,D., (2013). Nelson Mandela Bay. Business Guide. *Economic Overview*. Business Chamber.

Statistics South Africa, 2012(a). *Census 2011 Metadata*. Report no. 03-01-47. Pretoria: Statistics South Africa.

Statistics South Africa, 2012(b). *How the count was done*. Report no. 03-01-45. Pretoria: Statistics South Africa.

Statistics South Africa, 2012(c). *Gross Domestic Product, Annual and Regional estimates 2002-2011*. Statistical release no. P0441. Pretoria: Statistics South Africa.

Thadewald,T. and Buning. H., (2004). Jarque-Bera test and its competitors for testing normality: A power comparison. *School of Business and Economics Discussion Paper: Economics*, No. 2004/9.

The Constitution of Republic of South Africa, (1996). *Constitution of the Republic of South Africa, Act 108 of 1996*. Pretoria, Government Printer.

Vanderkamp, J., (1968). Interregional Mobility in Canada: A study of the Time Pattern of Migration. *The Canadian Journal of Economics*, 1(3), pp. 595-608.

Vargas-Silva, C ., (2011). Migration and Development. *Migration Observatory policy primer*, COMPAS, University of Oxford, UK, October 2011.

Vargas-Silva, C ., (2013). The Fiscal Impact of Immigration in the UK. *Migration Observatory briefing*, COMPAS, University of Oxford, UK, February 2013.

Venables, W. N. and Ripley, B. D., (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0. http://CRAN.R-project.org/package=MASS.

Veness, C., (2002-2015). Calculate distance, bearing and more between Latitude/Longitude points. Movable Type Scripts, [online], available: http://www.movable-type.co.uk/scripts/latlong.html, [25 April 2015].

Watts, C. R, Toth, G, Murphy, R. F. and Lovas, S., (2001). Domain movement in the epidermal growth factor family of peptides. *Journal of Molecular Structure (Theochem)*, 535, 171-182.

Appendices

Appendix A

Net-Internal Migration by Province

Table A.1: Net-Internal Migration by Province

Provinces	Net Internal migration_2001	Net Internal migration_2011
Western Cape	182009	189950
Eastern Cape	-253685	-315526
Northern Cape	-6616	-14115
Free State	-44713	-58780
KwaZulu-Natal	-78355	-107340
North West	-24021	30772
Gauteng	418108	550753
Mpumalanga	-29804	-21108
Limpopo	-162923	-254606

Appendix B

Net-Internal Migration

Table B.1: Net-Internal Migration in the district municipalities

CODE	Names	Net-Internal Migration	Province
TSH	City of Tshwane Metropolitan Municipality	232837	Gauteng
DC21	Ugu District Municipality	4656	KwaZulu-Natal
EKU	Ekurhuleni Metropolian Municipality	180471	Gauteng
DC16	Xhariep District Municipality	3725	Free State
CPT	City of Cape Town Metropolitan Municipality	79355	Western Cape
DC28	Uthungulu District Municipality	1932	KwaZulu-Natal
DC48	West Rand District Municipality	69672	Gauteng
DC5	Central Karoo District Municipality	811	Western Cape
DC31	Nkangala District Municipality	69127	Mpumalanga
DC45	John Taolo Gaetsewe District Municipality	503	Northern Cape
DC37	Bojanala District Municipality	46979	North West
DC6	Namakwa District Municipality	-1348	Northern Cape
JHB	City of Johannesburg Metropolian Municipality	39274	Gauteng
DC43	Sisonke District Municipality	-1405	KwaZulu-Natal
DC2	Cape Winelands District Municipality	32791	Western Cape
DC7	Pixley ka Seme District Municipality	-4842	Northern Cape
DC22	Umgungundlovu District Municipality	31369	KwaZulu-Natal
DC19	Thabo Mofutsanyane Cacadu District Municipality	-7663	Free State
DC40	Dr Kenneth Kaunda District Municipality	28535	North West
DC23	Uthukela District Municipality	-9270	KwaZulu-Natal
DC4	Eden District Municipality	27336	Western Cape
DC24	Umzinyathi District Municipality	-14263	KwaZulu-Natal
DC1	West Coast District Municipality	26683	Western Cape
DC9	Frances Baard District Municipality	-14836	Northern Cape
DC36	Waterberg District Municipality	24650	Limpopo
DC39	Dr Ruth Segomotsi Mompati District Municipality	-15204	North West
DC30	Gert Sibande District Municipality	22993	Mpumalanga
BUF	Buffalo City Metropolitan Municipality	-15735	Eastern Cape
DC3	Overberg District Municipality	22206	Western Cape
DC38	Ngaka Modiri Molema District Municipality	-18091	North West
DC10	Cacadu District Municipality	18162	Eastern Cape
DC14	Ukhahlamba District Municipality	-21159	Eastern Cape
DC8	Siyanda District Municipality	13602	Northern Cape
DC13	Chris Hani District Municipality	-28182	Eastern Cape
DC20	Fezile Dabi District Municipality	11152	Free State
DC42	Sedibeng District Municipality	-28550	Gauteng
DC29	iLembe District Municipality	10051	KwaZulu-Natal
DC47	Greater Sekhukhune District Municipality	-33377	Limpopo
MAN	Mangaung Metropolian Municipality	9238	Free State
ETH	eThekwini Metropolitan Municipality	-35519	KwaZulu-Natal
DC25	Amajuba District Municipality	5847	KwaZulu-Natal
DC18	Lejweleputswa District Municipality	-70695	Free State
DC26	Zululand District Municipality	-36136	KwaZulu-Natal
DC34	Vhembe District Municipality	-74779	Limpopo
DC44	Alfred Nzo District Municipality	-39196	Eastern Cape
DC32	Ehlanzeni District Municipality	-84587	Mpumalanga
DC12	Amathole District Municipality	-44965	Eastern Cape
DC35	Capricorn District Municipality	-91939	Limpopo
DC33	Mopani District Municipality	-53012	Limpopo
DC15	O.R.Tambo District Municipality	-103114	Eastern Cape
DC27	Umkhanyakude District Municipality	-61184	KwaZulu-Natal
NMA	Nelson Mandela Bay Metropolitan	-104906	Eastern Cape

Appendix C

Descriptive

Table C.1: Descriptive statistics

Variables	mean	max	median	std.dev	min	skewness	kurtosis
$Net - Internal\,Migration$	0	232837.0	-422.50	59146.12	-104906.0	1.38	4.14
pob	788479.50	3389278.0	646442.50	684131.99	7904.0	1.69	3.28
pow	88208.42	586495.0	35493.50	144687.70	1898.0	2.57	5.65
pod	215.16	2695.9	48.15	506.16	0.9	3.32	11.02
$nacstvs$	33.80	77.7	30.65	21.14	7.3	0.32	-1.40
tre	20.5	41.2	20.80	7.99	8.3	0.62	-0.06
ump	31.93	50.90	32.80	7.80	14.10	-0.22	-0.20
avh	79923.87	183263	74164	33097.97	37148	1.30	1.83
$illitrt$	23.82	39.50	23.85	7.80	8.90	-0.12	-0.84
GDP	$3.428711e^{+11}$	$1.010000e^{+12}$	$2.19000e^{+11}$	$2.557081e^{+11}$	$6.5259e^{+10}$	1.51	1.66
D_{ij}	806.7	2054.0	757.0	408.57	44.0	0.48	-0.39
$obp_{i/j}$	9.33	20.90	8.90	3.8	3.50	0.85	0.49
$ocr_{i/j}$	20.62	36.3	19.50	5.09	12.80	0.73	0.42
$ofp_{i/j}$	45.69	63.60	47.25	8.86	23.80	-0.51	-0.06
$poc_{i/j}$	88757.73	63.60	47.25	230728.10	1153	5.36	31.25
$poI_{i/j}$	24748.58	573334.0	3178.5	83446.37	300	5.69	33.46
$emp_{i/j}$	33.75	53.30	34.50	11.56	16.0	0.05	-1.14
$GDP_{i/j}$	3428711^{e+05}	101^{e+10}	219^{e+09}	2532852^{e+05}	65259^{e+06}	1.55	1.84
$CPI_{i/j}$	6.38	7.80	6.10	0.64	5.30	0.77	-0.32
$ad_{i/j}$	53192.12	207487	40462.5	42801.64	4375.0	1.90	3.53
$pwa_{i/j}$	88.87	99.40	95.65	12.92	49.10	-1.48	1.44
$Inftrdwell_{i/j}$	22.17	57.90	18.10	12.93	2.30	1.15	0.89
$pop_{i/j}$	9.955877^{e+05}	4.434827^{e+06}	7.513155^{e+05}	9.187444^{e+05}	7.1011^{e+04}	2.19	4.38

Table C.2: Descriptive statistics of the log variables

Variables	mean	max	median	std.dev	min	skewness	kurtosis
$log(D_{ij})$	6.54	7.63	6.63	0.60	3.78	-0.83	0.66
$log(obp_{i/j})$	2.15	3.04	2.19	0.41	1.25	-0.05	-0.63
$log(ocr_{i/j})$	3.00	3.59	2.97	0.24	2.55	0.13	-0.45
$log(ofp_{i/j})$	3.80	4.15	3.86	0.22	3.17	-1.08	1.04
$log(poc_{i/j})$	9.90	14.28	9.57	1.69	7.05	0.39	-0.54
$log(poI_{i/j})$	8.46	13.26	8.06	1.54	5.70	0.90	0.72
$log(emp_{i/j})$	3.45	3.98	3.53	0.38	2.77	-0.39	-1.04
$log(GDP_{i/j})$	26.32	27.64	26.11	0.71	24.90	-0.10	-0.15
$log(CPI_{i/j})$	1.85	2.05	1.81	0.10	1.67	0.61	-0.50
$log(ad_{i/j})$	10.60	12.24	10.61	0.77	8.38	-0.32	0.49
$log(pwa_{i/j})$	4.47	4.60	4.56	0.17	3.89	-1.85	3.01
$log(Inftrdwell_{i/j})$	2.93	4.06	2.90	0.62	0.83	-0.64	1.35
$log(pop_{i/j})$	13.48	15.3	13.53	0.83	11.17	-0.25	0.67

Appendix D

The R code

```
#import the migration data set
data =read.csv("H:/data_sets/data_c.csv",header =TRUE,sep =",")
data.out =data[c(8,12,15,16,18,19,20,25,26,27,28,33,34,37

,38 ,47,48, 49,50, 53,54,57,89)]



#Preparing for gravity model: Out-migrtion
This program replace the zero's by 0.1, because the log(0) is
undefined.

d =dim(data.out)
for(j in 1:d[1]){
 if(data.out[j,1]==0){
```

```
   data.out[j,1] =0.1  replace 0.1 in the place of zero,
  } else{
    data.out[j,1] =data.out[j,1]
     otherwise, leave the number.
  }
  }
  #Fitting a gravity model
 gravity model= lm(log(data.out[,1])~log(D_ij)
 +log(pop_{i})
+log(pop_{j}), data =data.out)
summary(gravity model)
extended.out =lm(log(data.out[,1])~.,

data=log(data.out[,-c(1,14,15)]))

summary(extended.out)

#Centering variable
The following program is centering all the independent variable
preparing for fitting a modified gravity model
with nonlinear term.
#The purpose of cenetering variables was to
minimise multicollinearity:

data_out =data.out[,-c(14,15)]
```

```
c_varb =data_out[,-1]
c=dim(c_varb)

for(k in 1:c[2]){
  c_varb[,k] =scale(log(c_varb[,k]),scale =F)
}

gr_out =lm(log(data_cvar[,1])~.,data =data_cvar[,-1])
summary(gr_out)

gr_out2sq =update(gr_out,~.+I(D_ij^2)+I(obpi^2)
+I(obpj^2)

+I(ocrj^2)+I(ofpi^2)+I(ofpj^2)+I(poci^2)+I(pocj^2)
+I(poIi^2)+I(poIj^2)+I(empi^2)+I(empj^2)+I(GDP_i^2)
+I(GDP_j^2)+I(CPI_i^2)+I(CPI_j^2)+I(adi^2)+I(adj^2)
+I(pwai ^2)+I(Inftrdwelli^2),data =data_cvar[,-1])


summary(gr_out2sq)


#Preparing for gravity model: In-migrtion

data_in2 =cbind(data.in[,1],out[c(2,14,15)])
```

```
min(data_in2[,1])
d2 =dim(data_in2)
for(j in 1:d2[1]){
    if(data_in2[j,1]==0){
        data_in2[j,1] =0.1
      } else{
          data_in2[j,1] =data_in2[j,1]
         }
   }

gravity_in =lm(log(data_in2[,1])~log(D_ij)+log(pop_i)
+log(pop_j), data =data_in2)
summary(gravity_in)

c_ivarb =data_in[,-c(1)]  Excluding the column with
in-migration entries,
ci=dim(c_ivarb)
for(h in 1:ci[2]){
  c_ivarb[,h] =scale(log(c_ivarb[,h]),scale =F)
}

gr_in =lm(log(data_icvar[,1])~.,data =data_icvar[,-1])
summary(gr_in)
in2 =update(gr_in,~.+I(D_ij^2)+I(obpi^2)
```

```
+I(obpj^2)+I(ocri^2)+I(ofpi^2)+I(ofpj^2)+I(poci^2)
+I(pocj^2)+I(poIi^2)+I(poIj^2)+I(empi^2)+I(empj^2)
+I(GDP_i^2)+I(GDP_j^2)+I(CPI_i^2)+I(CPI_j^2)

 +I(adi^2)+I(adj^2)+I(pwaj^2)
+I(Inftrdwell_i^2)+I(Inftrdwell_j^2))



#Fitting the Poisson, Negative Binomial and Gamma model
library(MASS)
poismodel =glm(out[,1]~.,family =poisson,data =out))
nbmodel =glm.nb(out[,1]~.,data =out[,-c(1)])

#Deleting rows with zero entries:
row_sub = apply(out3, 1, function(row) all(row !=0 ))

gammodelt =glm(out3[,1]~.,family=Gamma(log),data =out3[,-1])
summary(gammodel, dispersion =gamma.dispersion(gam_out))

#The same models (Poisson, NB and Gamma) were

used for modelling in-migration.

#Modelling Net-internal migration
```

```
dc_net =read.csv("H:/net.csv",header =TRUE,sep =",")


netmig =dc_net[c(4,15,17,19,22,23)]
modnmig<-lm(Net_internal migration~pod+nacstvs+pob

+pow+tre,


 data = netmig)
summary(modnmig)
anova(modnmig)
xy =dc_net[c(2,3)]
xy.sp =SpatialPoints(xy)
xy.cc =coordinates(xy.sp)

#Converting the Net-migratio data into Spatial data frame using
the "sp" package

xy.spdf =SpatialPointsDataFrame(xy.sp,netmig)
df1 =data.frame(xy,netmig)
coordinates(df1)=c("x_coor","y_coor" )

#Performing the Monte Carlo Significance Test
```

```
DM<-gw.dist(dp.locat=coordinates(df1))
bw<-bw.gwr(Net_migration~pod+nacstvs
+pob+pow+tre

,data=df1,dMat=DM, kernel="gaussian")

res.mont1<-montecarlo.gwr(Net_migration~pod

+nacstvs+pob+pow+tre, data = df1,dMat=DM, nsim=99,

kernel="gaussian", adaptive=FALSE, bw=6)



#Cluster Analysis
The  number of clusters were estimated  by following  R code.

datamgr <-read.csv("C:/Users/xolanij/Desktop/
data_net2.csv", sep=";", dec=",")

netcl =datamgr[c(1,4,6,7,8,15,17,18,19,20,22,23)]

netclus =scale(netcl[,-c(1,2,3,4,5,8,10)])

pamk.best <- pamk(netclus)
```

```
cat("number of clusters estimated by
optimum average silhouette width:", pamk.best$nc, "\n")

number of clusters estimated by optimum average

silhouette width: 2



#Print the name of the district municipalities that form the clusters.

clusmeancode =cbind(netclus,netcl[c(1)])
for (i in 1:2){
  print(paste("District municipalities in Cluster ",i))
  print(clusmeancode$CODE[fitkms$cluster==i])
  print (" ")
}
```

Appendix E

Sample results

```
 resettest(extended_in , power=2, type="regressor")
RESET test
data:  extended_in
RESET = 9.6884, df1 = 21, df2 = 2609, p-value <  2.2e-16

RESET test

data:  extended_out
RESET = 10.5934, df1 = 22, df2 = 2607, p-value < 2.2e-16

#Results of the Monte Carlo

bw<-bw.gwr(Net_migration~pod+nacstvs+pob+pow+tre,
```

```
data=df1,dMat=DM, kernel="gaussian")
Fixed bandwidth: 9.939891 CV score: 67246993225
Fixed bandwidth: 6.144419 CV score: 56758085760
Fixed bandwidth: 3.798688 CV score: 80700391253
Fixed bandwidth: 7.59416 CV score: 61456048023
Fixed bandwidth: 5.248429 CV score: 55551501230
Fixed bandwidth: 4.694678 CV score: 58472049266
Fixed bandwidth: 5.590667 CV score: 55553465354
Fixed bandwidth: 5.036915 CV score: 56100255724
Fixed bandwidth: 5.379153 CV score: 55448157468
Fixed bandwidth: 5.459944 CV score: 55453723699
Fixed bandwidth: 5.329221 CV score: 55469902081
Fixed bandwidth: 5.410012 CV score: 55444633740
Fixed bandwidth: 5.429084 CV score: 55446014667
Fixed bandwidth: 5.398225 CV score: 55445123384
Fixed bandwidth: 5.417297 CV score: 55444847909
Fixed bandwidth: 5.40551 CV score: 55444697873
Fixed bandwidth: 5.412795 CV score: 55444669369
Fixed bandwidth: 5.408292 CV score: 55444640419
Fixed bandwidth: 5.411075 CV score: 55444640586
Fixed bandwidth: 5.409355 CV score: 55444633698
Fixed bandwidth: 5.408949 CV score: 55444635274
Fixed bandwidth: 5.409606 CV score: 55444633336
Fixed bandwidth: 5.409761 CV score: 55444633346
Fixed bandwidth: 5.40951 CV score: 55444633419
```

```
Fixed bandwidth: 5.409665 CV score: 55444633319
Fixed bandwidth: 5.409702 CV score: 55444633321



#Tests based on the Monte Carlo significance test

                     p-value
(Intercept)             0.73
pod                     0.53
nacstvs                 0.66
pob                     0.05
pow                     0.01
tre                     0.54



#Cluster Analysis: Printing the district municipalities that

 form clusters
[1] "District municipalities in Cluster 1"
 [1] DC10 DC12 DC13 DC14 DC15 DC44 BUF  NMA  DC16 DC18 DC19
[12] DC20 MAN  DC42 DC48 DC21 DC22 DC23 DC24 DC27 DC28 DC43
[23] DC25 DC26 DC29 DC33 DC34 DC35 DC36 DC47 DC30 DC31 DC32
[34] DC37 DC38 DC39 DC40 DC6  DC7  DC8  DC9  DC45 DC1  DC2
```

```
[45] DC3  DC4  DC5
```

```
 [1] "District municipalities in Cluster 2"
[1] EKU JHB TSH ETH CPT
```