UNIVERSITY OF
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

Investigating the Spatial Distribution of Diabetes in Africa Using Both Classical and Bayesian Approaches

By

Adenike Oyelola Soogun
216065101

A dissertation submitted in fulfilment of the academic requirements of Master of Science in Statistics

School of Mathematics, Computer Science and Statistics

College of Agriculture, Engineering and Science.

Supervisor: Dr Siaka Lougue

November 2017

Disclaimer

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

Dedication

This work is dedicated to Almighty God, my husband

Dr Sunday Oluwambe Soogun and to my adorable children

Favour, Victor and Glory Soogun.

“If you can dream it you can be it”

Declaration

I, Adenike Oyelola Soogun, declare that this dissertation titled, ‘Investigating the Spatial Distribution of Diabetes in Africa Using Both Classical and Bayesian Approaches’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- No part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given.
- With the exception of such quotations, this dissertation is entirely my own work.
- All references that were used are duly acknowledged in the Bibliography.

______________________________ ______________________________

Adenike Oyelola Soogun (Student) Date

______________________________ ______________________________

Dr Siaka Lougue (Supervisor) Date

Abstract

Spatial analysis and cluster analysis are important statistical tools in decision making, planning, and information management, and are applicable in epidemiological and public health research. Patterns and variations in disease occurrence or epidemiology can be easily spotted and resolved through these visual and statistical procedures. The evolution of Bayesian statistics with its underlying benefits and accuracy over frequentist statistics prompted its application in this study.

The study aimed at investigating the pattern of Diabetes occurrence across Africa with the identification of some underlying high-risk factors associated with the diseases using both frequentist and Bayesian approaches.

Metadata from WHO, the World Bank data and other United Nations sources were extracted to form the data for the study. Diabetes prevalence metadata for the year 2015 from World Bank data with age 20-65 years was used. Spatial analysis of prevalence was used to examine the geographical distribution of the disease and its indicator factors on the continent. The data was further stratified into five regions and analysis was carried out both across the continent and across the regions. A spatial autocorrelation test with the use of Global Moran index (GMI) was used to determine the spatial pattern for the whole continent. Local Indicator Spatial Autocorrelation clustering map was used to identify the hotspots countries. Cluster analysis was performed to further investigate the clustering of the disease and to show similar groups of countries. Classical and Bayesian techniques form the statistical methods used to examine the impact of the moderators' variables on diabetes prevalence in Africa. Linear regression, fractional probit, Poisson and a Negative binomial regression model were used to investigate the risk factors. The analysis was done using R software using libraries "maptools", "ape", "spdep" "msclust" and SAS software.

The GMI result shows that there is spatial autocorrelation of diabetes prevalence in Africa, with a spatial pattern of clustering in the Southern and East Africa region and a spatial pattern of dispersion in the West, Central and North Africa regions. A positive spatial Autocorrelation was observed in the West, Central and East Africa region and a negative spatial autocorrelation within the North and Southern Africa region. The LISA cluster maps indicated that there are sixteen hotspot countries identified, which are: South Africa, Mali, Congo DRC, Ethiopia, Uganda, Madagascar, Cote d'Ivoire, Liberia, Senegal, Chad, Eritrea, Tanzania, Lesotho, Libya, and Morocco. Cluster analysis was used to strengthen the LISA Maps, where homogenous groups of countries with similar pattern were clustered. Likewise, because of the significance of some socioeconomic risk factors of diabetes (GDP, Population age, physician density per 1000 person and urbanization), from the fractional probit regression model, a group of clusters were created based on countries with high prevalence of those significant risk factors. No significant difference was observed between the classical analysis and Bayesian analysis of the disease using simple regression model, loglinear, fractional probit, Poisson and negative binomial model due to the over dispersion of the data and because of unavailability of non-informative prior.

In conclusion, diabetes is heterogeneous in Africa. However, homogenous grouping of the disease can help policy makers in forming a joint collaborative effort in the planning, control and effectively managing the disease in Africa, instead of fighting the disease in silos. At the population level, GDP, population age, urbanization, and physician density have been identified as significant and correlated socioeconomic risk factors of diabetes in Africa.

Keywords: Spatial distribution, Spatial Autocorrelation, Geographic weighted regression, Cluster Analysis, fractional probit, Poisson, Negative-binomial, Bayesian Statistics, Diabetes.

Acknowledgement

My profound gratitude goes to the almighty God for His strength all through this work.

I am indeed grateful to my supervisor Dr. Siaka Lougue whose encouragement, guidance and persistent criticism have helped in this work. This work could not have been accomplished without his continuous support. I am thankful to the College of Agriculture, Engineering, and Science (UKZN) for the bursary I received. Indeed, it went a long way in making this research output a success.

My sincere gratitude also goes to colleagues and friends at the University of KwaZulu-Natal for providing a pleasant, supportive and inspiring environment to study. To my extended family and friends, thanks for your continued support and encouragement.

Last but not the least, special thanks to my husband Dr Sunday Soogun whose support has been immeasurable all through, and especially who encouraged me to do this second Master's degree in the first place. It is a dream come true. To my children Favour, Victor and Glory, thanks for your understanding.

Oral Presentations at Conferences

1. Soogun, A.O. and Lougue S. “Spatial Correlation of Diabetes Across African Countries Using Meta-data”. Oral Presentation at Faculty of Science and Agriculture Post Graduate Research day, University of KwaZulu Natal, Durban. South Africa. September 2016.
2. Soogun, A.O. and Lougue S. “Spatial profile of Diabetes prevalence in Africa with correlated Socio-economic factors”. Oral presentation at the 28th International Population Conference of International Union for the Scientific Study of Population (IUSSP), 29th October - 5th November 2017, Cape Town, South Africa.
3. Soogun, A.O. and Lougue S. “Spatial Correlation of Diabetes prevalence in Africa using Meta-data”. Oral presentation at the 59th National Conference of South Africa Statistical Association, 27th - 30th November 2017. Ilanga Estate, Bloemfontein, Free State, South Africa.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

IDF	**I**nternational **D**iabetes **F**ederation
WHO	**W**orld **H**ealth **O**rganization
WDB	**W**orld **D**ata **B**ank
IGT	**I**mpaired **G**lucose **T**olerance
T2DM	**T**ype **2** **D**iabetes **M**ellitus
MENA	The **M**iddle **E**ast and **N**orth **A**frica
SEA	**S**outh **E**ast **A**sia
SCA	**S**outh and **C**entral **A**merica
WP	**W**estern **P**acific
NAC	**N**orth **A**merica **C**aribbean
AFR	**Af**rica **R**egion
SSA	**S**ub-**S**aharan **A**frica
GNI	**G**ross **N**et **I**ncome
GDP	**G**ross **D**omestic **P**roduct
MYS	**M**ean **Y**ear of **S**chooling
HDI	**H**uman **D**evelopment **I**ndex
GMI	**G**lobal **M**oran **I**ndex
LISA	**L**ocal **M**oran **I**ndices of **S**patial **A**utocorrelation
SE	**S**tandard **E**rror
d.f	**D**egree of **F**reedom
REML	**R**estricted **M**aximum **L**ikelihood
NB	**N**egative **B**inomial
NI	**N**on-**I**nformative prior

Chapter 1

Introduction

1.1 Background

Diabetes is a chronic disease which occurs when the pancreas does not produce enough insulin or when the body does not effectively use the insulin it produces. This then leads to an increased concentration of glucose in the blood (hyperglycaemia). There are three types of diabetes: type 1, type 2 and gestational diabetes. Type 1 diabetes, which was previously known as insulin-dependent or childhood-onset diabetes, is mainly characterized by a lack of insulin production. Type 2 diabetes, previously called non-insulin dependent or adult-onset diabetes, is characterized and caused by the ineffective use of insulin in the body, which often results from excess body weight and physical inactivity. Gestational diabetes, on the other hand, is a form of hyperglycemia that occurs during pregnancy (IDF, 2015; ADA, 2014).

The general symptoms of diabetes are polyuria, polydipsia, weight loss, blurred vision, and abnormally higher levels of glucose, a form of sugar in the blood (ADA, 2014). Symptoms can also be so mild as to escape notice. With Type 2 diabetes, many people don't get to know until long-term damage has been caused. Type 1 symptoms usually manifest quickly in a matter of days or weeks, and can be very severe, even to the point of causing sudden death. Common symptoms for both types of diabetes are hunger and fatigue, frequent urinating (average normal urination rate is 4 to 7 times in 24 hours, but an individual with diabetes may experience a much higher rate), dry mouth and itchy skin, blurred vision, slow-healing sores or cuts, nausea and vomiting (WHO, 2000). For many decades, diabetes diagnosis criteria have been based on glucose criteria, either through the FPG or the 75g Oral Glucose tolerance test (OGTT) (ADA, 2014).

According to the World Health Organisation (WHO) report of 2000, the criteria and methods of diabetes diagnosis include: a random venous plasma glucose concentration ≥ 11.1 mmol/l or, a fasting plasma glucose concentration ≥ 7.0 mmol/l or two hours plasma glucose concentration ≥11.1 mmol/l two hours after 75g anhydrous glucose in an OGTT. A Haemoglobin A1c (HbA1c) testing of 48mmol/mmol (6.5%) is recommended as a cut-off point for diagnosing diabetes but is not appropriate in the following situations: with children and young people, an individual taking medication that can cause a rapid rise in glucose level, pregnant women, and the presence of genetic factors. When the Haemoglobin A1c (HbA1c) is a less frequent diabetes test than home monitoring. An HbA1c of 6.5% is recommended as the cut point for diagnosing diabetes. Therefore, an HbA1c value of less than 6.5% does not exclude diabetes diagnosed using glucose test. However, finger-prick HbA1c is an alternative diagnostic test which should not be used except when the quality of the procedure is guaranteed. A person with <48mmol/mmol (6.5%) should be treated as at high risk of diabetes, with a repeat test conducted after six months or sooner if symptoms persist (WHO, 2000).

There are several options available in the treatment and management of diabetes. These include change of diet, active exercise, a constant check-up of blood sugar, administration of medications as injectable drugs (GLP-1 receptor agonists, pramlintide) to bring the blood sugar under control, insulin, and weight loss surgery. Different disease situations may necessitate some changes in the patient's medication.

In this 21st century, diabetes is one of the largest global health emergencies with one in every eleven adults having diabetes (IDF, 2015; WHO, 2000). An estimated 415 million adults currently have diabetes, with 318 million adults being diagnosed with impaired glucose tolerance which puts them at risk of developing the disease in future (IDF, 2015; Bonow and Gheorghiade, 2004; Day, 2016; Ogurtsova et al., 2017; Shaw et al., 2010; Whiting et al., 2011).

According to IDF (2015), the number of adults who died from diabetes globally is 5.0 million as compared to the numbers for HIV/AIDS which stand at 1.5 million, tuberculosis at 1.5 million, and malaria at 0.6 million. Therefore, this alarming number of death attributed to diabetes globally as compared to tuberculosis, HIV/AID and malaria necessitates an urgent intervention through the development and implementation of a cost-effective framework of improving the health condition of people with diabetes and prevention of new occurrences. In halting the rise of diabetes, a global diabetes awareness network by countries in the Africa continent through joint policy and collaborative effort became the background to this study (Day,C 2016).

Africa is the second-largest and second most-populous continent on earth with an estimated population in 2016 of 1.2 billion people. It is home to 54 recognized sovereign states and countries. In 2015, 14.2 million people (aged 20-79) were estimated to have diabetes in Africa with a regional prevalence of 2.1-6.7% (Day,C 2016). Therefore, a closer look at the countries within the Africa continent with a similar pattern of diabetes occurrence may aid in decreasing its prevalence.

Several factors account for diabetes prevalence both at the individual and country level. At the individual level, diabetes risk factors are categorized into two. The first category contains complex non-modifiable factors which include genetic, age, gender, ethnicity and family history. The second category comprises significant modifiable risk factors such as overweight, physical inactivity, sedentary behavior and change in dietary/carbohydrate intake, smoking, alcohol consumption, age distribution, and urbanization. At the country level, indicators of diabetes prevalence are Gross Domestic Product (GDP) and Gross National Income (GNI).

1.2 Research Problem

On the African continent where diabetes was once rare, there has been a tremendous increase in its prevalence due to an increase in economic development. According to IDF (2015), in Africa, the adult population (age 20-69) with diabetes is 441 million, with a projection of 926 million in 2040 and a regional prevalence of 3.2% with a projection of 3.7% in 2040.

In Africa, an increasing and alarming issue is the fact that people's lifestyles are becoming more sedentary, unhealthy eating habits are becoming widely adopted, and urbanization is increasing. These are environmental factors contributing to the rise of diabetes prevalence on the continent. Type 2 Diabetes is the common type of diabetes which affects 90% of the people who have diabetes (IDF, 2015). As was stated by Azevedo and Alla (2008), the potential severity of diabetes is such that some epidemiologists predict that its economic impact and the death toll will surpass the ravages of HIV and AIDS in the near future. However, many studies have been focusing on the estimating risk factors at the individual level but none at the population level. Likewise, many studies have been seen on developed nations like America and China on the spatial distribution of obesity, mortality but no study has been found on the spatial distribution of diabetes in Africa as a continent but few on the Sub-Sahara Africa countries. Therefore, more investigation needs to be performed on diabetes dynamics to establish the disease pattern and clustering and the likely socio-economic risk factors of diabetes prevalence across Africa. Since disease risk factors often exist in space and time, it would be appropriate to conduct the study in space and time for adequate understanding.

Statistical modelling has been one of the successful tools employed in analyzing and understanding the possible trends exhibited in infectious diseases. This study seeks to investigate the spatial pattern of diabetes in the Africa region with the aim of providing a suitable recommendation aimed at its eradication on the continent.

1.3 Significance of the Study

Today in Africa, diabetes and its complications are considered a pandemic. This has compelled African governments to start paying more attention to its impact and contributing factors, as thousands of Africans run the risk of dying young. In 2015, the number of death caused by diabetes in the continent was 321,000. The poor healthcare system in the most countries on the continent is a huge concern, with health expenditure due to diabetes of 3.4 Billion dollars in 2015 with a projection of 5.5. Billion dollars in 2040 (IDF, 2015). This study is significant, as it aims to reveal the underlying high-risk factors of diabetes and hotspots countries in Africa. Countries and regions with a similar pattern of the prevalence will be revealed and this could aid the approach in curbing the disease collectively. Furthermore, the study will reveal the competence and superiority of the Bayesian method as against classical statistical methods.

1.4 Research Questions

The study sought to provide answers to the under listed research questions:

1. How is Diabetes prevalence spatially distributed in Africa?
2. Which countries have a similar pattern in Diabetes prevalence?
3. What are the highest contributing risk factors of Diabetes prevalence in Africa?
4. How are the risk factors of Diabetes spatially distributed geographically in Africa?
5. How Bayesian statistics can perform compared to classical statistics.

1.5 Aim and Objectives

The study aimed at investigating the spatial distribution of diabetes prevalence across Africa with the identification of underlying risk factors associated with the diseases using both classical and Bayesian approach.

The specific objectives of the study are to:

i. Determine the relationship between identified factors of diabetes and its prevalence in Africa,

ii. Examine spatial configurations of diabetes prevalence levels and its underlying related factors in Africa,

iii. Compare the Bayesian statistics technique to classical statistics techniques as applied

1.6 Thesis Structure

The thesis is structured into six (6) chapters, summarised as follows:

In Chapter 1, the research background, research problem, significant of the study, research aim, objectives and questions are described

In Chapter 2, a review literature related to the epidemiology of diabetes prevalence, diabetes prevalence in the world and in Africa with its underlying risk factors is discussed.

In Chapter 3, we explain the statistical techniques used, such as cluster analysis, spatial analysis, spatial autocorrelation measure using Global and Local Moran indices, along with Local indicator spatial analysis (LISA) and cluster analysis, to determine the spatial pattern of diabetes across Africa using 2015 data. Classical and Bayesian methods of linear regression, Poisson, and negative binomial regression analysis are also discussed.

In Chapter 4, we present the findings of the data analysis.

In Chapter 5, the discussion, conclusion, and recommendation of the results are explained.

Chapter 2

Literature Review

2.0 Introduction

This chapter gives a literature review of diabetes prevalence. The epidemiology of diabetes is discussed, followed by a review of diabetes prevalence in different continents and regions in the world, as well as in Africa. The last section discusses the risk factors of diabetes.

2.1 Epidemiology of Diabetes

Diabetes is a chronic condition that occurs when the body cannot produce enough insulin or cannot use insulin and is diagnosed by the observation of a rise in the level of glucose in the blood (IDF, 2015; ADA, 2014). Insulin is a hormone produced in the pancreas which is required to transport glucose from the bloodstream into the body cells where it is used as energy. A person with diabetes is characterized by the lack or ineffectiveness of insulin in the body and the resulting circulation of glucose in the blood. Over time, this results in a high level of glucose in the blood, which is known as hyperglycaemia (high blood sugar), causing a huge damage to many tissues in the body and leading to the development of disabling or life health complications, and when no treatment is given or not early diagnosed, sudden death. People with diabetes face more health problems. Consistently high blood glucose levels can lead to serious diseases affecting the heart and blood vessels, eyes, kidneys, and nerves. People with diabetes are at an increased risk of developing numerous infections. In most high-income countries, diabetes is a leading cause of cardiovascular disease, blindness, kidney failure and lower-limb amputation.

According to IDF (2015), WHO (2000) and ADA (2014), diabetes is currently classified into four categories. Type 1 diabetes is caused by an autoimmune reaction in which the body's defense system attacks the insulin-producing beta-cells in the pancreas. This results in the body's inability to produce insulin. Type 1 can affect people at any age, but usually occurs among children or young adults. This type of diabetes requires a daily dose of insulin to control the level of glucose in the blood, and a lack of this can result in death. Type 1 often develops suddenly, with abnormal thirst and dry mouth, frequent urination, lack of energy, extreme tiredness, continuous hunger, sudden weight loss, and blurred vision as symptoms. Type 2 diabetes is the most common type of diabetes. It occurs in adults but is becoming increasingly seen in children and adolescents. With type 2 diabetes, the body can produce insulin but develops resistance so that the insulin is ineffective, resulting in an insufficient level of insulin in the body. Both insulin resistance and deficiency lead to high glucose levels. Symptoms of type 2 diabetes are similar to those of type 1, which include frequent urination, excessive thirst, weight loss, and blurred vision. Many people with type 2 diabetes can be unaware of the condition for a long time, as the symptoms are less marked than in type 1 diabetes and may take years before they are established. Although the cause of Type 2 diabetes is still unknown, there are several risk factors such as being overweight, physical inactivity, and poor nutrition, and diagnosis in middle or old age. The third category is known as gestational diabetes. It occurs during pregnancy tends to occur from the 24th week of pregnancy (slightly raised blood glucose level) and diabetes mellitus in pregnancy (women with substantially elevated blood glucose). The fourth category of diabetes is a less common type of diabetes that results from genetic defects, drug or chemical use, infections, or other diseases. This study data was on type 2 diabetes and this is our point of focus.

As noted in Day ,C (2016), Shaw et al. (2010), Bonow and Gheorghiade (2004), Ogurtsova et al. (2017), IDF (2015), and WHO (2000), diabetes imposes a large economic burden on the global health care system and economy at large. This burden can be measured in terms of direct and indirect medical costs associated with productivity loss, national GDP, GNI, and increase in total global diabetes health expenditure.

The incidence of diabetes, especially type 2, is rapidly growing in the world. In 1985, an estimated 30 million people suffered from this chronic disease, which, by the end of 2006, had increased to 230 million, representing 6% of the world population. Of this number, 80% is found in the developing world (Roglic et al., 2005; Roglic and Unwin, 2010). In 2000 it was predicted that 366 million people will have diabetes in the year 2030, just 20 million more than the 2011 estimate (Wild et., al. 2004).

It is estimated that during the next 35 years, diabetic world-wide prevalence will reach 25% of the world's population, with India being the hardest hit (IDF, 2015). For a long time, Africa was considered safe from many of the diseases that are called "diseases of affluence," which plague the Western world

(Day, C 2016, Whiting et al., 2011). Similarly, there was a time when Africa was thought to be a continent relatively free of diabetes mellitus. However, diabetes is now considered to be a common disease in Africa, a situation that seemed to have remained virtually static until recent years. Indeed, from 1959 to the mid-1980s, medical statistics showed that the prevalence of diabetes in Africa was equal to or less than 1.4%, apart from South Africa, where the rate was estimated to be as high as 3.6% in 2001 (Motala, 2002; Motala et al., 2003; Omar et al., 1993). As at 1994, the continent-wide prevalence of diabetes mellitus stood at 3 million and was then predicted to double or triple by the year 2010 (Peer et al., 2014, Sobngwi et al., 2012, Sobngwi et al., 2001, Kengne et al., 2013). Approximately 7.1 million Africans were said to be suffering from diabetes at the end of 2000, a figure that was expected to rise to 18.6 million by 2030 (Wild et al., 2004).

According to Organisation (2000), an estimated 2.9 million people died globally from diabetes in 2000, i.e. a case-fatality rate (CFR) of 0.0161. Also, in 2000, the prevalence of diabetes in the WHO African Region was estimated at 7.02 million people, out of which about 0.702 million (10%) people had type 1 diabetes and 6.318 million (90%) had type 2 diabetes (IDF, 2015). It was estimated that about 113,100 people died from diabetes-related causes, 561,600 were permanently disabled, and 6,458,400 experienced temporary disablement (IDF, 2015, Whiting et al., 2011).

By 2030, the prevalence of type 2 diabetes in the Middle East, Indian, South Asia and Sub Sahara Africa is expected to increase by more than 150% (Hossain et al., 2007). Diabetes has been associated with poverty in low and high-income countries, compared to middle-income countries where it is associated with overconsumption of high fats foods with low or no physical activity due to urbanization (Aspray et al., 2000; Assah et. al., 2011).

Globally in 2013, it is estimated that almost 382 million people suffered from diabetes, a prevalence of 8.3%. North America and the Caribbean is the region with the highest prevalence of 11%, having 37 million people with diabetes. It is followed by the Middle East and North Africa with a prevalence of 9.2% (35 million people). Western Pacific is the region with the highest number of people living with diabetes (138 million). However, its prevalence is 8.6%, close to that of the World. (IDF, 2015, Alwan, 2011, Bonow and Gheorghiade, 2004, Shaw et al., 2010).

According to IDF (2015), the top 10 countries with the highest prevalence of diabetes in 2013 are Tokelau (37.5%), Federated States of Micronesia (35%), Marshall Islands (34.9%), Kiribati (28.8%), Cook Islands (25.7%), Vanuatu (24%), Saudi Arabia (23.9%), Nauru (23.3%), Kuwait (23.1%) and Qatar (22.9%). It is interesting to highlight that 35 out of 219 countries (16% of the total) have very high prevalence of diabetes of 12% or higher. These countries are located mainly in Western Pacific, and the Middle East and North Africa regions. Africa is the region with the lowest prevalence of diabetes (4.9%), having Réunion (15.4%), Seychelles (12.1%) and Gabon (10.7%) as the top three countries with high prevalence and 10 out of 48 countries with prevalence of diabetes higher than the upper quartile (6.3%) prevalence (IDF, 2015). In 2015, Europe had 56 million people with diabetes (8.5%) having Turkey in the upper extreme of the prevalence of diabetes with 14.9%, four percentage points higher that Montenegro (ranked number two) with 10.1% of prevalence. Likewise, in North America and the Caribbean, Belize (15.9%), Guyana (15.8%) and Curacao (14.5%) are the top three countries with the higher prevalence of diabetes. At the same time, this region presents the highest values of the prevalence of Impaired Glucose Tolerance (IGT) with a median of 12% (IDF, 2015, Riste et al., 2001).

As stated by (Færch et., al. 2016), Type 2 diabetes is a heterogeneous disease with large variation in the relative contributions of insulin resistance and beta cell dysfunction between subgroup and individual. Many studies ascertain that magnitude and geographic distribution of the prevalence of diabetes heterogeneous (Barker, 2011; Tuomi, 2014; Schwitzgebel, 2014).

2.2 Prevalence of Diabetes in the World

From the diabetes prevalence statistics given in the previous section, the importance of learning about risks and warning signs of diabetes in order to take actions to prevent it and seek health care when it is developed becomes evident. However, diagnosis of diabetes is both a personal and health system and services responsibility. Diabetes risk assessment and testing must be integrated into primary health care with universal health coverage.

As stated by IDF (2015), Majeed et., al. (2013), Matiner (2013), 35 out of 219 (that is 16 % of the total) countries in the Western Pacific, Middle East and North Africa region have a high prevalence of diabetes. In the Middle East and North Africa, approximately 34.4. Million people or 9.1% of adults ages 20-79 are living with diabetes with over 40% of undiagnosed cases. Countries with high diabetes prevalence include Saudi Arabia (raw diabetes prevalence of 17.6%) and Kuwait (14.3 %). Diabetes was responsible for 342,000 deaths in 2015, and over half (51.3) of all death from diabetes in the region occurred in people under 60 years. Azeem et. al., (2013) mentioned that the prevalence of diabetes is also very high by international standards in countries within the region which are not as affluent as the Gulf States, such as Egypt, Lebanon, and Oman. Likewise, the largest number of cases is seen in Egypt, Pakistan, and Iran, reflecting their status as the countries with largest populations in the MENA region.

In Europe, the number of people with diabetes is estimated to be 59.8 million (aged 20-79), including 23.5 million undiagnosed cases (IDF, 2015). While the Europe region has the second-lowest age-adjusted comparative diabetes prevalence rate of any IDF region (after the Africa Region), there are still many countries with relatively high diabetes prevalence rates (IDF, 2015). Turkey has the highest age-adjusted comparative prevalence of diabetes with 12.8% and the third highest number of people with diabetes in the Europe region with 6.3 million, after Germany (6.5 million) and Russia (12.1 million) (IDF, 2015). Tamayo et. al., (2013), in their study, showed that in 2013, the number of people with diabetes was estimated to be 56 million, with an overall estimated prevalence of 8.5%, and with diabetes being responsible for 10% of total health expenditure in 2010. Age is an important risk factor for diabetes. In the Europe region, 30.8% of the general population were aged between 20 and 79 years in 2015, and this is expected to increase to 35.6% by 2040. In 2015, approximately 627 000 people aged 20 -79 died from diabetes in the Europe region (IDF, 2015).

According to IDF (2015), the North America and Caribbean (NAC) region has the highest prevalence of diabetes compared to other IDF regions. An estimated 44.3 million people aged 20-79 lived with diabetes in the region in 2015, of which 13.3 million (29.9%) are undiagnosed, with 12.9% of the adult population being affected. The USA has the highest prevalence at 29.3 million, followed by Mexico (11.5 million) and Canada (2.5 million). In the region, the total number of deaths in 2015 attributable to diabetes was 324,000. More men (173,000) than women (151,000) died from diabetes-related causes. In addition, diabetes mortality in this region was not limited to older age groups. Over one-third (38.3%) of deaths occurred in adults under 60 years. In the USA, more than 219,000 people died from diabetes in 2015. 14% of the region’s total expenditure is spent on diabetes. In North America and the Caribbean, Belize (15.9%), Guyana (15.8%) and Curacao (14.5%) are the top three countries with the highest prevalence of diabetes. At the same time, this region presents the highest values of the prevalence of Impaired Glucose Tolerance (IGT) with a median of 12%.

According to IDF (2015), in the South and Central America (SCA) region, an estimated 29.6 million people or 9.4% of the adult population had diabetes in 2015. 11.5 million (39.0%) were undiagnosed. Over 82% of people with diabetes lived in urban areas and 81% with diabetes were living in middle-income countries. Puerto Rico has the highest prevalence of diabetes among adults (12.1 % age-adjusted comparative prevalence and 14.2% raw prevalence). Brazil has the highest number of people with diabetes (14.3 million). In 2015, 247 000 died because of diabetes (122 100 men and 125 400 women). Over 42.7% of these deaths occurred in people under 60 years. Diabetes health spending in the region was estimated at between USD34.6 billion and USD 59.9 billion, accounting for 5.0% of the global total. Brazil spent at least USD21.8 billion on people with diabetes.

Furthermore, the South East-Asia (SEA) region, consisting of India, Sri Lanka, Bangladesh, Bhutan, and Mauritius and Maldives, was home to more than 72 million adults with diabetes in 2013 and is expected to exceed 123 million in 2035, with a substantial number (24.3million) of people also having impaired glucose tolerance (IGT) (IDF, 2015). A sharp increase in the prevalence has been observed in the SEA Region, both in urban and rural areas, which is mostly associated with the lifestyle transitions towards urbanization and industrialization. The highest prevalence in the South East-Asia region is found in Mauritius (14.8%) and followed by India (9.1%). Study have highlighted that Asian Indians have a higher risk of diabetes compared to other Asians, whether they are living in their land of birth or in an affluent foreign country (Ramachandran et. al., 2010).

According to IDF (2015), in the Western Pacific region, 9.3% of adults aged 20-69 were estimated to be living with diabetes in 2015, equivalent to 153 million people. Over half (52.1%) of these were undiagnosed. 61.6% live in the cities and 90.2% live in low or middle-income countries. The Western Pacific is home to 36.9% of the total number of people with diabetes in the world, with China having the highest number of people (110 million) with diabetes in the world. 1.9 Million Deaths occurred among adults, the highest number of deaths in the IDF region. Over 44.9% of diabetes deaths occurred in people under the age of 60 years.

2.3 Diabetes Prevalence in Africa

The Africa Region (AFR), where diabetes was once rare, has witnessed an increase in the prevalence (Peer et al., 2013). As mentioned by Levitt (2008), diabetes is an increasing problem in sub-Sahara Africa with Type 2 the most common form. The number of people suffering from diabetes in Africa remains uncertain, although as stated by IDF (2015) estimate from 2000 put the figure at 7.5 million diabetic adults between 20 and 79 years of age. Likewise, it is estimated that the diabetes population will double over the next twenty-five years in Africa (IDF 2015). This raises enormous health care questions, as all African countries are already struggling to cope with the diabetes burden. Awareness is regarded as being poor, and the concentrations of the disease vary considerably between different ethnic groups. Type 1 diabetes, although still rare in many areas, is becoming increasingly more prevalent. IGT is also becoming problematic and exceeds 30 per cent in many African countries.

As stated by Motala (2002), Ogurtsova et al. (2017), Peer et al. (2014), IDF (2015), and WHO (2014), in the world, Africa is the region with the lowest prevalence of diabetes (4.9%), having Réunion (15.4%), Seychelles (12.1%) and Gabon (10.7%) as the top three countries with highest prevalence and 10 out of 48 countries with prevalence of diabetes higher than the upper quartile (6.3%). The only high-income countries in the region are Equatorial Guinea and Seychelles, both of which have GNIs of over ID22 000 per capita. The Central African Republic has the world's lowest GNI of ID610 per capita.

According to Motala (2002), Motala et al. (2003a), and IDF (2015), an estimated 14.2 million adults aged 20-79 have diabetes in the Africa Region, representing a regional prevalence of 2.1-6.7%. This is expected to increase to 34.2 million adults by 2040, more than double the number in 2015. The Africa Region has the highest proportion of undiagnosed diabetes – over two thirds (66.7%) of people with diabetes are unaware they have the disease. The majority (58.8%) of people with diabetes live in cities, even though the population in the region is predominantly (61.3%) rural.

As stated by IDF (2015), some of Africa's most populous countries have the highest number of people with diabetes, including South Africa (2.3 million), Democratic Republic of Congo (1.8 million), Nigeria (1.6 million), and Ethiopia (1.3 Million). Nearly half of all adults with diabetes in the region live in these four countries. Increase in urbanization and population age pose an ever-growing threat of type 2 diabetes in this region.

In 2015, The Africa Region accounted for 0.5% health expenditure on diabetes with over USD 3.4 billion spent on health care with a projection of USD 5.5 billion in 2040, the lowest of any region (IDF 2015). This is equivalent to 7% of the region's total health budget per person with diabetes per year. An estimated number of 321,121 deaths was recorded due to diabetes with 79% under age 60 years old. This shows that investment, research and health systems are slow to respond to this burden and primary focus has remain on infectious disease. From the Africa population of 441 million, an estimated 14.2 million number of adult age 20-79 is living with diabetes with a projection of 34.2 million in 2040, making a regional prevalence of 3.2% and a projection of 3.7% in 2040. Among the IDF regions, Africa region with 49 diverse sub-Sahara countries and territories has the highest proportion of undiagnosed diabetes; with over two third (66.7%) of people with diabetes are unaware they have the disease. Majority of people (58.8%) with diabetes lives in the cities, even though the population in the region is predominantly rural (61.3%) (IDF Atlas 2015).

However, in 2013, the African adult population (aged 20–79 years) was around 374 million, and the prevalence of diabetes in this age group was estimated to be around 9.2%, equivalent to 34.6 million people with diabetes. This figure is expected to increase to 67.9 million by 2035 or 11.6% of the adult population. Of the 34 million people affected by diabetes, nearly 17 million were estimated to be undiagnosed and therefore at considerable risk of diabetes complications and poor health outcomes (IDF 2013). The prevalence is rising due to the rising rate of obesity, physical inactivity and urbanisation. As mentioned by Levitt (2008), the current morbidity rate of diabetes in Africa is primarily due to high rate of microvascular complication.

2.4 Evolution of Diabetes Prevalence in Africa

Diabetes was regarded as a rare disease in Sub-Sahara Africa (SSA) prior to the 1990s (Mbanya and Ramaiya 2006). Since the 1990s, demographic and epidemiological transitions, as well as urbanization, have rendered diabetes one of the Non-Communicable Disease burdens in SSA. There are 10.4 million individuals with diabetes in SSA, representing 4.2% of the global population with diabetes (Kengne et. al., 2013).

Studies by Tobin et. al., (2015) show that the trends in diabetes prevalence have paralleled high rates of overweight and obesity, which are leading risk factors for the development of diabetes. As stated by Kengne et.al., (2013) by 2025, it is estimated that the prevalence of diabetes will increase by 80% to reach 18.7 million in this region, with a higher prevalence in the urban areas. Among the aging population coupled with rapid urbanization, is expected to lead to the increasing prevalence of diabetes in SSA.

2.5 Risk Factor of Type 2 Diabetes

The diabetes epidemic is rising due to rapid globalization and urbanization, increase in sedentary lifestyle, and changes in diet in line with the worldwide rise in obesity and overweight. Peer et al (2014), classified factors that drive the development of type 2 diabetes into two: a complex gene-environment interaction of non-modifiable (genetics, age, gender, ethnicity and family history) and modifiable risk factors which include obesity, excessive alcohol ingestion, poor dietary habits and physical inactivity. The non-modifiable risk factors include aging and genetic predisposition. Similarly, Whiting and colleagues (2003) noted that the contextual, clinical, and health system challenges to the delivery of health care for diabetes in Africa is influenced by several factors, including poor patient attendance at health clinics, short consultation time with physicians (leaving little or no time for patient education), inadequate staff, limited staff training, poor control of blood glucose and blood pressure, inadequate referral systems, and almost non-existent patient education. Some of these factors in the context of African countries are discussed briefly in this section.

2.5.1 Demographics

According to IDF (2009), the population of sub-Saharan Africa is set to grow from around 860 million in 2010 to more than 1.3 billion by 2030. For age groups above 40 years, the population size will double. People aged 45-59 years are 8.5 times more likely to develop diabetes than those aged 15-29 years, and those above the age of 60 are 12.5 times more likely to develop diabetes. Based on the present prevalence rates in sub-Saharan Africa, demographic changes alone will account for an increase of 9.5 million people with diabetes between 2010 and 2030 (Azevedo and Alla (2008); Sobngwi et al. (2012)).

The age of onset of type 2 diabetes is also decreasing in sub-Saharan Africa. Peak occurrence of type 2 diabetes is between the ages of 20 and 44, already 40 years lower than the peak age of occurrence in high-income countries, and it is predicted to fall further. Studies from Levitt et.al., (1999), Omar et. al. (1994), Ramaiya et. al. (1991), and McLarty et. al. (1989) showed that the highest diabetes prevalence is in people of Indian origin, followed by native Africans.

According to Jamison D.T, et al (2006), age and ethnicity are the two main non-modifiable risk factors of diabetes in Africa. In most developed communities, the peak of occurrence falls in the age group of 65 years or older, whereas in developing countries it is in the age group 45 to 64, and in Sub-Saharan Africa, it is in the age groups 20 to 44 and 45 to 64 years.

A strong relationship has been found between population age and diabetes. Diabetes has also been found to increase the risk of dementia (Shaw et al., 2010, Booth et al., 2006, Biessels et al., 2002, Wang et al., 2014). Studies show that diabetes is rising globally, particularly in Africa due to population ageing and rapid urbanisation of (Hall et al., 2011, Assah et al., 2011, Mbanya et al., 2010, Kanmogne et al., 2010, Lim et al., 2012, De Ramirez et al., 2010, Wang et al., 2014). Since the risk of developing diabetes increases with age, the global aging of the population, especially Africa, is a major driver of the global rise in diabetes. According to IDF (2015), by 2035, diabetes peak in Africa is expected to be in the oldest individuals.

Despite the HIV epidemic, the number of people with diabetes in Africa is expected to grow because of changing demography (Levitt, 2008). Therefore, a concerted multi-sectorial effort is essential in ensuring improvement in health care delivery for people with diabetes in Africa.

2.5.2 Poverty

According to the World Bank (2015), poverty (otherwise known as extreme or absolute poverty) is defined as living with an income of less than $1.90 per day. It is measured in the international dollar, with adjustment on purchasing power parity (PPP) adjustment. It is calculated and measured as the ratio of the share of the population living below a certain poverty line over the total number of population.

Studies from Indian and America shows that poverty and depression has an impact and increase diabetes (Ramachandran (2002), Levine (2011), De Groot (2003). Holtgrave and Crosby (2006) in an exploratory study found that there is a strong correlation between poverty and diabetes. Everson et al. (2002) found a similar pattern of epidemiological evidence for the relationship between socio-economic status, depression, obesity and diabetes.

2.5.3 Urbanization

The Africa region is currently experiencing the most rapid increase in urbanization and changes in lifestyle (Motala, (2002); (Assah et al., 2011)). The continent is undergoing urbanization faster than any other region, with the urban population growing at an average annual rate of 4.5% (Parnell and Pieterse, 2014). Studies have also shown a global rise in diabetes prevalence, particularly in Africa at large, due to population aging and rapid urbanization. Studies on the African region show that diabetes is more prevalent in urban compared to rural areas (Mbanya et al., 2010, Motala, 2002, Motala et al., 2003a, Peer et al., 2014). According to Godfrey and Julien (2005), a major factor increasing diabetes prevalence in Africa is urbanisation, which is because of continuous movement of people from rural to urban areas, particularly in the sub-Saharan Africa. Studies have showed that high prevalence of diabetes is associated with urbanisation lifestyle (Ramachandran et al., 2008, Ramachandran et al., 1999, Esteghamati et al., 2008, Whiting et al., 2011). This migration is inevitably associated with a shift in lifestyle from a relatively healthy traditional pattern to the urban scenario of increased food quantity and reduced quality, low level of exercise, smoking, and increased alcohol availability.

It is predicted that 45% of the population will live in urban areas in 2025 (UN, 2008). This urbanization process has a significant impact on the prevalence of diabetes, as urban residents have a 1.5 to 4 times higher prevalence of diabetes than rural residents (Jamison et. al., 2006). Currently, 68% of people with diabetes in sub-Saharan Africa live in urban areas, and this number is expected to increase to

78% in 2030. Type 2 diabetes in sub-Saharan Africa primarily affects the poorest people living in urban areas Azevedo and Alla (2008), (Sobngwi et al., 2012, Mbanya et al., 2010). Increase in urbanization and population age will pose an ever-growing threat in the increase of diabetes. Ramachandran et. al. (2007) has shown that urbanization plays a significant role in increasing the burden of cardiovascular diseases, a class of diseases which diabetes falls under.

2.5.4 Overweight/Obesity

Driven by rapid globalization and urbanization with subsequent changes in diet and the adoption of a sedentary lifestyle, the diabetes epidemic has expanded in line with the worldwide increase in overweight and obesity (Assah e., al. 2009). Being overweight or obese massively increases the risk of diabetes (Mbanya et. al., 2010).

Studies by Assah et., al. (2009) and Rotimi et. al. (1995), described overweight/obesity as a major and well-known modifiable risk factor for diabetes. A high growing prevalence of overweight and obesity has been observed in SSA. For example, in a meta-analysis of obesity among West African populations, the prevalence of obesity was 10.0% (95% CI, 6.0-15.0) (Abubakari and Bhopal, 2008; Abubakari et. al., 2008). A study in Benin found that abdominal obesity was positively associated with increased probability of metabolic syndrome (Ntandou et. al., 2009). Across many sub-Saharan African countries, obesity has been linked to both urban residence and wealth – the more wealth a person has, the more likely he or she is to be overweight or obese due to nutritional transition (Ntandou et. al., 2009), transitions in energy expenditure due to urbanization, and other unknown factors. As stated by Sobngwi et. al. (2004) in a study which explored the effects of lifetime exposure to an urban environment in Cameroon in relation to obesity and other cardiovascular risk factors, urbanization is associated with a drastic decrease in physical activity and changes in dietary habits. Mbanya et. al. (2006), Dugas et. al. (2009), Njelekela (2009), Holdsworth (2006), and Duda et., al. (2007) have found that the prevalence of obesity is mostly among women in West Africa, South Africa, and Tanzania. However, in South Africa, it has been reported that 61% of the population is overweight and obese (SAMRC, 2014), with an estimate of 2 million people having diabetes (Adele and Fiona 2014). In addition, a complex set of cultural, psychosocial, and biological factors influences the maintenance of healthy weight (Scott et. al., 2013; Popkin, 2006), particularly in the Africa region where access to food remains a daily challenge.

This shift in dietary habits in the last decade has been driven by a change in lifestyle and socio-economic status. These changes have affected diet, physical activity and ultimately the prevalence of obesity, leading to an increase in both diabetes and IGT.

2.5.5 Diet and Alcohol, Tobacco and CVD Risk

Because of foreign direct investment from transnational food companies, processed foods have become easily available, contributing to the rise of diabetes. According to Peer et.al. (2014), Tuei et. al. (2010), and Madu et al. (2003), the intake of unhealthy foods that are high in fat, refined carbohydrates, and sugar contributes to increased energy imbalance and subsequent obesity and diabetes.

Likewise, studies have shown that alcohol consumption is also correlated with an increased risk of glucose intolerance (GI) and diabetes (Peer et. al., 2014; Tuei et. al, 2010). However, a few studies have reported that alcohol was independently related to diabetes in rural South Africa (Motala et., al. 2008), and in Kenya, frequent alcohol intake in men was associated with glucose intolerance (Christensen et. al, 2009).

Some studies have revealed that an emerging diabetes risk factor is associated with smoking and psychosocial factors, environmental pollutants, and biomarkers of metabolic pathways (Tamayo et. al., 2013, Rotella & Mannucci, 2013). Prospective studies also observed that exposure to components of traffic and industry-related air pollutants such as Nitrogen Oxide (NO_2) and fine particulate matter increases the risk of type 2 diabetes (Rjagopalan & Brook, 2012). Therefore, an increase in the level of alcohol consumption, unhealthy dieting, and smoking are contributing factors to higher diabetes prevalence in Africa.

2.5.6 Physical Activity

A physically inactive lifestyle is prevalent in the Africa region (Peer el., al 2014), and can be attributed to rapid urbanization and transition in socio-economic life (Tuei et. al., 2010, Unwin et. al., 2010). According to WHO, insufficient physical activity is defined as less than 150 minutes' moderate physical activity per week or its equivalent, which is present in about a quarter of men and a third of women in Africa region (WHO, 2011). Studies have shown that there is a significant relationship between physical inactivity and diabetes and obesity (Sobngwi et. al., 2002, Assah et al. (2011)). Physical activity is more common in rural than urban regions of Africa because rural populations rely on walking for transport and often have intense agricultural activities as their main occupation. In Sub-Saharan Africa, walking time and pace is drastically reduced (by factors of 2 to 4 for walking at a slow pace and 6 to more than 10 for walking at a brisk pace) in an urban community as compared with a rural community. Abubakari et., al. (2009) stated that one in seven people in West Africa is physically inactive. Some studies report that up to three-quarters of urban residents are sedentary in their daily occupation and half of government workers do not undertake any leisure time physical activity. An elevated level of physical inactivity increases the risk of developing diabetes (Tuei et. al., 2010). A world health survey data indicated that 7-19% of the population in West Africa countries is physically inactive. Studies from Cameroon, Kenya and in sub-Saharan Africa have found that higher physical activity levels are inversely related to abnormal glucose tolerance (Assah et. al., 2011; Christensen, 2009; Aspray et. al., 2009).

2.5.7 Access to Care

The healthcare system in Africa is often challenged by insufficient resources to provide adequate patient care. Both lack of institutional resources and up-to-date practical information for health care providers often jeopardize patient care. A review by Motala (2002) noted that the increasing diabetes trends in Africa are influenced by inadequate health care infrastructure, inadequate supply of medications, and lack of available healthcare facilities and providers.

Watkins et al. (2001) suggest that the management of chronic disorders such as diabetes in rural African communities could be improved by decentralizing care to local village healthcare facilities to improve access to treatment and reduce mortality. This proved to be effective in improving diabetes control in a rural Ethiopian village. Gil et al. (2008) attributed the lack of glycemic control among diabetics in rural Ethiopia to geographically scattered populations, shortage of drugs, and insulin. Also, a lack of diabetes team care is a major factor behind these serious issues of diabetic control and complications.

2.6. Spatial Clustering of diabetes

Understanding the local and global pattern of diabetes prevalence and its associated socio0economic risk factor can assist in the design of location specific intervention. Technology development such as GIS and the advancement of spatial statistics have allowed the application of not only disease mapping but also spatial analysis, such as spatial clustering in epidemiological research (Green et., al. 2003, Jemal et., al. 2002). Recent advances in spatial analytical techniques such as exploratory data analysis implemented in Geo Data software and statistical software like R provide the capability to uncover spatial clusters (Anselin et., al 2007). These techniques have been used for spatial analysis on different disease such as obesity, physical activity, cancer, the pattern for prescribing cardiovascular drugs, malaria, and sexually transmitted disease (Goovaerts, 2010; Michimi and Wimberly (2010); Schuurman 2009, Chen and Wen, 2010).

Several studies have used spatial clustering methods to identified clusters of deaths due to NCD, such as cancer in developed country, for example Rosenger et., Jemal et., al 2002. There are few studies of spatial clustering of diabetes that has been conducted using United State county level data (Shrestha et., al. 2013, Hipp and Chalise (2016), Gartner et., al. 2016). Several studies have been carried out on different diseases and especially infectious, NCD and high mortality disease, to assist public health planning and providing location specific intervention. For example in Denmark spatial cluster of leukaemia and type 1 diabetes has been studied (Schmiedel et., al. 2011), also studies on spatial clustering in some Africa countries and sub-Sahara countries do exist on different diseases such as on cholera in Ghana (Frank et., al. 2000), on HIV mortality in South Africa (Namoshan et., al. 2013) and also on tuberculosis in as urban West Africa country (Touray et., al. 2010). As mentioned by Namoshan et., al. 2013, there are several examples of work that have used spatial clustering methods to help understand the epidemiology of infectious disease in rural Africa settings. For example, study by Nomashan et., al. (2013) demonstrated substantial geographical heterogeneity in the prevalence of HIV infection in the rural area of Mthubathuba in KwaZulu-Natal province in South Africa. As cited by Namoshan et., al 2013, (Snow et., al. 1999) demonstrated a space time clustering of seven childhood morbidity on the Kenyan coast with seasonal peaks in incidence of severe malaria. Similar studies on micro epidemiology of malaria were done in Nouna, a rural area in Burkina Faso (Sankoh et., al. 2001) and the results are likely to help better understand the observed clustering of mortality in the area. However, we are aware of no previous studies on spatial clustering of diabetes and its correlates in Africa as a continent using meta-data. Therefore, for a NCD related cases like diabetes, spatial clustering has not been previously done both at the local level of individual countries or at the national level on the continent, this has necessitated the study.

2.6 Conclusion

This chapter reviewed previous studies on diabetes by various researchers on the risk factors and the prevalence of diabetes. From all the past studies reviewed, it was found that risk factors such as population age, obesity, alcohol consumption, urbanization, physical inactivity, and access to care, all have a significant association with diabetes prevalence. This justifies the selection of these variables as risk factors for diabetes.

Chapter 3

Methodology

3.0 Introduction

This chapter introduces the data set and the methodologies used in achieving the objectives. Some simple descriptive statistics were conducted to explore the relationship that exists between each independent variable and diabetes. Other statistical techniques such as cluster analysis, spatial analysis, classical and Bayesian statistical methods were also used.

3.1 Data Building

The dataset used in this study is a collection of data from previous studies. Most of the variables included in the data are from the World Bank, derived from a UN source (see Appendix A). The summary table of definitions and sources of each indicator are shown in Appendix B. The diabetes prevalence (20-69 years ages) variables for 2015 were extracted from World Bank data. The next section explains the statistical methodology used for the data analysis.

3.2. Exploratory data analysis

3.2.1: Shapiro–Wilk Normality Test

This is a test of normality in frequentist statistics. It whether a sample $y_1, \dots, y_n$ came from a normally distributed population. The test statistics is

$$W = \frac{\left(\sum_{i=1}^{n} b_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.0}$$

Where $y_{(i)}$ is the *ith* order statistics, $\bar{y}$ is the sample mean, and b_i is the constant. The null hypothesis of this test is that the population is normally distributed ($H_o\colon P_i \sim N$). Thus, the null hypothesis is rejected if the *p*-value is less than the chosen alpha level, and we conclude that the samples are not from a normally distributed population. On the contrary, if the *p*-value is greater than the chosen alpha level (in this study we choose $\alpha = 0.05$), then the null hypothesis that the data came from a normally distributed population cannot be rejected.

3.2.2. Spearman Rho Correlation Test

Spearman's rho, denoted as ρ (rho) or (*r*), is a non-parametric measure of statistical dependency (degree of association) between two variables. It is used to assess the relationship between two variables. Spearman's test is like the Pearson correlation, but while Pearson correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). The formula is given as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3.1}$$

Where ρ is the spearman correlation, d_i the difference between the ranks corresponding values X_i and Y_i and n is the number of value in each data set.

3.2.3. Ordinary Least Square Regression (OLS)

Ordinary Least Square (OLS) linear regression is a statistical technique used for the analysis and modelling of linear relationships between a response variable and one or more predictor variables. If the relationship between two variables appears to be linear, then a straight line can be fit to the data to model the relationship. For one of this study objective of determining the relationship between diabetes prevalence (response variables) and the risk factors (predictors). The linear equation or equation for a straight line for a bivariate regression takes the following form:

$$y = mx + c$$

where y , is the response (dependent) variable, is m the gradient (slope), x is the predictor

(Independent) variable and c is the intercept. The modelling application of OLS linear regression allows one to predict the value of the response variable for varying inputs of the predictor variable given the slope and intercept coefficients of the line of best fit

3.2.4. Geographic Weighted Regression (GWR)

Geographically weighted regression (GWR) is a spatial analysis technique that takes non-stationary variables or data into consideration and then model the local relationships between the predictors and an outcome of interest (Brunsdon et., al. 1996, Brunsdon et., al. 1998, Fotheringham et., al. 2002, Goovaerts, 2008).

Consider a global regression model written as:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \qquad (3.2.1)$$

GWR is an extension of the traditional regression model, which allow local parameter rather than global parameters to be estimated so that the model above equation 3.2.1 can be rewritten as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k (u_i, v_i) x_{ik} + \varepsilon_i \qquad (3.2.2)$$

Where ((u_i, v_i) denotes the coordinates of the ith point in space and $\beta_k (u_i v_i)$ is a realisation of the continuous function $\beta_k(u, v)$ at point i. That is, we allow a continuous surface of parameter values, and the measurements of this surface are taken at certain points to denote the spatial variability of the surface. In this case equation 3.2.1 is the special case of equation 3.2.2 in which the parameters are assumed to be spatially invariant. Thus, the GWR in equation 3.2.2 recognises that spatial variations in relationship might exist and therefore provides a way in which they can be measured.

GWR is an outgrowth of ordinary least square regression (OLS) that add a level of sophistication by allowing the relationship between the independent and dependent variables to vary by locality

(Wheeler and Paez 2010). GWR allows visualization of stimulus response relationships and if that relationship varies in space and it also accounts for spatial autocorrelation of variables (Mitchell, 2012, Fotherigham et., al. 2002). In summary, GWR constructs a separate OLS equation for every location in the data set, which incorporates the dependent and explanatory variables of locations falling within the bandwidth of each target location. Bandwidth can be manually entered by the user with the option of selecting the default or adaptive.

Also, GWR can be applied under the assumptions that the strength and direction of the relationship between a dependent variable and its predictors may be modified by contextual factors (Mitchell 2012). However, GWR has some limitations, which include the problem of multicollinearity and the approaches to calculation goodness of fit statistics (Charlton and Fotheringham 2009, Leung et., al. 2000, Wheeler and Tiefelsdorf 2005)

3.3. Cluster Analysis

3.3.1. Definition

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) based on a set of measured variables into several different groups such that similar subjects are placed in the same group (Cornish R., 2000). Cluster analysis groups objects (countries) based on the information found in the data describing the objects or their relationship. It is a convenient method for identifying a homogenous group of objects called clusters. The clusuters in this study are African countries. The aim is that countries in the same group will be similar (or related) to one another and different from (or unrelated) to those in other groups. As stated by Tango (2010), clusters should exhibit high internal homogeneity and high external heterogeneity. The objectives of cluster analysis are exploratoration and relationship identification which are not possible with observed data. This similarity measure can be done either by distance, correlation or association.

3.3.2. Cluster Criteria for Continuous Data

One of the possible ways of extracting information from data is to find homogeneity clusters or similar units (Romesburg, 2004). Therefore, an informative partitioning of the objects will produce groups of similar patterns.

As stated by (Tango, 2010), the most common clustering criteria derived from an *n x p* matrix *X* of continuous data makes use of a decomposition of the *p x p* dispersion matrix *G*, given by:

$$G = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (X_{m\mathrm{l}} - \bar{X})\,(X_{ml} - \bar{X})' \qquad (3.3)$$

Where X_{ml} is the p-dimensional vector of observation of the lth object in group *m*, and $\bar{X}$ is the *p*-dimensional vector of overall sample means of each variable. The total dispersion matrix can then be partitioned into the within-group dispersion matrix, denoted as:

$$T = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (X_{ml} - \bar{X}_m)\,(X_{\mathrm{ml}} - \bar{X}_m\,)' \qquad (3.4)$$

Where $\bar{X}_m$ is the *p*-dimensional vector of sampled means within group m and the between-group dispersion matrix, denoted as:

$$W = \sum_{m=1}^{g} n_m\,(\bar{X}_m - \bar{X})\,(\bar{X}_m - \bar{X}\,)' \qquad (3.5)$$

So that we have,

$$G = T + W \qquad (3.6)$$

Because of this partitioning, the within-group and between-group variation needs to be minimised through taking into consideration the Euclidean distance between individual and group means, given as:

$$E = \sum_{m=1}^{g}\sum_{l=1}^{n_m}(X_{ml} - \bar{X}_m)\,(X_{\mathrm{ml}} - \bar{X}_m\,)'$$

$$\sum_{m=1}^{g}\sum_{l=1}^{n_m} d_{ml,m}^2 \tag{3.7}$$

Where $d_{ml,m}$is the Euclidean distance between the lth individual in the mth group and the mean of the mth group, also derived based on distance matrix:

$$E = \sum_{m=1}^{g}\frac{1}{2n}\sum_{l=1}^{n_m}\sum_{v=1}^{n_m} d_{ml,mv}^2 \tag{3.8}$$

Where $d_{ml,mv}$ is the Euclidean distance between the lth and in vth individual in the mth group.

3.4. Clustering Techniques

In this section, we explain the most renowned clustering methods, types of clustering and the techniques in clustering. The main reason for having many clustering methods is the fact that the notion of “cluster” can be defined differently or measured differently or apprehended in different angles (Estivill-Castro, 2000). Consequently, many clustering methods have been developed, each of which uses a different induction principle. Therefore, a precise definition of “similarity” (or “closeness “, or proximity”) is required in cluster analysis. However, it is natural to employ the familiar concept of distance when the grouping is based on variables. Therefore, this study groups the African countries by analyzing the similarity in the pattern of diabetes and associated predictors.

Farley and Raftery (1998) suggest dividing the clustering methods into two main groups: hierarchical and non-hierarchical (or partitioning) methods. Nonetheless, this study uses both methods as suggested by (Romesburg, 2004). The most commonly used non-hierarchical method is the k-mean method (Tango, 2010) which is also used in this study.

The hierarchical method constructs the clusters by recursively partitioning the instances in an either top-down or bottom up fashion. This method can be subdivided as agglomerative or divisive. The agglomerative method is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The divisive method, on the other hand, is a top down approach where all observations start in one cluster and splits are performed concurrently as one moves down the hierarchy. The result of hierarchical clustering is a dendrogram representing the nested grouping of objects and similarity levels at which groupings change. Clustering is obtained by cutting the dendrogram at the desired similarity level. The emerging division of clusters is performed according to some similarity measure, chosen to optimise some criterion. The hierarchical clustering methods could be further divided according to how the similarity measure is calculated. These are single-link, complete-link and average-link clustering (Jain et. al., 1999).

Partitioning or non-hierarchical methods relocate instances or objects by moving them from one cluster to another, starting from an initial partitioning. This method requires that the number of clusters be pre-set by the user. Diverse types of partitioning methods include error-minimizing algorithms, graphic-theoretic, density-based, or model-based methods, grid-based methods and soft-computing methods. The method of soft-computing is classified into k-means, simulated annealing (SA), genetic algorithm (GA), and centroid methods.

In choosing which technique to use, an empirical study of K-Means, GA, and SA was presented by Al-Sultan and Khan (1996). GA and SA were judged comparable in terms of solution quality, and all were better than K-mean. However, the K-mean method is the most efficient in terms of execution time, as against other methods.

3.4.2. K-Means Clustering

This is a prototype-based, non-hierarchical clustering technique that attempts to find a user-specified number of clusters (*k*), which are represented by their centroids (Tango, 2010). Prototype-based clustering techniques create a one-level partitioning of the data objects. There are many of such techniques, but two most used are K-means and K-medoid. The K-mean defines a prototype in terms of a centroid, which is usually the mean of a group of points and is typically applied to objects in a continuous *n*-dimensional space. This study uses K-means, which is one of the oldest and most widely-used clustering algorithms (Tango, 2010).

The aim of K-means clustering is to partition an observation into *k* clusters. Therefore, the observation is said to belong to the cluster with the nearest mean which serves as a prototype of the cluster. The strengths and weaknesses of K-means clustering are that it is efficient, simple to apply, and can be used for a wide variety of data types. However, it cannot handle non-globular clusters or clusters of varied sizes and densities, although it can typically find pure sub-clusters if a large number of clusters is specified. In addition, the K-means technique has trouble clustering data that contains outliers. Outlier detection and removal can be of significant help in such cases. However, K-mean is restricted to data for which there is a notion of a center (centroid). A related technique, K-medoid clustering, does not have such restrictions but is very expensive.

Clustering can be treated as an optimization problem. One way to solve this problem is to find a global optimum. This is done by enumerating all possible ways of dividing the points into clusters and then choose the set of clusters that best satisfies the objective function, for example, minimising the total sum of squares (SSE). This approach is computationally intractable, and so a more practical approach is needed. One approach to optimizing a global aim function is to rely on algorithms that are often good but not optimal. An example of this approach is the K-means clustering algorithm which tries to minimize the sum of the squared distances (error) between objects and their cluster centers. The assumption here is that the data is one-dimensional, that is, $dis\,(x, y) = (x - y)^2$, which does not essentially change anything, but greatly simplifies the notation.

3.4.3. K-Means Algorithm

The K-means algorithm is simple, and the first step is to choose the initial centroid, where *K* is a user-specific parameter, namely the number of desired clusters. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. The assignment is repeated and update steps until no points change clusters or until the centroid remains the same.

The K-means algorithm is described as:

1. Select *K* points as initial centroids
2. Repeat
3. Form *K* clusters by assigning each point to its closest centroid
4. Recompute the centroid of each cluster
5. Until centroids do not change

For some combination of proximity functions and types of centroids, K-means usually converges to a solution, that is, K-means reaches a state in which no points shift from one cluster to another, and hence, the centroid is stable. Because most of the convergence occurs at the initial steps, however, the condition (step 5) of the algorithm is often replaced by a weaker condition, for example, repeat until only 1% of the points change clusters.

In assigning a point to the closest centroid, a proximity measure that will quantify the notion of "closest" is needed for the specific data under consideration. Euclidean (L_2) distance is often used in Euclidean space, while cosine similarity is more appropriate for documents. However, there are several types of proximity measures that are appropriate for a given type of data. For example, Manhattan (L_2) distance can be used for Euclidean data, while the Jaccard measure is often used for document.

Since the algorithm repeatedly calculates the similarity of each point to each centroid, the similarity measures used in K-mean are relatively simple. Consider some data whose proximity measure is Euclidean distance. The objective function which measures the quality of clustering uses the sum of squared error (SSE), also known as scattering. The error of each data point is calculated, that is, its Euclidean distance to the closest centroid, and then the total sum of squared errors is computed.

3.4.4. K-Means as an Algorithm to Minimise the Sum of Squared Errors (SSE)

This section shows how the centroid for the K-means algorithm can be mathematically derived when the proximity function is Euclidean distance and the objective is to minimise the SSE. Specifically, we investigate how we can best update a cluster centroid to minimise the cluster's SSE. Consider the geometrical representation of the object set ϑ by a cloud of points designated as

$$N(\mathrm{I}) = \{o_i \mid i \in I\} \tag{3.7}$$

Equally weighted in space $\mathbb{R}^p$ and endowed with the usual Euclidean distance measure.

The initial state of the algorithm can be either a partition

$$P^0 = \{ p_1^0, p_2^0, \dots \dots p_k^o, \dots \dots, p_K^o \} \tag{3.10}$$

of ϑ or a set

$$G^0 = \{ g_1^0, g_2^o, \dots \dots g_k^o, \dots \dots, g_K^o \} \tag{3.11}$$

of K distinct points, which define the K attraction centers. This can be determined at random or from expert knowledge. Without loss of generality, it is assumed that we start with $G^0 = \{ g_1^0, g_2^o, \dots \dots g_k^o, \dots \dots, g_K^o \}$

In satisfying the proximity requirement, the first encountered in selected. ϑ is then divided in K cluster, denoted as:

$$P^0 = \{ p_1^0, p_2^0, \dots \dots p_k^o, \dots \dots, p_K^o \}$$

Partition obtained.

Steinley (2003), suggested the procedure in k-means methods as follows:

Step 1: randomly partition the object into k clusters (specify the number of clusters to work with).

Step 2: calculate initial centroid from this partition (mean within a cluster is calculated).

Step 3: assign each data point to its nearest centroid (iteratively, to minimise the within-cluster) until no major differences are found.

Step 4: also, a further use of locally optimal k solution can be defined by combining information regarding the cluster problem (number of variables, number of clusters)

Step 6: finally, from the number of *k* defined, empirically derive the cluster membership.

3.4.5. Hierarchical method (Ward's and Centroid method)

The five most popular hierarchical methods of clustering are single linkage, complete linkage, average linkage, centroid methods, Ward's method and Mahalanobis distance (Kaufman and Rousseeuw, 2009).

Ward's method simply joins clusters with a small number of observations and is strongly biased toward producing clusters with the same shape and width, and roughly the same number of observations. The aim in Ward's method is to join cases into clusters such that the variance within a cluster is minimised. This is done such that each case begins as its own cluster. Clusters are then merged in such a way as to reduce the variability within a cluster. In this method, at each stage, the average similarity of the cluster is measured, and the difference between each case within a cluster and the average similarity is calculated and squared. Therefore, the sum of square deviation is used as a measure of error within a cluster. The result produces both a dendrogram and other statistical analysis results which are shown in Chapter Four.

Generally, hierarchical cluster analysis provides an excellent framework with which to compare any set of cluster solutions. The method is used in predicting and judging the number of clusters that should be retained or considered.

This study uses K-mean, Ward's and Centroid methods of cluster analysis to compare the results. The results of these methods are shown in Chapter Four.

3.4.6. Conclusion

Generally, K-means aims at partitioning a given data set into distinct, exclusive clusters so that the data points in each group are quite like each other. The underlying idea of K-means clustering is that the cluster that contains the smallest possible within-cluster variation of all the observation in relation to each other is said to be a good cluster. The most common way of defining the variation is by using the Euclidean or Mahalanobis distance (to compensate for correlated variables).

3.5. Spatial Analysis

In this section, we discuss the spatial configuration/distribution of diabetes prevalence through clustering with the aim of showing the pattern of a disease cluster. The section is divided into three sub-sections. The first section explains the importance of investigating disease clustering and its classification. The second and third sections describe the statistical methods for computing spatial analysis of a geographical distribution area. The subsequent sections deal with the estimation of the parameters.

3.5.2. Introduction

In epidemiology studies, it is important to evaluate and investigate whether a disease is randomly distributed or clustered over time and/or space after adjusting for known confounding factors that may contribute to the study of the disease (Tango, 2010). Furthermore, since the 1980s, interest has grown in the study of relationship between risk of a disease and proximity of residence to the source of the disease. In 1990, “Guidelines for investigating clusters of health events” was issued by the Centers for Disease Control and Prevention. A statistical summary of methods used in assessing clusters of health events was provided as a resource for researchers and investigator as a basis for analysis. This section introduces statistical methods for detecting disease clustering which are widely used in the literature.

3.5.3. Classification of Disease Clustering

According to Tango (2010), disease clustering is classified into one of three groups: temporal, spatial or space-time clustering. Temporal clustering examines the question of whether cases tend to be located close to each other in time (Kaplan et. al., 1982), while spatial clustering examines the question of whether cases tend to be located close to each other in space. Space-time clustering examines the question of whether cases that are close in space are also close in time (Tango, 2010).

To investigate whether clustering is real and significant, many different tests have been proposed for different purposes. Besag and Newell (1991) classified these tests into

a) General tests (designed for investigating the question of whether clustering occurs over the study region), and
b) Focused tests (designed for assessing the clustering around a prefixed point) (Michelozzi et. al., 2002).

General tests were further classified by Kulldorff (1988) into

a) Global clustering tests: designed for evaluating cases which tend to come in groups or are located close to each other no matter when and where they occur.
b) Cluster detection: designed both for detecting localized clusters and evaluating their significance ((Arslan et al., 2013, Xiang and Song, 2016, Yazdy et al., 2015, Condoleo et al., 2016, Gartner et al., 2016).

In general, the type of data used for cluster investigation can be classified into two types:

a) individual geographical location data on coordinates of disease or incidence by residence, usually by street address, zip code or post code unit;
b) regional count data, which includes number of cases and population at risk in specific areas. However, due to the restriction of clinical confidentiality, obtaining individual data can be impossible.

This study considered the problem of detecting disease clustering in space based on regional count data (diabetes prevalence data), which is illustrated in the section below. In this section, firstly, we describe the construction of neighbors and spatial weights which are a prerequisite in determining the spatial autocorrelation, which is an important aspect in spatial statistics that measure the similarity (correlation) between nearby observations. We then go into ways of measuring spatial autocorrelation.

3.5.4. Spatial Neighbours and Weights

3.5.4.1. Spatial Neighbours

This is the first step in spatial autocorrelation. It simply identifies which regions or states are neighbors of a given region or state. Here, the first step is to define which relationships between observations are to be given a non-zero weight, i.e. choose the neighbor criterion to be used. Secondly, assign weights to the identified neighbor links. This takes the form of a square symmetric *R* X *R* matrix with (*i, j*) element equal to 1 if regions *i* and *j* are neighbors of one another (or spatially related) and zero otherwise.

Conventionally, the diagonal matrixes are set to zero. The matrix can be constructed using different approaches such as linear continuity, rook contiguity, queen contiguity and distance base (Renard, 2011).

3.5.4.2. Spatial Weights

A slight transformation of a spatial neighbour is referred to as a spatial weight. The most common transformation is called row standardization, in which the rows of the neighbour's matrix are made to sum to unity.

Let W with elements $\widetilde{w_{ij}}$ be a spatial neighbour matrix. To row-standardize, we divide each element in a row by the sum of the elements in the row. This gives a spatial weights matrix W, with element w_{ij}, which can be defined by:

$$w_{ij=} \frac{\widetilde{w}_{ij}}{\sum_j \widetilde{w}_{ij}} \tag{3.12}$$

The library package "spdep" was used in R for the analysis of the nb2listw function.

3.5.5. Spatial Autocorrelation Tests

Spatial autocorrelation tests are one of the many tests which examine the entire population area for spatial autocorrelation, assuming the spatial process is the same everywhere. Spatial autocorrelation measures how much close objects are in comparison with other close objects. It helps to understand the degree to which one object is similar to other nearby objects. (Fischer and Getis, 2009, Pfeiffer et al., 2008, Olsen et al., 2010).

One of the reasons why spatial autocorrelation is important is because statistics relies on observations being independent from one another. Therefore, if autocorrelation exists in a map, then there is a violation of the facts that observations are independent from one another. Since observations may not be independent when made at various locations, measurements may be closer in value when made at the nearby location than measurements made at a location that is far apart. This phenomenon is referred to as spatial autocorrelation. Spatial autocorrelation measures the correlation of a variable with itself through space. Also, spatial autocorrelation indicates if there is clustering or dispersion in a map. (Druck et al., 2004, Fotheringham and Rogerson, 2013, Fischer and Getis, 2009).

Spatial autocorrelation is defined as the coincidence between similarity in value location (Fischer and Getis, 2009). Statistically, this can be explained as, if y_i and y_j are realizations of random variable y, indexed by spatial locations, then we have spatial autocorrelation if

$$Corr\ (y_i, y_j)\ = E\ (y_i y_j) - E(y_i)E(y_j)\ \ \neq 0 \tag{3.13}$$

A positive spatial autocorrelation is said to occur when similar values occur near one another (or say when similar values cluster together in a map), while a negative spatial autocorrelation occurs when dissimilar values occur near one another (or say cluster together in a map). As stated by (Renard, 2011), the test statistics usually used for spatial autocorrelation are the Moran Index, Geary's C, Gestis-Ord G, and Mantel test. This study uses only Moran Indices due to its popularity. It is discussed in the next section.

3.5.5.1. Moran Indices Test

Moran’s I indices were developed by Patrick Moran in 1950 as a tool for spatial exploratory data analysis. It is a diagnostic statistical tool used to detect spatial autocorrelation in a dataset, given that one has a weight matrix w with entries w_{ij} representing distances between observations X_i and X_j (Elliot et al., 2000).

It follows the basic form for Global Moran Index (GMI) with similarity between regions *i* and *j* defined as the product of the respective difference between $X_i \ and \ X_j$. The statistic is defined for a data vector *Y* by dividing the basic form by the sample variance. This tool has been subsequently used in almost all studies employing spatial autocorrelation (Arslan et al., 2013, Xiang and Song, 2016).

Moran I take the form of a correlation coefficient where the mean of a variable is subtracted from each sample value in the numerator. It is calculated as a ratio of the product of the variable of interest and its spatial lag, with the cross-product of the variable of interest and adjusted for the spatial weights used. This results in coefficients ranging from (-1) to (+1), where a value between (0) and (+1) indicates a positive association between variables, a value between (0) and (-1) indicates negative association, and (0) indicates no correlation between variables. Therefore, a positive spatial autocorrelation is said to occur when Moran’s I is close to +1 and if close to -1, then we say a negative spatial autocorrelation occurs. While a positive Moran’s I hints at data is clustered, a negative Moran’s I implies data is dispersed. (Fischer and Getis, 2009, Fotheringham and Rogerson, 2013).

In the study of spatial patterns, it may logically follow that close observation should or is likely to be similar than those far apart. It is usual to associate a weight to each pair x_i and x_j which quantifies it. In its simplest form the weight usually takes the values 1 for close neighbours, and 0 otherwise. We set $w_{ii} = 0$, and the weights are sometimes referred to as a neighbouring function. The formula for Moran I is given as:

$$\mathrm{I(d)} = \frac{\frac{1}{n}\sum_i^n \llbracket \sum_j^n w_{ij} \, (x_i - \bar{x})\big(x_j - \bar{x}\big)\big)}{\frac{1}{n}\sum_i^n (x_i - \bar{x})^2} \tag{3.14}$$

where I($_d$) = Moran correlation coefficient as a function of distance

w_{ij} = spatial weights of the link between i and j, a matrix of weighted values

1 = x_i and $x_j$are within a given distance class, for $x_i \neq x_j$

x_i and $x_{j=}$ value of variable at location i and j

$\bar{x}$ = the mean of variable of interest

n = sample size, and S_0 is the sum of all ${w_{(ij)}}^2$

therefore, equation 3.14 can also be rewritten as equation 3.15 below:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2} \tag{3.15}$$

Though Moran I is similar to Pearson correlation, it is not bounded on [-1, 1] because of the spatial weights. In R, the function to compute Moran I is “Moran. Test”.

3.5.6. General Test for Detecting Spatial Clustering (Moran I Indices and Local Indexes)

3.5.6.1. Moran I Indices Spatial Clustering

GMI is a spatial statistics tool used to describe the overall spatial pattern of attributes over definite geography. In this case, GMI is used to determine the spatial correlation of diabetes prevalence among the entire population across the Africa continent (Aselin, 1996; Daniella et., al., 2016). The goal of GMI in spatial autocorrelation is to summarize the degree to which similar observation tend to occur near each other. That is, we want to check if the same pattern (of diabetes) or process occurs over the entire geographic area. However, it is good to know that global statistics only suggests that there is clustering but does not identify the areas of clusters. Therefore, GMI is often used first to determine if there is an evidence of spatial association. Therefore, the similarities of values at location B_i and B_j

are weighted by the proximity of i and j. The weight w_{ij} defines proximity. The extent of similarity is represented by the weighted average of similarity between the populations.

Thus, the GMI tool is an inferential statistic which provides a test of the null hypothesis, which states that the spatial distribution shows a complete randomness of the attributes being studied. That is, the attribute value of the location does not depend on the values of neighboring locations.

3.5.6.2. Monte Carlo Simulation

There is currently no simple and systematic way and method of significance testing in spatial clustering, these methods are normal permutation and Monte carlo simulation. Monte Carlo simulation was used in order to minimise and understand error in the map. This is a permutation test for Moran I statistics, calculated by using random permutations for the given spatial weighting scheme, to establish the rank of the observed statistics in relation to the n_{sim} simulated values.

This approach repeats randomization of the observation a large number of times (for example $N_{sim} \sim 999$). A Moran's I statistic is calculated for each randomization and can be compared to the Moran's I. The function "moran.mc" is used to carry out the test in R.

3.5.6.3. Moran Correlograms

Spatial correlograms are used to examine the patterns in spatial autocorrelation. The correlograms of the Moran's I statistic are used to determine the appropriate number of neighbors or distance. They are used to show how correlated pairs of spatial observations are when we increase the distance (lag) between them. This approach is used to calculate Moran's I based on the number of *k*'s for a range of *k* lags (also known as spatial lag).

3.5.6.4. Local Indexes of Spatial Autocorrelation (LISA)

Global statistics can establish if there is clustering, but do not identify the areas of particular clusters. In spatial analysis, the global test is often used first to determine if there is evidence of spatial association. Once that is detected and established, it is necessary to detect local areas of similar values. This is done by estimating local statistics such as Local Indexes of Spatial Autocorrelation (LISA).

The aim of the LISA test is generally to detect local areas of similar values. It therefore requires local statistics. LISA tests are decompositions of global indicators into the contribution of each individual observation, indicating the extent of significant spatial clustering of similar values located around that observation. LISA can be used to detect clusters, it allow for a classification of the significant locations as high-high and low-low spatial clusters and high-low and low-high spatial outliers. In this case, a reference to high and low is relative to the mean of the variable and should not be interpreted in an absolute sense. Usually four types of spatially associated cluster can be identified. These are:

High-High (denoted as H-H) : this is a cluster region where those above mean prevalence are neigboured by other similar countries

Low-Low (denoted as L-L) : this is the cluster states where those with below mean prevalence is neigboured by other similar countries

Low-High (denoted as L-H and High –Low (denoted as H-L) : are cluster countries where countries of isolated low and high prevalence with similar neighbors).

Similarly, from Local Moran's I also known as Hot spot concentration is on the spatial concentration of high-low and low-high values that is the spatial outliers

As a result, the sum of LISA's for all observation is proportional to the Global Moran I, denoted as:

$$I_t = \sum_t \sum_{j=1}^{T} w_{ij} z_{ij} \tag{3.16}$$

where z_{ij} is the elements of a spatial contiguity matrix. Under the null hypothesis of no spatial association, the moment for I_t statistics can be derived for a randomnisation hypothesis.

Generally, with LISAs, each test gives an indication of the extent of significant spatial clustering of similar values located around that observation of interest. Therefore, the location of a particular observation is identified and defined as a neighborhood, and then formalized with the spatial adjacency weight matrix, *W*. However, it should be noted that *W* can be based on the sharing distance from one location to another or share a border in full or partially.

Waller and Gotway (2004), (Elliott and Wartenberg, 2004) describe spatial outliers as a spatially referenced object whose non-spatial attributes are significantly different from those of other spatially referenced objects in its spatial neighbourhood. The two types of spatial outliers are multi-dimensional space-based outliers which use Euclidean distances to define their spatial neighbourhoods and graph-based outliers which use graph connectivity.

The results of the LISA maps clusters which is used to identify the outliers (hotspots) is shown in chapter four.

3.4.6. Conclusion

In investigating spatial correlation, spatial autocorrelation test is suggested using GMI. However, with limitation of GMI which only shows the global pattern, additional method of Moran I was used to investigate spatial pattern as the regional level which helps in detecting the similarity function of the prevalence among the regions. There is a need to identify the outliers (hotspots) countries in other to help formulate network collaboration in the control over diabetes.

3.6. Distribution for Count data

3.6.2. Modelling Count Data

Most analyses of public health data involve disease counts, proportions, or rates as outcome variables rather than continuous outcomes (Waller and Gotway, 2004). Spatial analysis often focuses on counts from geographical areas with relatively few subjects at risk, and few expected cases. Such cases require modeling appropriate count outcomes. Count data are typically used to model the number of occurrence of an event within a fixed period. Examples of count data may include:

- The number of goals scored by a team
- The number of murders in a city

3.6.3. The Poisson distribution

The most popular distribution used in modeling count data is the Poisson distribution, which is derived under three assumptions:

1. The probability of one event happening in a short interval is proportional to the length of the interval.
2. The number of events in non-overlapping intervals is independent, and
3. The probability of two events happening in a short interval is negligible in comparison to the probability of a single event happening.

The probability mass function of the Poisson (λ) distribution is

$$\Pr(Y = y) \; = \; \frac{e^{-\lambda} \, \lambda^{y}}{y!} \tag{3.17}$$

For $\lambda > 0$, The mean and variance of a Poisson distribution are shown as

$$E(Y) = Var(Y) = \lambda \tag{3.18}$$

The assumption of a Poisson distribution having mean equal to its variance is known as the equidispersion assumption. This assumption implies that the Poisson distribution does not allow for the variance to be adjusted independently of the mean. Likewise, in the presence of underdispersion data (variance is less than the mean) or over-dispersion data (variance is greater than the mean), the Poisson distribution is not an appropriate model and one has to use another parametric model, with an additional parameter compared to Poisson. In the case of over-dispersed data, one of the distributions that can provide a better fit is the negative binomial distribution.

3.6.4. The Negative Binomial Distribution

The probability mass function of the negative binomial (μ, k) distribution is

$$P(Y = y|\mu, k) = \; k^{y}(1-k)^{\mu}\binom{y+\mu-1}{y} \quad \text{y=0,1,2, …} \tag{3.19}$$

the mean and variance of an NB(μ, k) are respectively

$$E[Y] = \frac{k\mu}{(1-k)'}$$

$$(3.20)$$

$$V[Y] = \frac{k\mu}{(1-k)^{2}}$$

An alternative formulation of the negative binomial distribution is

$$P(Y = y|\mu, k) = \frac{\Gamma\left(\frac{1}{k}+y\right)}{\Gamma\left(\frac{1}{k}\right)y!}\left(\frac{k\mu}{1+k\mu}\right)^{y}\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}} \tag{3.212}$$

Thus, the mean and variance for the NB (μ, k) is

$$E[Y] = \; \mu$$

$$(3.22)$$

$$V[Y] = \mu + \; k\mu^{2}$$

The first parameter in equation 3.20 of the formulation is the mean of the distribution, whereas the second (equation 3.22) is referred to as the dispersion parameter. A large value of *k* is a sign of over-dispersion, but when $k \rightarrow 0$, the variance of the distribution in equation 3.22 is equal to the mean, and that gives us a special case of Poisson.

3.6.5. The Fractional Probit Model

The Fractional Probit Model was introduced by Bliss Chaster (Bliss C.I 1934) as a fast method for computing maximum likelihood estimates as proposed by Ronald Fisher. The model is a type of regression model where the dependent variable can take only two values (say 0 or 1). The aim of this is to estimate the probability that observations with a characteristic will fall into a specific one of the categories (0 or 1).

Suppose we have a continuous dependent variable *y* in (0,1) and a vector of independent variables (x). Assume we want to fit a regression to the mean of *y* conditional on *x*, that is, *E (y | x)*, and because *y* is in [0,1], we want to restrict that *E (y | x)* is also in [0,1]. Thus, the fractional Probit objective is achieved by using the following model:

- Probit: $E\ (y|x) = \emptyset\ (x, \beta)$
- Heteroskedastic Probit: $E\ (y|x) = \emptyset(x\beta \exp(Z\gamma))$
- Logit: $E(y|x) = \exp(x\beta)\,/(1 + \exp(x\beta))$

The fractional regression implements quasi-likelihood estimators, which implies that there is no need to know the distribution to obtain consistent parameter estimates, that there is need for the correct specification of the conditional mean, and that by default, a robust standard error is computed.

3.6.6. Regression Model for Count Data

3.6.6.1. Linear Regression Model

In classical terms, linear regression is also known as a least square regression model. This is a statistical method that allows us to summarize and study the relationship between two continuous variables. One variable, denoted as *X*, is regarded as a predictor, explanatory, or independent variable, and the other, denoted as *Y*, is regarded as the response, outcome, or dependent variable. A linear regression model with a single predictor variable is known as a simple linear regression model (Kutner et al., 2004, Seber and Lee, 2012, Montgomery et al., 2015, Williams, 1959).

The simple linear regression model is written as:

$$Y = X\beta + \epsilon \tag{3.23}$$

where

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} \cdots & x_{1k} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

Y is the response variable, β_0 is the intercept, β is the vector of the coefficient estimates of the random variables (an unknown pparameter that needs to be estimated) and ε is the error term (or residual) used to capture the deviation of the data from the model. The aim is to find the values for the parameters β_a $\{a = 1, 2, \ldots\ldots k\}$ which would provide a best fit for the data. The regression must follow the following assumptions: linear relationship, independence of errors with each other and the covariates, multivariate normality, no or little multicollinearity, no autocorrelation and homoscedasticity (errors must have zero mean and constant variance) (Kutner et al., 2004, Seber and Lee, 2012, Montgomery et al., 2015).

Applying the zero-mean assumption of the errors in equation 3.23, the expectation of the random matrix is defined as:

$$E[Y] = [E\{Y_{ij}\}] \tag{3.24}$$

Where $i = (1, \ldots\ldots n; j = 1, \ldots \ldots p)$. Least squares regression describes the behaviour of the location of the conditional distribution, using the mean of the distribution to represent its central tendency. The residual ε_i are defined as the difference between the observed and the estimated values. Minimizing the sum of the square residuals,

$$\sum_{i=1}^{n} r(Y - X^T\hat{\beta}) = \sum_{i=1}^{n} (Y - X^T\hat{\beta})^2 \tag{3.25}$$

Where $r(\mu) = \mu^2$ is the quadratic loss function, which gives the least squares estimator $\hat{\beta}$ by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad where\ X \neq X^T \tag{3.36}$$

Also, the additional assumption that the errors (residual) ε_i follow a Gaussian distribution

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2 I_n) \tag{3.27}$$

Where I_n is the *n x n* identity matrix, it provides a framework for testing the significance of the coefficient found in equation 3.26. Under this assumption, the least square estimator is also the maximum likelihood estimator. By taking the expectations, with respect to ε_i in equations 3.25.and 3.26, as well as noting that the linear function of a normally distributed random variable is normally distributed itself, we can rewrite the model in 3.26 as:

$$X \sim N(\mu, \sigma^2 I_n), \qquad where\ \mu = X^T \beta \tag{3.28}$$

Therefore, the model in 3.26 represents the relationship between the mean of $y_{i.}$ for $i = 1,2, \dots n$, and the covariates linearly.

3.6.6.2. Generalized Linear Model

The family of generalized linear models (GLMs) provides a collection of models extending basic concepts from linear regression to applications where error terms follow any wide range of distributions, including binomial and Poisson for modeling count data (Waller and Gotway, 2004).

Thus, equation 3.28 refers to data that are normally distributed, but can be generalized to any distribution of the exponential family (Nelder and Baker, 1972, McCullagh and Nelder, 1989). GLMs consist of three components:

1. a probability distribution that belongs to the exponential family of distributions (known as a random component which defines the distribution of error terms)
2. a linear predictor $\rho_{i=} \ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \ = X^T \hat{\beta}$ (also known as a systematic component defining the linear combination of explanatory variables) and

3. A link function which defines the relationship between the systematic and random components, given as: $E[Y] = \mu_i = \vartheta^{-1}(\rho_i)$.

GLM parameters generally require an iterative procedure rather than closed-form solutions for linear models (McCullagh and Nelder, 1989, Waller and Gotway, 2004). A GLM can be used for data that are not normally distributed and for cases where the relationship between the mean of the response variable and the covariates is not linear. The GLM includes many important distributions such as Gaussian, Poisson, Gamma, and Inverse-Gamma (Cameron and Trivedi, 2013, Kutner et al., 2004, Seber and Lee, 2012, Montgomery et al., 2015, Waller and Gotway, 2004).

3.6.2.1. Poisson regression

This is a special case of GLM commonly used to model count data. Poisson regression has been used for modelling count data in many fields such as public health (Arslan et al., 2013, Duncan et al., 2002, Xiang and Song, 2016) , Epidemiology (Best et al., 2000, Frome and Checkoway, 1985, Zou, 2004, Gartner et al., 2016, Hanewinckel et al., 2010, Sobngwi et al., 2001), insurance (Boucher and Denuit, 2006, Christiansen and Morris, 1997, Ismail and Jemain, 2007) and many other research areas. The canonical link function is logarithm.

The model is specified as:

$$\Pr(Y = y) = \frac{e^{-\lambda}\,\lambda^{y}}{y!} \tag{3.29}$$

For $\lambda > 0$, the mean and variance of a poison distribution are shown as

$$E(Y) = Var(Y) = \lambda \tag{3.30}$$

The likelihood function is given as

$$L(\beta|y,x) = \prod_{1}^{N} \Pr(y_i|u_i) = \prod_{i=1}^{N} \frac{\exp(-u_i)\, {u_i}^{y_i}}{y!} \qquad (3.31)$$

The assumptions are:

Response *Y* has a poison distribution, $Y \sim Pois(\lambda), (E(Y) \ = \lambda, and \ \ Var(\,Y) = \lambda$

With the assumption that the mean is equal to the variance, any factor that affects one will affect the other. This poses a problem when the data exhibits a different behavior. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data (Preston, 2005). Statistically, an important motivation for the Poisson distribution, however, lies in the relationship between the mean and the variance. Most of the proposed approaches to this problem focus on over-dispersion (Ismail and Jemain, 2007, Berk and MacDonald, 2008).

3.6.2.2. Negative binomial

One of the ways to handle the situation posed by a Poisson regression is to fit a parametric model that is more dispersed than Poisson. A natural choice is the negative binomial (NB), given as:

$$P(Y = y|\mu_i, k) = \frac{\Gamma\left(\frac{1}{k} + y_i\right)}{\Gamma\left(\frac{1}{k}\right) y_i!} \left(\frac{k\mu_i}{1 + k\mu_i}\right)^{y} \left(\frac{1}{1 + k\mu_i}\right)^{\frac{1}{k}}, \qquad (3.42)$$

$$\log\{\mu_i\} = \ X^T \beta \qquad (3.33)$$

Where the parameters $\mu_i \ and \ k$ represent the mean and the dispersion of the negative binomial. The respective mean and variance of this model are:

$$E[Y_i] = \exp\{X^T\beta\}, \qquad E[Y_i] = \exp\{X^T\beta\}, \tag{3.34}$$

$$V[Y_i] = \exp\{X^T\beta\} + \; k\, exp\{X^T\beta\}^2 \tag{3.35}$$

The variance of a negative binomial is a quadratic function of its mean. The negative binomial approaches the Poisson $(\mu_i\,)$ model for $k \to 0$.

The negative binomial PDF can be described as the probability of observing y failures before *kth* success in a series of Bernoulli trials. Under such description, r is a positive integer (Hilbe, 2011). However, there is no compelling mathematical reason to limit this parameter to integers.

Negative binomial is a generalization of Poisson regression. It loosens the highly restricted assumption that the variance is equal to the mean. this is based on the mixture of the poison -gamma mixture distribution. This model is popular because it models the Poison heterogeneity with a gamma distribution.

Given the negative binomial PDF with parameter ((μ, k):

$$f(y|\,\mu, k) = \binom{y_i + k - 1}{k - 1} \mu_i^k\,(1 - \mu_i)^{y_i} \tag{3.36}$$

Or

$$f\;(y|\,\mu, k) = \frac{(y_i + k - 1)}{y_i!\;(k - 1)!}\,\mu_i^k\,(1 - \mu_i)^{y_i} \tag{3.37}$$

Converting the NB PDF into exponential family form results in

$$(y|\, \mu, k) = \exp\left\{y_i\, ln(1-\mu_i)\ + k\ln(\mu_i) + \ln\binom{y_i+k-1}{k-1}\right\} \tag{3.38}$$

Where

$\exp\{y_i\, ln(1-\mu_i)\}$ is the link, and $k\ln(\mu_i) + \ln\binom{y_i+k-1}{k-1}$ is the cumulant.

Thus, the canonical link and cumulant can easily be extracted from a PDF when expressed in exponential family form. This gives:

$$\theta_i\ =\ \ln{(1-\ \mu_i)} \rightarrow \mu_i = 1 - exp(\theta)_i \tag{3.59}$$

$$b(\theta_{i)}\ =\ -\mathrm{k}\ln\mu_i \rightarrow -k\ (1-\exp(\theta_i)$$

$$=\ \alpha_i\emptyset(scale) = 1$$

Therefore, the first and second derivates, with respect to θ, respectively yield the mean and variance functions, given as:

$$\mathrm{E}[Y] = b'\ (\theta_i) = \frac{\delta b}{\delta k_i}\frac{\delta k_i}{\delta\theta_i} = \ -\frac{r}{k_i}\big(-\ (1-k_i)\big) = \frac{r(1-k_i)}{k_i} = \ \mu_i$$

$$\mathrm{V[Y]}= b''(\theta_i) = \frac{\delta^2 b}{\delta k_i^2}\left(\frac{\delta k_i}{\delta\theta_i}\right)^2 + \frac{\delta b}{\delta k_i}\frac{\delta^2 k_i}{\delta\theta_i^2} = \frac{\mu}{k_i^2}\ (1-k_i)^2 + \frac{-\mu}{k_i}\ \big(-(1+k_i)\big)$$

$$= \frac{r\ (1-k_i)}{k_i^2} \tag{3.40}$$

$V(\mu)$ therefore equals $r(1-k)/k^2$. Assume we now parameterize k and μ in terms of π and γ

$$(1 - k_i/(\gamma k_i) = \pi_i \tag{3.41}$$

$$(1 - k_i/(k_i) = \gamma\pi_i$$

$$k_i = 1/(1 + \gamma\pi_i) \tag{3.42}$$

Where $\gamma = 1/r$ given the defined values of $\mu\, and\, \gamma$. negative binomial PDF can then be re-parametrization such that

$$f(y|\,\pi, \gamma) = \begin{pmatrix} y_i + \frac{1}{\gamma} - 1 \\ \gamma - 1 \end{pmatrix} \left(\frac{1}{1 +\, \gamma\pi_i}\right)^{\frac{1}{\gamma}} \left(\frac{\gamma\pi_i}{1 + \gamma\pi_i}\right)^{y_i} \tag{3.43}$$

This is identical to equation 3.59, which we derived via Poisson-gamma mixture.

3.7. Principles of Bayesian Methods

The Bayesian approach to univariate and multivariate linear regression with normal errors has long been of interest in areas such as econometrics and epidemiology (Koop, 2003; Poirier, 1995; Zellner, 1971). The Bayesian statistical model incorporates prior non-information about the unknown parameters. Incorporating such prior knowledge into statistical analysis of diabetes prevalence has the potential of enhancing the quality of the statistical results.

As described by Lesaffre & Lawson (2012), Rafia & Schlaifer (1961), Bolstad (2007), Bernado & Smith (1994), and Congdon (2006), the fundamental idea of Bayesian analysis framework is that `the unknown model parameters are the random variables, hence specified by prior distribution.'

The measures of degree of belief is interpreted as the probability statement. The underlying Bayesian analysis is from Bayes' theorem, which is used to revive the beliefs about the parameters considering the observed data to obtain the posterior distribution. The posterior distribution gives relative weights to each parameter value after analyzing the sample data.

The relationship between the prior distribution, the observed data, and the posterior distribution is expressed as *posterior = prior X likelihood* (Press, 1989; Bolstad, 2007; Gill, 2009; Congdon, 2006; Gellman & Rubin, 1996).

The Bayesian approach has several features and advantages as compared to frequentist (or classical) statistical analysis. It allows a consistent way of modifying one's belief about the parameter given the data that occurred, which implies that inference is based on actual data and not on all possible data sets that might have occurred (as in the case of the frequentist approach), as presented by Rao (2011), Raiffa & Schlaifer (1961), and Bolstad (2007). Likewise, according to Congdon (2006) and Lesaffre & Lawson (2012), it provides a way of formalizing the process of learning from data to update beliefs in accord with recent notions of knowledge synthesis. It can also assess the probability on both nested and non-nested models (unlike the classical approach). Using modern sampling methods, it readily adapts to complex random effect models that are more difficult to fit using classical methods (Carlin et al., 2001; Gilks et. al., 1993). The approach also provides a way of improving estimation in sparse data sets by borrowing strength (Richardson & Best 2003; Stroud, 1994). Bayesian methods may also improve on classical estimators in terms of the precision of the estimate (Congdon 2006). Allowing a parameter to be a random variable enables one to make a probability statement about it (the parameter) posterior to observing the data. However, the use of both classical and Bayesian approach has gained popularity in many practical analyses in which design-based inferences can be derived from the Bayesian perspective, using classical models with noninformative prior distribution.

However, the arguments and debate around the drawbacks and benefits of the classical and Bayesian approach in statistical analysis emanate mainly from the differing fundamental interpretation of probability. As described in Bolstad (2007), classical methods define probability as long-run tendencies of events that eventually converge on some true population proportion, whereas Bayesian interprets probability as a “degree of belief”.

The Bayesian philosophy implies that prior distribution is descriptions of relative likelihoods of events based on past knowledge, personal intuition, expert opinion, or from the posterior distribution. The prior knowledge being updated by conditioning on newly observed data.

It is argued that the Bayesian analysis paradigm works than a frequentist approach in biological, health and social science research (Bostad 2007; Press, 1989). This is mainly due to the availability of immense prior data information on most of the phenomena in these fields.

Suppose we have an observable random vector y with probability mass (or density for continuous) function $(y|\theta)$, where θ denotes an unobservable parameter. The Bayes' theorem as stated by Press (1989), asserts that the probability function of θ , for a given value of y os expressed as

$$p(\theta|y) = \begin{cases} \dfrac{f\ (y|\theta)g(\theta)}{\sum_{\theta} f(y|\theta)g(\theta)} & for\ a\ discrete\ parameter \\ \\ \dfrac{f(y|\theta)g(\theta)}{\int f(y|\theta)g(\theta)\, d\theta} & for\ a\ continous\ parameter \qquad . \end{cases} \tag{3.44}$$

Here, $p(\theta|y)$ is called the posterior probability function of θ given the observed data, and $g(\theta)$ is the prior probability function θ. This study has continuous parameters; therefore, the focus is on Bayesian inference for continuous data. Therefore, the equation 3.42 can be rewritten considering only the continuous parameters, given by Bayes theorem:

$$p(\theta|y) \ = \frac{p(y|\theta)g(\theta)}{p(y)} \tag{3.45}$$

where the normalizing constant is

$$p(y) \ = \ \int p(y|\theta)\ g(\theta)\, d\theta$$

Equation 3.45. Is the marginal probability of the observed data given the model that is the likelihood and the prior while ignoring the constant gives:

$$p(\theta|y) \quad \propto \quad p(y|\theta)\ X\ g(\theta) \tag{3.66}$$

more colloquially,

$$posterior \propto Likelihood\ X\ prior$$

where ∝ denotes the proportionality and $p\ (y|\theta)$ denotes the likelihood function $g\ (\theta)$ is the prior and $p\ (\theta|y)$ is the posterior distribution.

3.7.2. The Likelihood

Fisher (1992) first introduced the concept of likelihood, expressed the plausibility of the observed data given as a function of the parameters of a stochastic model. Likelihood contain information provided by the observed sample.

Let us consider observed data denoted as $y = (y_i, \dots \dots .., y_n)$, and an unobservable parameter θ, the likelihood function, denoted by $p\ (y|\theta)$ is given as

$$p(y|\theta) = \prod_{i=1}^{n} p(y|\theta) \tag{3.47}$$

However, the likelihood function can be viewed as representing the plausibility of θ considering the data, where the value of θ that maximizes $p(y|\theta)$ is called the maximum likelihood estimate (MLE). The observed data y affects the posterior distribution through $p(y|\theta)$. It is necassry to note that the bayesian inference is based on the assigned probability due to the observed data, and not due to other imaginary data (in the classical case) that might have been observed. Therefore, adherence to the likelihood principle implies that inferences are conditional on observed data since the likelihood function is parameterized by the data.

3.7.3. Choice of Prior Distribution

In Bayesian analysis, incorporating prior information is argued to make the approach more suitable to much empirical research (Lesaffre & Lawson 2012; Press, 1989). Often, priors are expressed probabilistically or by using a distribution by which their parameters are called hyperparameters.

The pivotal point in the Bayesian paradigm is the choice of a prior, with an option of choosing between informative and non-informative (Lesaffre & Lawson, 2012; Press, 1989). The term "prior" means that the prior knowledge should be specified independently of the collected data (Cox, 1999; Lesaffre & Lawson 2006). In modeling the data, at first the prior could represent lack of information (non-informative), but in the second step, the prior can be informative. However, the subjectivity surrounding the choice of a prior, and the problems stemming from model misspecification provide a basis for most of the critics of Bayes methods (De Finetti, 1937; Raiffa & Sclaifer, 1961).

The choice of a prior distribution in Bayesian inference is significant. In Bayesian methods, the prior distributions generate the most controversial issue. This choice of the prior may affect Bayesian estimation, as a strong prior can have a profound influence on Bayesian estimation. Therefore, in the absence of any prior information from any solid and scientific evidence or sources, one may use non-informative priors. This will be further discussed in the next section.

The Bayesian technique is flexible compared to classical techniques because the prior can be informative or non-informative. An informative prior distribution describes the available (prior) knowledge about the model parameters. Some prior knowledge is available basically in all research. An informative prior is one that summarises the evidence about the parameters concerned from various sources and usually has a significant impact on the results. Prior information provides specific and definite information about a parameter, which usually comes in form of historical data mainly from past studies and expert knowledge (Lesaffre & Lawson, 2012; Congdon, 2006; Spielgelhalter et. al., 2003; Vail et. al., 2001). Box & Tiao (1973) argue that we are almost never in the state of ignorance. The challenge is to incorporate the available little prior knowledge into a probabilistic framework. Therefore, the question of whether to include the beliefs of one expert or a community of experts forms the main question in Bayesian analysis.

3.7.4. The Prior Distribution

As stated earlier, the Bayesian approach is based on a prior belief about the parameter of interest. However, this prior belief can be described in terms of a density function, which is then referred to as prior distribution. When the sample size is low or poor, the prior distribution will dominate the analysis. In contrast, for a large sample size, the likelihood function will effectively contribute to the relative risk estimation.

In the Bayesian paradigm, the parameters are stochastic, and thus can be assigned a prior distribution accordingly. Thus, the prior distribution is a distribution assigned to a given parameter θ before the observed data x_i. Given a single parameter, the prior distribution can be denoted as $P(\theta)$. However, for a parameter vector, similar notation can be used for the joint distribution provided θ is well understood to be a vector parameter.

The prior distribution has different properties. One of such properties is that, it can be improper or proper. A prior distribution is said to be an improper prior under the condition where the integration of the prior distribution of a random variable θ over its range Ω is infinite (that is, the normalising constant is infinite) ((Lesaffre and Lawson, 2012); (Lawson, 2013). This is denoted mathematically as:

$$\int_{\Omega} P(\theta)d\theta = \infty \tag{3.48}$$

As noted by Lesaffre and Lawson (2012), the improper prior is a limitation of any prior distribution, but it is not necessarily being the case that some improper prior leads to the same property in the posterior distribution.

3.7.5. Non-Informative Priors

A non-informative prior, also referred to as a vague, subjective, objective or reference prior, is defined by Tiao and Box (1973), (Bernardo and Rueda, 2002, Congdon, 2014) as one that provides little or no information relative to the experiment or expresses a minimal effect relative to the data. A non-informative prior is often regarded as a formal representation of ignorance. Non-informative priors that are not a distribution or that have an infinite area under the curve are called improper priors (Lesaffre & Lawson, 2012; Condgon, 2006; Press, 1989; Bernardo & Smith, 1994).

As mentioned earlier, the choice of a non-informative prior distribution can depend upon that the prior distribution of a variance parameters must have a positive real line. Such possible cases are often seen in the gamma, inverse gamma or uniform distribution families.

3.7.6. Conjugate Prior

Similarly, a combination of a prior distribution and likelihood function can lead to the same distribution family in the posterior $P(\theta|y)$ for a prior distribution $P(\theta)$. Thus, such prior and posterior distributions are referred to as a conjugate distribution, and for that likelihood, the prior is called a conjugate prior. For example, the Poisson likelihood with parameter θ and the gamma prior distribution for θ is also a gamma distribution. The same holds for the binomial likelihood, and the beta prior distribution for the probability of success, as well as the normal data likelihood with a normal prior distribution for the mean (Lesaffre and Lawson, 2012).

All members of the exponential family have conjugate priors. By examining the kernel of the prior likelihood product, the conjugacy prior can be identified. However, the prior likelihood must be the same for the prior distribution. In affirmation to this, Lesaffre and Lawson (2012) stated often time conjugacy always guarantees a proper posterior distribution.

3.7.6.1. **Jefferey's Prior**

The Jefferey's prior is a non-informative prior distribution on parameter θ that is proportional to the square root of the determinant of the Fisher information $I(\theta)$. Thus, Jefferey's prior for a single parameter θ is defined as:

$$P(\theta) \propto \sqrt{I\ (\theta)} \text{ where } I\ (\theta) = -E\left[\tfrac{d^2\ \ln P(y|\theta)}{d\theta^2}\right] \tag{3.49}$$

Hence, the Jeffery's prior can a further be defined as

$$P(\theta) \propto \sqrt{I\ (\theta)} = \sqrt{\left[\left(\frac{d}{d\theta} \ln P(\ y|\ \theta)\right)^2\right]} \tag{3.50}$$

To find such vague or flat prior for a given distribution, Jeffreys prior was developed.

From the equation 3.50 above, the Jeffrey's prior for the Poisson mean θ of the positive real value can be defined and denoted as:

$$P(\theta) = \sqrt{\sum_{x=0}^{+\infty} P\ (\ y\ |\ \theta)\ \left(\frac{y-\ \theta}{\theta}\right)^2\ \frac{1}{\sqrt{\theta}}} \tag{3.51}$$

Therefore, the prior is known to be improper and not non-informative (Lawson, 2013).

Also, the Jeffreys prior for the normal distribution of the positive real value X with fixed mean θ is:

$$P(\theta) = \sqrt{\int_{-\infty}^{+\infty} P\ (\ y\ |\theta\ )\ \left(\frac{y-\theta}{\sigma^2}\right)^2\ dy} = \sqrt{\frac{\sigma^2}{\sigma^4}} \propto 1 \tag{3.52}$$

The prior does not depend upon θ. Likewise, the Jeffreys prior for the normal distribution of the positive real value y with standard deviation, $\sigma > 0$ is:

$$P(\sigma) = \sqrt{\int_{-\infty}^{+\infty} P\,(\,y\,|\theta\,)\left(\frac{(y-\ \theta)^2 y - \sigma^2}{\sigma^3}\right)^2 dy} \;=\; \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma} \qquad (3.73)$$

Also, using equation 3.53, Jeffery's prior for binomial likelihood with parameter θ and sample size n is denoted as:

$$P\,(\,\theta) = \sqrt{\frac{n}{\theta(1-\theta}} = n^{1/2}\,\theta^{-1/2}\,(1-\theta)^{-1/2} \qquad (3.54)$$

Thus, equation 3.53 above can be written as $P\,(\,\theta) \propto \theta^{-1/2}\,(1-\theta)^{-1/2}$, where $\theta \sim Beta\left(\frac{1}{2},\frac{1}{2}\right)$. Hence, the data formation helps in determining the prior in equation 3.53, but not the real observed data, since we are averaging over y.

However, it should be noted that it is sometimes more important to be more informative with the prior distribution if the likelihood function has little information about the identification of the parameter. In this case, identification can only come from the prior specification. Therefore, Lesaffre and Lawson (2012) described identification as an issue relating to the ability to differentiate between parameters and with a parametric model.

3.7.6.2. **The Posterior Distribution**

In the Bayesian approach, the posterior distribution contains all the information of interest, as it combines the prior information and the likelihood of the data. After the data has been observed and the prior assumptions have been made, the outcome is the posterior distribution which describes the behaviour of the parameter of interest.

The posterior distribution is obtained from the product of the likelihood and the prior distribution, denoted mathematically as:

$$P(\theta \mid y) = \frac{P\,(\,y\,|\theta)P(\,\theta)}{\int P(y|\theta)P(\theta)d\theta} \qquad (3.55)$$

Where $\int P(y|\theta)P(\theta)d\theta$ is known as the normalizing constant, which is equal to the marginal pdf of y with no information about θ, which can also be taken as a constant. Because of this the distribution can be taken as a proportion, which is denote mathematically as:

$$P(\theta \mid y) \propto P(y \mid \theta)\, P(\theta) \tag{3.56}$$

Where the constant of proportionality is the reciprocal of $\int P(y|\theta)P(\theta)\, d\theta$. From equation 3.56, the right-hand side is the product of the likelihood and the prior distribution.

Hence, in disease mapping where often the data are counts or can be transformed to count, the assumption, as explained by Lesaffre and Lawson (2012), is that the likelihood follows a Poisson distribution with a common relative risk parameter and a single gamma prior distribution.

The posterior distribution is usually presented as (a) summary measures for location and variability; (b) interval estimators for the parameters of interest, and (c) the posterior predictive distribution (PPP) used to predict future observations.

The most popularly used measures of location are the posterior mean, posterior median, and posterior mode. The posterior mean is defined by $\bar{\theta} = \int \theta p\ (\theta|y)\ d\theta$, which minimizes the squares loss, that is $\int \ \left(\theta - \hat{\theta}\right)^2\ p(\theta|y)\ d\theta$. The posterior median is the solution to the equation $0.5 = \int_{\ \bar{\theta}_M} P(\theta|y)\ d\theta$. The posterior mode is defined by $\hat{\theta}_M$ $\arg max_\theta\ p\ (\theta|y)$, and gives the value of θ for which $p\ (\theta|y)$ is maximal. A measure of variability which determines the shape of the distribution is the posterior variance $\bar{\sigma}^2$ (together with the posterior standard deviation $\bar{\sigma}$), which is defined as $\bar{\sigma}^2 = \int \left(\theta - \hat{\theta}\right)^2 p\ (\theta|y)$. A range of plausible parameter values of θ, which is termed credible interval with probability $1 - \alpha$, can also be obtained from the posterior distribution. Formally, an interval $[a, b]$ is a $100\ (1 - \alpha)\%$ credibility interval for θ if

$$P\ (a \le \theta \le b|y) = 1 - \alpha.$$

It is worth nothing that PPD is the distribution of unobserved observations conditional on the observed data. Suppose $p\ (y|\theta)$ is the distribution of y and assume an i.i.d. sample $y \equiv \{y_1, \dots\dots, y_n\}$ is available and suppose we wish to predict future observation $\tilde{y}$ or sets of observations $\widetilde{y},$ that is, we wish to obtain the distribution of $\widetilde{y},$ as the PPD and is given as :

$$p\ (\tilde{y}\ |\theta)\ =\ \int p\ \ (\tilde{y}\ |\theta)\ \ p\ (\theta|y)\ \ d\theta \tag{3.57}$$

Many examples of posterior distributions and PPD under such distributions such as binomial for binary and Poisson for count data were given by Lesaffre & Lawson (2012).

3.7.7. Bayesian Linear Regression model

The Bayesian linear regression model (BLRM), is a statistical approach to linear regression in which the statistical analysis undergoes the context of Bayesian analysis. A BLRM combines prior information on the regression parameters and residual variance with the derived normal regression likelihood.

The general linear regression model assumes a linear relationship between a response y y and d regressors $x_1, x_2, \dots\dots, x_d$, for a sample of n observations. Thus, mathematically given as

$$y = X^T\beta +\ \varepsilon_i \quad (i = 1, \dots, n) \tag{3.58}$$

With $\beta^T\ = (\beta_0, \beta_1, \dots\dots, \beta_n\ ),\ \left(x^T\ =\ (1, x_1, x_2, \dots\dots, x_n)\right),\ and\ \ \varepsilon_i \sim N(0, \sigma^2)$

In matrix notation, the classical normal linear regression model is written as:

$$y = X\beta +\ \varepsilon \tag{3.59}$$

where y is the $n\ x\ 1$ (n times 1 matrix) vector of responses, X is the $n\ x\ (d + 1)$ design matrix with rows x_i^T and ε the $n\ x\ 1$ vector of the random errors, assuming a normal distribution $N(0, \sigma^2, I)$ with the I identity $n\ x\ 1$ matrix.

Bayesian linear regression in a Bayesian framework is then expressed in words as the response data point y from a multivariate normal distribution that has a mean which is equal to the product of the β coefficients and the predictors X, and a variance of $\sigma^2. I$ is the identity matrix, which is necessary because the distribution is multivariate. Therefore, the BLRM is expressed mathematically as:

$$y \sim N(\mathrm{X}\beta, \sigma^2\, I) \tag{3.60}$$

3.7.7.1. The Likelihood

From the assumption in section 3.6.6 above the likelihood is:

$$Y = Z + \varepsilon \tag{3.61}$$

The classical estimate of β is the Maximum Likelihood Estimate (MLE), which is also equal to the least square estimate (LSE), given as $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Furthermore, the residual variability of the observed response to the fitted response is thus expressed as $SSE = (y - X\beta)^T \left(y - X\hat{\beta}\right)$, which is called the residual sum pf square and the mean sum *pf* square is thus given as $S^2 = {}^{SSE}/_{(n - d - 1)}$ ·

3.7.7.2. Prior

A BRLM combines prior distribution on the regression parameters and the residual variance with the above-derived normal regression likelihood. Prior information can be either non-informative or formative or conjugate, as discussed in section 3.6.4 and 3.6.5.

If we possess prior knowledge about the parameter, then we can choose a prior distribution that reflects this, otherwise a noninformative prior can be used. Let us consider a non-informative prior for the parameters (β, σ^2), the choice will then be:

$$p\,(\beta, \sigma^2) \propto \sigma^{-2} \tag{3.62}$$

A popular non-informative (NI) prior for the BLRM is obtained from Jefferey's multi parameter rule, which is the product of a flat prior for the regression parameters and the classical Jefferey prior for the scale parameter. Often a normal prior with large variance is recommended and chosen for the regression coefficients β, σ^2 and with a small ε.

However, a prior $p\ (\beta, \sigma^2)$ is conjugate to the likelihood function if it has the same functional form with respect to β and σ^2. Since the log-likelihood is quadratic in β, it can then be re-written in such that the likelihood becomes normal in $(\beta - \hat{\beta})$, stated as:

$$(y - X\beta)^T(y - X\beta) = \left(y - X\,\hat{\beta}\right)^T\left(y - X\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)^T(X^T X)\left(\beta - \hat{\beta}\right) \qquad (3.63)$$

Then, the likelihood can now be rewritten as

$$p\ (y \mid X, \beta, \sigma^2) \propto (\sigma^2)^{v/2} \exp\left(-\frac{vs^2}{2\sigma^2}\right)(\sigma^2)^{-n-v/_2}\, exp\left[-\frac{1}{2\sigma^2}\left(\beta - \hat{\beta}\right)^T (X^T X)\left(\beta - \hat{\beta}\right)\right], \qquad (3.64)$$

Where $vs^2 = \left(y - X\hat{\beta}\right)^T\left(y - X\hat{\beta}\right)$ and $v = n - k$, and k is the number of regression coefficient.

This then suggests a form for the prior given as:

$$p(\beta, \sigma^2) = p(\sigma^2)p(\beta|\sigma^2), \qquad (3.65)$$

Where $p(\sigma^2)$ is an inverse-gamma distribution given as:

$$p(\sigma^2) \propto\ (\sigma^2)p(\beta|\sigma^2) \qquad (3.66)$$

And the conditional prior density $p(\beta|\sigma^2)$ is a normal distribution, denoted as:

$$p(\beta|\sigma^2) \propto (\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta-\mu_0)^T\right) \vartheta_0 (\beta-\mu_0) \tag{3.67}$$

And the conditional prior distribution for equation (3.67) is $N\ (\mu_0, \sigma^2, \vartheta_0^{-1})$

3.7.7.3. Posterior

Having specified the prior, the posterior distribution can be stated as:

$$\begin{aligned} p(\beta,\sigma^2\,|y,X) &\propto p(y|X,\beta,\sigma^2)\, p(\beta|\sigma^2)\, p(\sigma^2) \\ \propto (\sigma^2)^{-\frac{n}{2}} &\exp\left(-\frac{1}{2\sigma^2}\ (y-X\beta)^T\ (y \right. \\ \left. - X\beta)\right) &(\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}\ (\beta-\mu_0)^T \vartheta_0\ (\beta \right. \\ \left. - \mu_0)(\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)\right) \end{aligned} \tag{3.68}$$

Because all models are conditional on the design matrix X, X will be omitted in the notation of the posterior. With some re-arrangement as stated by Lesaffre and Lawson (2012), the posterior can then be re-written so that the posterior mean μ_n of the parameter vector β can be expressed in terms of the least square estimator $\hat{\beta}$ and the prior mean μ_0, with the strength of the prior indicated by the prior precision ϑ_0.

The joint posterior distribution is $(d+2)$-dimensional: $(d+1)$ dimensions pertain to the regression parameters β and one dimension pertains to the residual variance σ^2. With the choice of a non-informative prior in section 3.7.4, the posterior distribution can be mathematically derived.

3.7.8. Bayesian Poisson Regression

In this section, we present a Bayesian regression model for Poisson regression. For a spatial-structured, over-dispersed, count or clustered data, there are several assumptions that can be implemented in modeling such data in Bayesian statistics (Lesaffre & Lawson, 2012; Ntzoufras (2011)):

A1: The n regions are unique, not related to each other. Under this assumption, information from region j on θ_i $(j \neq i)$ does not provide any in is equal to SMR with asymptotic variance $\theta_i/_{e_i}$. In this case, SMR shows a high variability for a sparsely populated region (with small e_i)and the Bayesian estimate of θ_i $d$epends only on y_i.

A2: The counts $y_1, \dots\dots . y_n$ are simple random sample of a population and hence each SMR_i estimate same risk θ. Since the reference region is internal, that means that $\theta_1 = \cdots = \theta_n \; \theta = 1.$ For an external reference region, that is when $\theta \neq 1,$ and the Poisson assumption, the MLE of θ is $\sum_i y_i \, / \sum_i e_i$.

A3: The regions share some common environmental conditions, especially when they are close together geographically. For example, climatological conditions, air pollution, and life style do not change drastically when crossing the border of a region. Which implies that the $\theta_i \, s$ are related. To express this, we assume that $\{\theta_1, \dots\dots ., \theta_n\}$ is a random sample from $p(\theta_. \, |.)$, which is called the prior of θ_i.

The choice between assumptions A1 and A3 depends on subjective arguments. Assumption A2 can be tested statically. Because of sample heterogeneity, A3 seems reasonable and was recommended. Therefore, assumption A3 is a compromise between A1 and A2, implying that $\theta_1, \dots .., \theta_n$ have a common distribution $p(\theta|\lambda)$, with λ a vector of hyperparameters. Under these assumptions, the set $\theta = \{\theta_1, \dots .., \theta_n\}$ is called exchangeable. The subject within the population are exchangeable but not between regions. Thus, exchangeability is applied on the two levels of the hierarchy. However, the implication of exchangeability is that θ_i are estimated using information from all the region. This phenomenon is called “borrowing strength “from other regions."

The standard Poisson regression model was adopted:

$$\lambda(x) = X^T\beta,\ \ \theta(x) = e^{\lambda(x)}, y \tag{3.69}$$

$$= Y_i \ \sim \ Poisson\ (\lambda_i, \theta_i)$$

$$\text{For } i = 1, 2, \dots \dots \dots, n$$

where the density function is given by

$$g(\theta|\alpha, \beta) = \frac{\beta(\beta\theta)^{\alpha-1}\ exp(-\beta\theta)}{\Gamma(\alpha)}$$

3.7.8.1. **Likelihood**

The marginal likelihood of the Poisson model is given by integrating the random effect θ_i. The conditional likelihood is given as $f(y|\ \lambda_1, \lambda_2, \alpha_1, \beta_2, \theta\ )$.

3.7.8.2. **The Prior**

With a choice of assumption A3 above, the prior choice that is computationally convenient is the conjugate for the Poisson distribution, with the prior mean and variance of the parameter θ_is is $E(\theta) = \frac{\alpha}{\beta}\ and\ Var(\theta) = \frac{\alpha}{\beta^2}$ respectively. Without specifying a prior distribution for the hyperparameters $\alpha\ \ and\ \beta$, the above model is a frequentist hierarchical model. Generally, the choice of the hyperprior $p(\alpha\ ,\beta)$ depends on our prior belief about the hyperparameters. The Poisson-gamma model is a two-level hierarchical model with the following levels:

- Level 1: $y_i\ |\ \theta_i \sim Poisson\ (\theta_i y_i)\ \ for\ \ i = 1, \dots., n$
- Prior: $(\alpha,\ \beta) \sim\ p(\alpha, \beta)$

The distribution of θ_i is often referred to as a prior in the Bayesian literature, since in the paradigm, there is no difference between parameters and latent random variables. Likewise, in a more general setting, we may have m_i counts y_i for each θ_i, in which case the first level of the above two level hierarchical model becomes:

- Level 1: $y_{ij} \mid \theta_i \sim Poisson\ (\theta_i e_i)\ for\ j = 1, \dots\dots, m\ ; i = 1, \dots\dots, n$

3.7.8.3. **The Posterior Distribution**

Based on the distributional assumptions made above, the posterior distribution can be derived. Let y reperesent the set of observed counts $\{y_1, \dots\dots, y_n\}$, then the joint posterior distribution is

$$p(\alpha, \beta, \mid y) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \alpha, \beta) \prod_{i=1}^{n} p(\theta_i \mid \alpha, \beta)\ p(\alpha, \beta) \tag{3.70}$$

By assuming that $y_i \sim Poisson\ (\theta_i e_i)$, implicit hierarchical independence is assumed for the counts, and thus,

$$p(\alpha, \beta, \theta \mid y) \propto \prod_{i=1}^{n} \frac{(\theta_i e_i)^{y_i}}{y_i\,!} \exp(-\theta_i e_i) \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)}\ \theta_i^{\alpha-1} e^{-\beta\theta_i}\, p(\alpha, \beta) \tag{3.71}$$

However, a Gibbs sampling approach requires the determination of the full conditionals. Choosing an independent exponential prior for the hyperparameters: $p(\alpha) = \lambda_\alpha \exp(-\lambda_\alpha \alpha)$ and $p(\beta) = \lambda_\beta \exp(-\lambda_\beta \beta$, for this choice, the full conditionals are:

$$p(\theta_i \mid \boldsymbol{\theta_i}, \alpha, \beta, y) \propto \theta_i^{y_i + \alpha - 1} \exp[-\,(e_i + \beta)\theta_i]\ \ (i = 1, \dots, n),$$

$$p(\alpha \mid \theta, \beta, y) \propto \frac{(\beta^n \prod \theta_i)^{\alpha-1}}{\Gamma\,(\alpha)^n} \exp(-\lambda_\alpha \alpha),$$

$$p(\beta \mid \theta, \alpha, y) \propto \beta^{n\alpha} \exp\left[-\left(\sum \theta_i + \lambda_\beta\right)\beta\right],$$

With θ_i is equal to θ without θ_i (Cameron and Trivedi, 2013).

3.7.9. Bayesian Negative-Binomial

The standard model for over-dispersion count data is the negative binomial (Collaboration, 2010, Cameron and Trivedi, 2013, Barron, 1992, Lachin, 2011).

In regression analysis of count data, the main objective is to explain the relationship between the data and its covariates. Based on the estimated regression coefficient, the future count can be replicated (or rather predicted) from the explained covariates (Christensen and Waagepetersen, 2002, Cameron and Trivedi, 2013).

3.7.9.1. **Likelihood**

A discrete random variable Y follows a negative binomial (NB) distribution $Y \sim NB\ (\mu, k)$ with the probability function given as:

$$f_{NB}\ (y\colon \mu, k) = \frac{\Gamma\ (y+k)}{y!\ \Gamma\ (k)}\ \mu^k\ (1-\mu)^y \tag{3.72}$$

Where $y = 0,1,2, \ldots, k$ and $k > 0$, a positive parameter. The mean and the variance of this distribution are equal to $E(Y) = \ \mu\ = \log(\mu) = X^T\beta$ and $Var\ (Y) = \ \mu + k\mu^2$.

3.7.9.2. **Prior**

The normal distribution is assumed to be the natural prior for the process model and the embedded linear regression coefficient or say the parameter *r* is assumed to follow a gamma distribution because of its positive nature and the dispersion:

$$k \sim gamma\ (0.001, 0.001)$$

$$beta \sim\ dnorm\ (0, 0.001)$$

From the prior, the posterior mean and variance can then be computed.

3.8. Bayesian Markov Chain Monte Carlo Methods (MCMC) Methods: Parameter Estimation

This section briefly describes the elementary theory of the Markov Chain Monte Carlo (MCMC) algorithm. The MCMC algorithm is widely used to perform posterior inference in the case where the product of the likelihood and the prior are analytically intractable (Congdon, 2014, Ntzoufras, 2011). MCMC is a technique used for drawing samples that will approximate a posterior distribution with respect to its complexity. However, this draw will not be independent, but from a Markov chain.

MCMC techniques have contributed greatly to the development and propagation of Bayesian theory. The aim of MCMC algorithm is to combine a Markov process with Monte Carlo simulation techniques, to generate random numbers of a stationary distribution similar to a posterior distribution. The summary statistics of the MCMC samples will approximate the properties of the posterior distributions. However, there are assumptions and requirements to be fulfilled to ensure that MCMC random numbers generated converge to the distribution. There are multiple methods of MCMC, which can be implemented in the popular Win BUGS software (Ntzoufras, 2011), as well as in R and SAS software. This method involved a set of steps that use iterative simulation of parameter values from the posterior distribution based on a Markov chain (MC) formulation.

A Markov chain is a sequence of random variables $(\theta_n)_{n=0}^{\infty}$ in which the future is independent of the past, given the present. This can be written as:

$$\pi\ (\theta_{n+1} \mid \theta_0, \theta_1, \dots \dots, \theta_n) = \ \pi\ (\theta_{n+1}\ | \theta_n) \tag{3.73}$$

For MCMC methods to function, we should use a Markov chain that has a stationary posterior distribution $\pi(\theta) = p(\theta\ |y)$. The main aim of this method is that after a period, they will consist of a sample from the posterior distribution.

MCMC is carried out starting from one chain and subsequently increasing the number of chains to ensure the convergence of the iterative process to a stable estimation of posteriors. MCMC conveniently generates many independent posterior estimates, handling non-normal and skewed posterior distributions, such as those that arise when testing indirect or interactive effects (Ntzoufras, 2011). Therefore, in relation to classical (traditional or frequentist) methods, MCMC also sounds like other forms of simulation such as bootstrapping which may be a helpful analogy (Lawson, 2013).

3.8.2. The MCMC Algorithm

According to (Ntzoufras, 2011) Markov chain is stochastic process $\theta^{(1)}, \theta^{(2)}, \dots \dots, \theta^{(T)}\}$, such that,

$$f\left(\theta^{(t+1)} \mid \theta^{(t)}, \dots . \theta^{(1)}\right) = f\left(\theta^{t+1} \mid \theta^{(t)}\right) \tag{3.84}$$

That is the distribution of θ at sequence t+1 given all the preceding θ values (for times t, t-1, …, 1) depend only on the value $\theta^{(t)}$ of the previous sequence *t*. Let us consider the case where, $f\left(\theta^{t+1} \mid \theta^{(t)}\right)$ is independent of time t. We also consider the situation where a homogenous Markov chain is irreducible, aperiodic, and positive recurrent. As $t \rightarrow \infty$, the distribution of $\theta^{(t)}$ converges to its equilibrium distribution, which is independent of the initial values of the chain $\theta^{(0)}$.

Therefore, a sample from $f(\theta \mid y)$ the posterior must be generated. It is called an ergodic MC, which then follows the construction of a Markov chain with two desired properties:

1. $f\left(\theta^{(t+1)} \mid \theta^{(t)}\right)$ should be easy to generate from, and
2. The equilibrium of distribution of the selected Markov chain must be the posterior distribution of interest$f\ (\theta \mid y)$. This condition occurs when the MC is irreducible, aperiodic and positive-recurrent (Nummelin 2004).

Therefore, we describe an MCMC algorithm for which this property holds. The standard approach to Bayesian inference using MCMC is as follows:

1. Select an initial value θ^0.
2. Generate *S* values until the equilibrium distribution is reached
3. Monitor the convergence of the algorithm using the convergence diagnostic. If convergence diagnostic fails, we further generate more samples.
4. Cut off the first *R* observations. This is referring to as Burn-in period. Here the first R iterations are eliminated from the samples to avoid the influence of the initial values. However, Ntzoufras (2011) stated that if the sample generated is large enough, the effect of the burn-in periodS on the calculation of the posterior will be minimal, therefore a large refresh value is

recommended for simple models while for a complicated model (while the algorithm will be slower) a low refresh value is recommended.

5. Consider as the sample for the posterior analysis
6. Plot the posterior distribution.
7. Finally, obtain summaries of the posterior distribution (that is mean, median, standard deviation, quantiles, and correlations)

From the above, steps 6-7 are referred to as the convergence diagnostic. A convergence diagnostic is a statistical test used in identifying cases where convergence is not achieved. The MCMC output provides us with a random sample.

From this sample, for any function $J(\theta)$ of parameter of interest θ, a sample of the desired parameter $J(\theta)$ by simply considering $J(\theta^{(1)}), J(\theta^{(2)}), \dots, \dots., J(\theta^{(S')})$. A posterior summary $J(\theta)$ can also be obtained from the sample distribution using the traditional sample estimates (Ntzoufras, 2011). Thus, the posterior mean $\widehat{E}\ (J(\theta)|\beta$ can be estimated by

$$\widehat{E}\ (J(\theta)\mid\beta = \frac{1}{S'}\sum_{s=1}^{S'} J(\theta^{s}) \tag{3.95}$$

And the posterior standard deviation $\widehat{SD}$ by

$$\widehat{SD}\ (J(\theta)|\beta\ = \frac{1}{S'}\sum_{s=1}^{S'}\left[J(\theta^{s}) - \widehat{E}\ (J(\theta)\mid\beta\right]^{2} \tag{3.106}$$

Other measures of interests such as the posterior median or quantiles (25% and 95% credible intervals) can be obtained in the similar form (Ntzoufras, 2011). Monitor correlations between the parameters can also be obtained, as well as produce plots of the marginal posterior distributions such as histograms, density plots, boxplots, error boxplots.

According to (Ntzoufras, 2011), the two most popular MCMC algorithms are the Metropolis-Hastings (Metropolis et., al., 1953; Hasting, 1970) and the Gibbs Sampler (German and German, 1984). These two algorithms have many variants with further extensions that have been developed which are more advanced and sometimes more specific to some problems. The choice between the two methods often depends on the problem at hand. Other MCMC algorithms that exist recently in the MCMC literature are the slice sampler (Hidgon, 1998; Damien et. al., 1999; Neal, 2003), the reversible MCM (RJMCMC) algorithm (Green, 1995), and perfect Sampling (Propp and Wilson, 1996; Merller 1999; (Ntzoufras, 2011).

This study uses the Metropolis-Hastings algorithm because the full conditional is easy to sample, and it is most popular, with accurate implementations in SAS (Lesaffre & Lawson 2006; Overt & Casella 2004). The next section gives a brief general discussion of both methods.

3.8.3. The Metropolis-Hastings Algorithm (M-H)

The Metropolis Hastings (M-H) is a MCMC technique that was introduced by Metropolis et al. (1953) and generalized by Hastings in 1970.

Suppose we have π as a vector-valued parameter $P(\theta|y)$ the target posterior distribution from which we wish to generate the sample of size T. in Bayesian paradigm, the M-H algorithm follows a step by step iterative steps, as follows:

1. Begin with initial values $\theta^{(0)}$.
2. For t = 1,..., T, repeat the following steps
    - Set $\theta = \theta^{(t-1)}$,
    - generate entirely new values denoted as π' from proposal distribution $q(\theta' \,|\theta)$
    - calculate the ratio

$$\beta = \min\left(1, \frac{P(\theta'|y)q(\theta|\pi\theta')}{P(\theta|y)q(\theta'|\theta)}\right) \tag{3.117}$$

- update $\theta^{(t)} = \theta'$ with probability β or $\theta^{(t)} = \theta^{(t-1)}$, with probability $1 - \beta$

An important characteristic of this algorithm to note is that it convergence depends on the proposal distribution. Thus, practically, it is essential to carefully choose the proposal distribution, as a poor choice can considerably delay convergence to the targeted distribution.

3.8.4. The Gibbs Sampler (GS)

The Gibbs Sampler (GS) was introduced by German and German (1984). It is a special case of the single-component Metropolis-Hastings algorithm which is usually cited as a separate simulation technique due to its convenience and popularity. GS became popular in the statistical world when Gelfand and Smith (1990), showed its ability to tackle complex estimation problems in a Bayesian manner.

One advantage of GS is that at each step, random variables must be generated from unidimensional distributions from which a wide range of computational tools emerge (Gilk,1996). However, it can be ineffective if the parameter space is complicated or highly correlated (Ntzoufras, 2011). GS is an algorithm for simulating samples from the posterior distribution. The algorithm can be summarized below:

1. set initial values θ^0.
2. For $i = 1, \ldots\ldots G$. repeat the steps below:
   - (i) set $\theta = \theta^{i-1}$
   - (ii) for $j = 1, \ldots\ldots d$, update θ_j from $\theta_j \sim P\left(\theta_j \,|\theta_{\ /j}\,, y\right)$, where

     $$\theta_{\ /j} = \theta_1, \ldots . \theta_{j-1}, \theta_{j+1}, \ldots \ldots \ldots \ldots \theta_j$$

   - (iii) set $\theta^i = \theta$ and save it as generated set of values at $i + 1$ iteration of the algorithm.

Hence, given a particular state of the chain $\theta^{(g)}$, we generate new values by

$$\theta_1^i \ from \qquad P\,(\theta_1\,|\theta_2^{i-1}, \ldots, \theta_p^{i-1}\ \ , y)$$

$$\theta_2^i \ from \qquad P\,(\theta_2\,|\theta_1^{i-1}, \theta_3^{i-1} \ldots, \theta_p^{i-1}\ \ , y)$$

$$\theta_3^i \ from \qquad P\,(\theta_3\,|\theta_1^{i-1}, \theta_2^{i}, \theta_4^{i-1} \ldots, \theta_p^{i-1}\ \ , y)$$

$$. \quad . \qquad\qquad .$$

$$. \quad . \qquad\qquad .$$

$$\theta_j^i \ from \qquad P\left(\theta_j \left|\theta_1^g, \theta_2^g \ldots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \ldots . \theta_j^{g-1} \quad , y\right)\right.$$

$$. \quad . \qquad .$$

$$. \quad . \qquad .$$

$$\theta_p^i \ from \qquad P\left(\theta_p \left|\theta_1^g, \theta_2^g \ldots, \theta_{p-1}^i, \ldots . \theta_{p-1}^g, y\right.\right.$$

As stated by Ntzoufras (2011), generating values from

$$P\left(\theta_j \left|\theta_{\ /j}, y\right) = P\left(\theta_j \left|\theta_1^g, \theta_2^g \ldots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \ldots . \theta_j^{g-1} \quad , y\right)\right.\right.$$

is relatively easy since it is a univariate distribution and can written as $P\left(\theta_j \left|\theta_{\ /j}, y\right) \propto P(\theta|y)\right.$, where all the variables except θ_j are kept constant at their given values.

3.8.5. Gibbs Sampler versus Metropolis-Hastings

There are advantages and disadvantages of using Gibbs sampler and Metropolis Hastings methods. (Lawson, 2013, Lesaffre and Lawson, 2012) stated that the Gibbs Sampler gives a single new value for each estimate at each interaction. On the contrary, each step of M-H does not require the evaluation of a conditional distribution, though the acceptance of a new value is not guaranteed.

GS can provide faster convergence chains when computations at the conditional distribution at each iteration are not time consuming. On the other hand, Metropolis-Hastings steps are usually faster at each iteration, but do not guarantee exploration acceptance of the generated value. Using a more global proposal in Metropolis Hastings algorithm may produce a higher efficiency, but the drawback stands with the choice of the proposal distribution.

It is worthy of note that GS is a special case of M-H algorithm which uses Open BUGS and JAGS software for Bayesian inference analysis. On the other hand, M-H is easy to understand and can use other software like SAS, Stata and SPSS.

Unlike GS, different proposals can be used for Metropolis-Hastings. As a result, most research is being conducted on this basis for efficient sampling of the posterior distribution.

3.8.6. Monte Carlo Error

In MCMC analysis, an important measure that must be reported and monitored from the output is the Monte Carlo error (denoted as MC error). The MC error measures the variability of each estimate due to the simulation. The MC error must be low to calculate the parameter of interest with increased precision.

The MC error is proportional to the inverse of the generated sample size that can be controlled by the user. Therefore, for enough iterations, *T*, the quantity of interest can be estimated with increased precision. Two common ways of estimating MC errors are the batch mean and window estimator methods. The first one is simple and easy to implement, and thus widely used, while the latter is more precise. For the batch mean method, the resulting output sample is partitioned into *K* batches, which must be sufficiently large to enable the estimation of the variance consistently and eliminate autocorrelation (Ntzoufras, 2011).

The Monte Carlo error of the posterior mean of $G(\theta)$ is calculated thus:

First, calculate each batch mean $\overline{G\theta_b}$ by

$$\overline{G\theta_b} = \frac{1}{v} \sum_{t=(b-1)v+1}^{bv} G(\theta^t) \tag{3.78}$$

For each batch *b=1, … , K* and the overall sample mean by

$$\overline{G(\theta)} = \frac{1}{T' \, \Sigma_{t=1}^{T'}} G(\theta^t) = \frac{1}{K} \sum\nolimits_{b=1}^{K} \overline{G\theta_b} \tag{3.79}$$

If we keep $\theta^{(1)}, \dots, \theta^{(T')}$ observations, then an estimate of the MC error is simply given by the standard deviation of the batch means estimates $\overline{G\theta_b}$

$$\text{MCE } [G(\theta)] = \widehat{SE}$$

$$MCE\ [G\ (\theta)] = \widehat{SE}\ \left[\overline{G(\theta)}\right] = \sqrt{\frac{1}{K}}\ \overline{\widehat{SD}}\ [\overline{G\theta_b}]$$

$$= \sqrt{\tfrac{1}{K\ (K-1)}\ \textstyle\sum_{b=1}^{K} \left(\overline{G(\theta)_b}\right)\ - G(\theta)^2}$$

Since the procedure for calculating the MC error for any other posterior quantity of interest is equivalent to $\widehat{U} = U(\theta^1, \ldots\ldots, \theta^s)$, to estimate the corresponding Monte Carlo error, we calculate $\widehat{U}_b = U\left(\theta^{((b-1)v+1)}, \ldots\ldots, \theta^{bv}\right)$ from each batch $b = 1, \ldots\ldots, k$, then get the MC error by

$$MCE\ (\widehat{U}) = \sqrt{\frac{1}{K(K-1)} \sum_{b=1}^{k} (\widehat{U}_b - \widehat{U})^2} \qquad (3.80)$$

The batch mean estimators of the MCMC errors are discussed in more detail by Hasting (1970), Geyer (1992), Roberts (1996), Carlin and Lois (2000), and Givens and Hoesting (2008).

The second method, known as window estimator, is based on the expression of the variance in autocorrelated samples given by Roberts (1996).

3.8.7. Assessing MCMC Convergence Diagnostic

The convergence of an algorithm is a term used to determine whether an algorithm has reached its equilibrium or target distribution. Several methods can be used to monitor convergence. The simplest and reliable way is to monitor the Monte Carlo (MC) error. A small value of the MC error (less than 0.5) is an indication that the quantity of interest has been calculated with precision. This was discussed explicitly in section 3.7.6. The second way is to monitor the trace plots (that is the plot of iteration versus the generated values). We assume convergence if all the values are within a strong periodicity, and most importantly, tendency zone.

A major decision to make is the number of iterations to be used to represent the posterior density and to ensure that Markov chain converged. However, the convergence of a model does not necessarily indicate a good model. It is just the starting point of model assessment.

Suppose we have *I* iterations and after discarding (burning) some, the remain is I'. Using the batch mean method, the MC error is calculated by first partitioning the resulting output into *K* batches, usually *K* = 30 or *K* = 50. To enable consistency in estimation and eliminate autocorrelation, both the number of batches and the sample size $s = {}^{I'}\!/_{K}$. The MC error of the posterior mean $\overline{T\ (x)}$ is calculated firstly by calculating batch means denoted by $\overline{T(x)_b}$.

where

$$\overline{T(x)_b} \ = \frac{1}{s} \sum_{t=(b-2)s+1}^{bs} T\left(X^{(t)}\right) \tag{3.81}$$

For each batch b = 1, 2…., K and the overall sample mean is given as

$$\mathrm{n}\ \overline{T\ (x)} \ = \frac{1}{K} \sum_{b=1}^{K} \overline{T(x)_b} \tag{3.122}$$

Assume we keep X^1 , … . . $X^{T'}$ observations. Then the MC error is given as:

$$MCE[T(x)] = SE\left[\,\overline{T\ (x)}\right] \ = \sqrt{\frac{1}{K\ (K-1)} \sum_{b=1}^{K} \left(T(X) - \overline{T(x)_b}\ \right)^2} \tag{3.133}$$

However, the window estimator method is based on the expression of variance in an autocorrelated samples as:

$$MCE[T(X)] = \frac{SD\ [\ (T\ (x)\ \ ]}{\sqrt{I'}} \ = \sqrt{1 + 2\ \sum_{k=1}^{\infty} \widehat{\rho k}\ \ [T(X)]} \tag{3.84}$$

Where (ρ k) ̂[T(X)] is the estimated autocorrelation of lag *k*. When *k* is large, the autocorrelation estimation cannot be reliable due to the small number of observations that remains. For a sufficiently large k, the autocorrelation should be close to zero. In this method, a window ω was identified after which autocorrelation are considerably low and discard ρ k with k > ω from the preceding MC error estimate. Thus, equation 3.82 is modified as

$$MCE[T(X)] = \frac{SD\ [\ (T\ (x)\ \ ]}{\sqrt{I}'} \ = \sqrt{1 + 2\ \sum_{k=1}^{\omega} \widehat{\rho k}\ \ [T(X)]} \tag{3.85}$$

The output of the algorithm produces a variety of results which are used for discussion and further conclusion. The summary of the outputs are discussed below.

- **Autocorrelation Function (ACF) Plots:** Autocorrelation is a situation where the estimated parameters in the chain are correlated. Congdon (2010) noted that the absence of vanishing autocorrelation with high lags is an indication of less information about the posterior, because for each iteration a high sample size is required to cover the parameter space. However, autocorrelation can be reduced by "thinning". Thinning involves storing of samples from *jth* iteration, where j>1 is the value of the field thinned (Congdon, 2010). When a long run is being carried out, thinning reduces MCMC errors. Likewise, (Lawson, 2013) mentioned "over-relaxation" as another way of reducing correlation. This generates multiple samples at each iteration and allows selection of one that is negatively correlated to the current value.
- **Gelman and Rubin Multiple Chain Convergence:** This is based on using two or more parallel chains with different starting values (Lawson et al., 2003, Ntzoufras, 2011). However, a multiple-chain convergence diagnostic is an evidence of the robustness across different subspaces ((Lawson et al., 2003). Thus, the convergence of a Markov Chain can be improved by standardizing covariates and the unstructured random effect (Condgon, 2010)
- **Kernel Density Plots:** For marginal posterior distributions that are approximately normal, a satisfactory density plot for a convergence chain should look more bell-shaped (Lawson et al., 2003, Ntzoufras, 2011)
- **Node Statistics:** This tool gives the posterior summary, which gives the estimates of posterior mean, standard deviation, MC errors, median, the quantiles, total number of iteration

generated sample size and the number of iterations that the sample started (the burn-in period) with. Here, focus is on the MC error, which measures the variation of the mean of the parameter of interest due to the simulation. The MC error should not be greater than 0.05 and increasing the number of iterations will decrease the MC error (Ntzoufras, 2011).

- **Running quantiles**: Monitoring the evolution of selected quantiles is essential. This plot is an indication of stability in the stated quantiles, which also implies convergence
- **CODA:** Here two windows are created for each chain. The first is the output file with all the generated values, sequentially stored for each monitored parameter with their corresponding iteration number. This file must be stored as text file for further transformation analysis in CODA library in R. The second windows are the index file with details of each nodes stored.

  The Markov Chain Monte Carlo diagnosis of the model output as explained above will be discussed below:

3.9. Comparison of Bayesian and Frequentist Approaches

Both Bayesian and frequentist statistical inference approaches allow one to evaluate evidence about competing hypothesis. The goal of this comparison is to explain the difference between the *p*-value in frequentist and a posterior distribution probability in Bayesian.

The Bayesian school of thought models' uncertainty using a probability distribution, as against hypothesis. Therefore, the ability to make inference depends on one's degree of confidence in the chosen prior, and the effectiveness of the findings to alternate prior distributions is highly relevant and important. However, the frequentist school of thought only uses conditional distributions of data given specific hypotheses. The presumption is that some hypothesis (parameters specifying the conditional distribution of the data) is true and that the observed data is sampled from that distribution. Particularly, the frequentist approach does not depend on a subjective prior that may vary from one researcher to another. These two schools of thought can be further contrasted as follows:

Bayesian inference

- Uses probabilities for both data and hypothesis.
- Depends on the prior and likelihood of observed data.

- Requires one to know or contrast a “subjective prior”.
- Has dominated statistical practice before the 20th century.
- Computation may be intensive due to integration over many parameters.

Frequentist inference

- Does not use or give the probability of a hypothesis (No prior or posterior).
- Depend on the likelihood P(D/H) for both observed and unobserved data.
- Dominated statistical practice during the 20th century.
- Less computationally intensive.

Frequentist measures like the *p*-value and confidence interval continue to dominate research, especially in the life sciences. However, in the current era of powerful computers and big data, Bayesian methods have undergone an enormous renaissance in research. The main critique of Bayesian methods is hinged on the subjective prior which may often lead to different posteriors and conclusions, unlike a straight objective frequentist approach.

The goal of a statistical inference is to help make a decision. The decision rule, however, combines both the expected and the observed data into a statistical framework for making decisions. In a decision-making rule, the frequentist will consider the expected variable given a hypothesis, whereas the Bayesian combines that expectation with a posterior distribution. In the long run, a significant theoretical result is that for any decision rule, there is a Bayesian decision rule which is more precise and better.

3.10. Conclusion

The Bayesian method is becoming very popular in statistical data analysis. Bayesian statistics differ from classical methods in terms of the use of prior information and likelihood functions, which give more power to the analysis.

Chapter 4

Results

3.0. Introduction

This chapter provides results of the application of statistical techniques to determine the spatial clustering and distribution of diabetes and the impact of its relative risk factors in Africa. Cluster analysis, Moran I indices and LISA approaches was used to determine which countries have similar pattern of diabetes and their spatial configuration and detect countries with a high risk of diabetes in Africa. Bayesian statistical analysis techniques were used to determine the risk factors that have more impact on diabetes prevalence in Africa.

4.1.1. Data Exploratory Analysis Results

The African continent with its 53 countries is the study area. The response variable considered in this study is the diabetes prevalence rate of adults between 20-69 years old. The predictors are prevalence of obesity, GDP, prevalence of alcohol consumption, percentage of physical activity, percentage of people aged 15-69, percentage of population growth, age dependency ratio, physician density per person, health expenditure, GNI, MYS, and HDI.

4.1.2. Summary Statistics of Diabetes and predictors

Table 4.1. Shows the summary statistics of variables in the study. The mean and median estimates of diabetes prevalence, obesity, GDP, age dependency ratio, urbanization, health expenditure shows that most of their values are concentrated on the high scale. In the same line of analogy, alcohol, and MYS are on the low scale. These variables are almost symmetrical and normally distributed. All the variables are positively skewed except HDI.

Table 4.1: Summary statistics of variables in the data set

Variable	Mean	Median	Min	Max	95% CI	Skewness
Diabetes	5.36	4.050	0.80	22.3	(2.3, 7.5)	2.159
Age dependency	6.23	5.670	4.34	13.45	(5.1, 6.7)	2.14
Alcohol	3.13	1.830	0.00	10.72	(0.5,5.3)	0.92
GDP	2767.810	1091.113	255.04	18918.28	(582.6,3393.6)	2.609
GNI	4838.38	772.207	0.00	23300.31	(1353.8,6032.3)	1.783
HDI	0.50	0.020	0.00	0.78	(0.43,0.59)	-1.054
Health expenditure	5.83	5.587	0.00	11.09	(4.04,7.25)	0.348
MYS	4.79	0.306	0.00	9.94	(3.12,6.23)	0.111
Obesity	11.50	9.250	2.60	33.10	(6.53,15.03)	1.248
Physical inactivity	17.66	19.050	0.00	46.90	(7.13, 25.25)	0.179
Physician density	0.31	0.104	0.00	2.83	(0.05,0.30)	3.129
Population age (20-75)	56.60	55.036	46.95	71.11	(52.55,59.24)	0.906

From Table 4.1. Mean prevalence of diabetes is 5.36 with the median of 4.050, 95% Confidence Interval (CI) is (2.3, 7.5) and positive skewness of 2.159. All the risk factors were positively skewed with HDI negatively skewed.

4.1.3. Spatial Pattern of Diabetes and Predictors

The geographic description of each variable in the study is made and visually presented on the African map using R software and library "maptools" and "spdep".

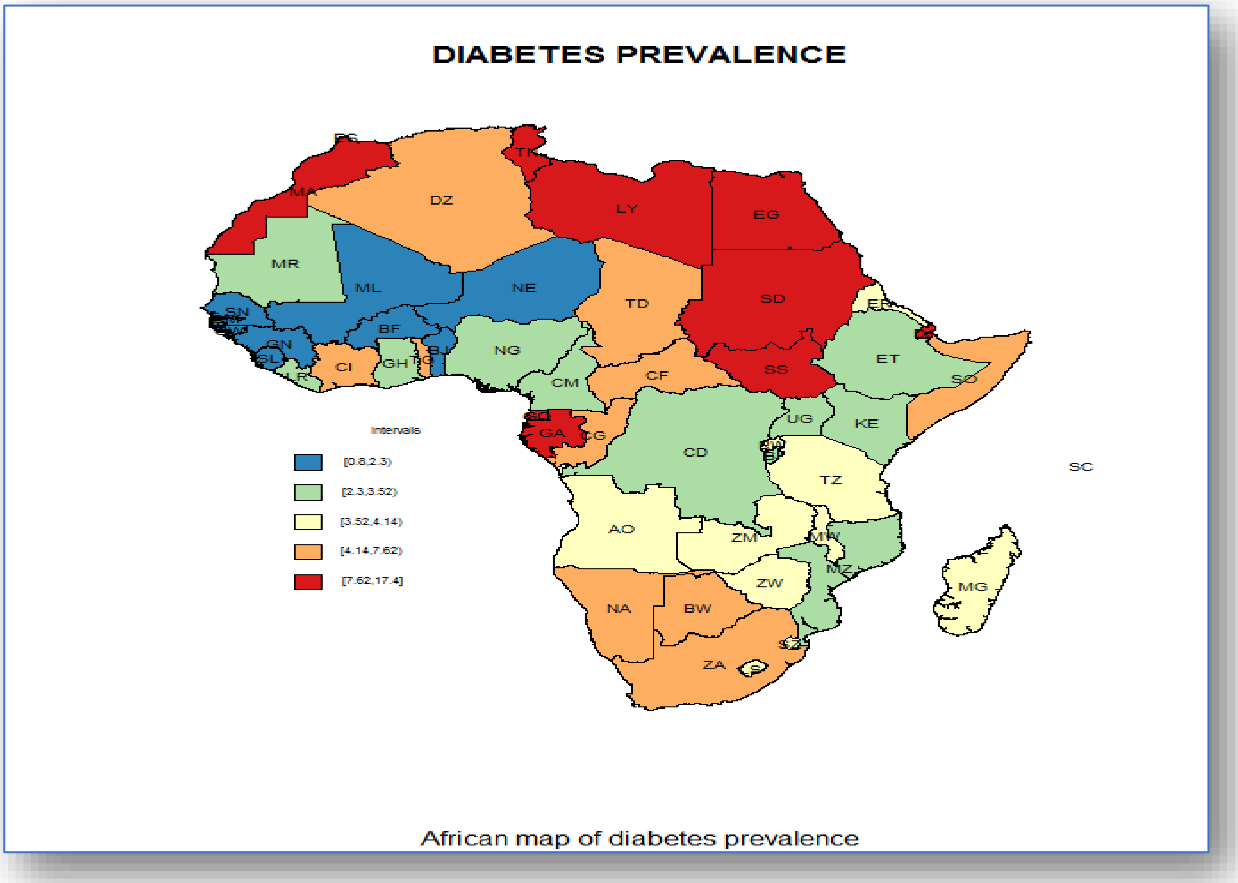


Figure 4.1. Spatial Distribution of Diabetes Prevalence in Africa

From Figure 4.1 above, it can be visualized that the prevalence of diabetes is not uniformly distributed in the continent. Countries with the higher diabetes prevalence are Mauritius (22.3%), Seychelles (17.4), Egypt (16.7), Libya (10.4%), Comoros (9.9%), Tunisia (9.6%), Sudan (8.9%), Djibouti (8.4%), South Sudan (8.1%), Morocco (7.1%), Gabon (7.8%), Equatorial Guinea (7.7, South Africa (7.6) and Algeria (7.5%). The countries with lower prevalence (less than 2.3%) in the continent are Benin (0.8%), Gambia (2%), Burkina Faso (2.2%), Niger (2.2%), Mali (2.2%), Senegal (2.2%), Sierra Leone (2.2%) , Guinea (2.2%,Guinea Bissau (2.2), Cabo Verde (2.3%), Ghana (2.3%), Liberia (2.3%), Mauritania (2.3%), Nigeria (2.3%), Sao Tome Principe (2.3%) and Kenya (2.4%). Overall, the highest prevalence is seen within the North, Central and South regions, with the West Africa region having the lowest prevalence of diabetes. The map clearly expresses some regional patterns which will be investigated in the following sections.

Because of the regional pattern exhibited in the Figure 4.1, a summary statistic of the continent region is necessary. This is expressed through a bar plot. A bar plot is a method of summarising a dataset. A bar plot of the median diabetes prevalence by country region in Figure 4.2 shows the that North region has the highest median prevalence of 8.90%, followed by Central Africa region with 6.45%, South region has 4.10%, East region has 3.85% and the lowest is in West Africa 2.20%.

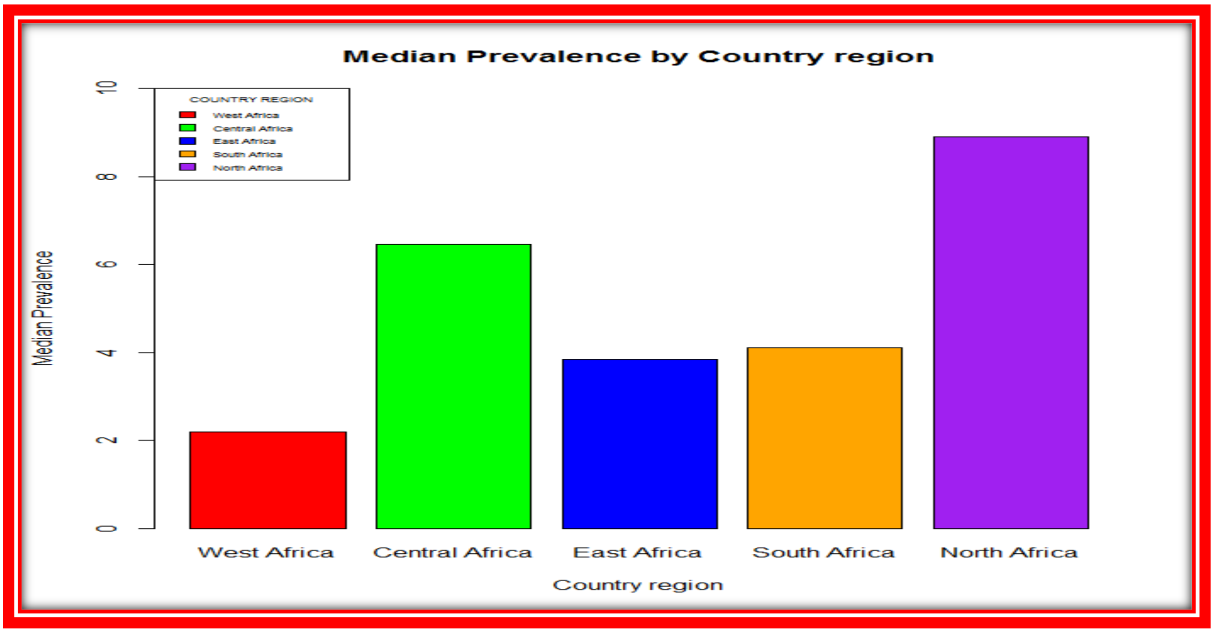


Figure 4.2. Bar plot of median prevalence

4.1.4. Spatial distribution of socio economic variables

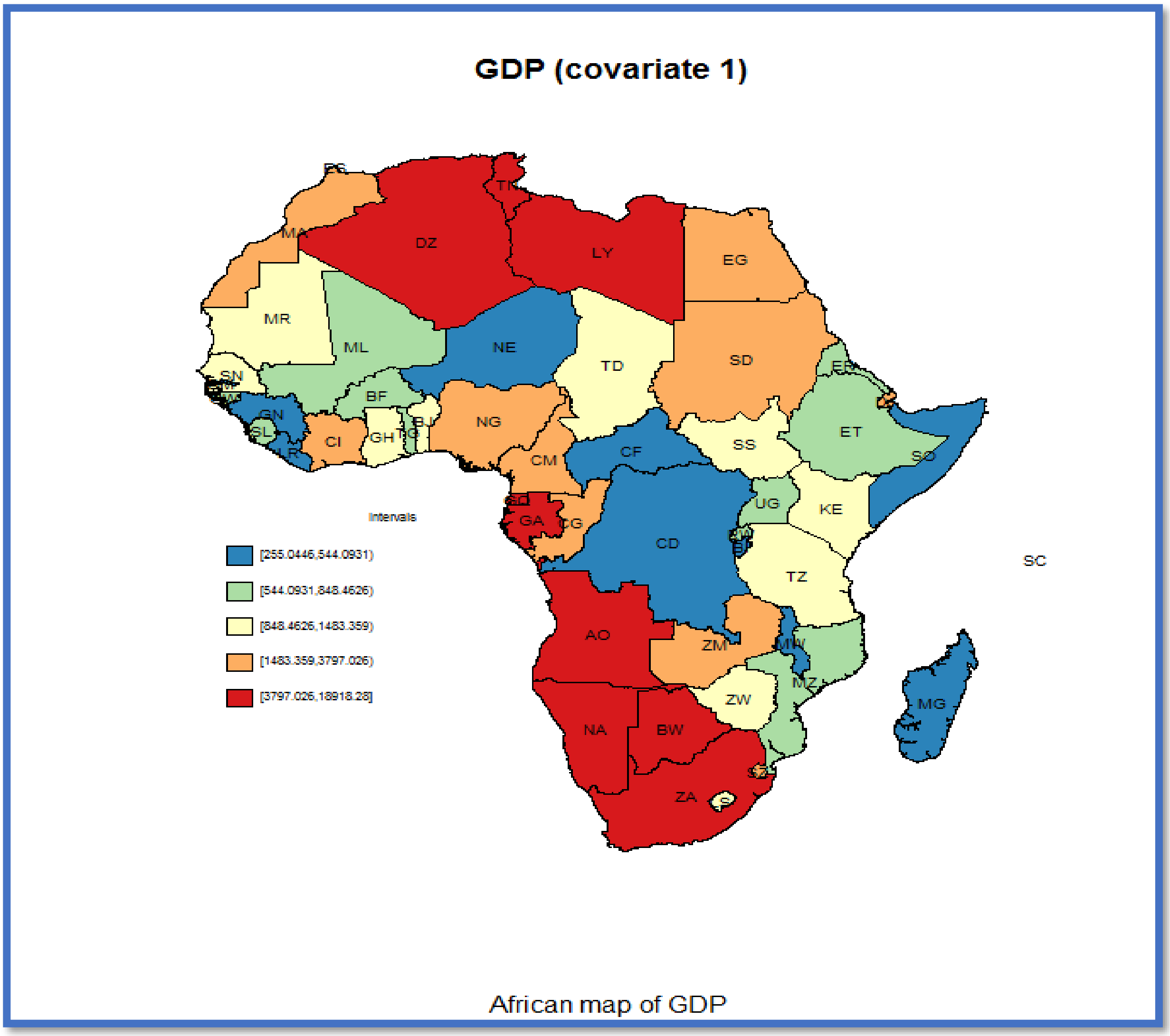


Figure 4.3. Spatial Distribution of African countries GDP

Figure 4.3 shows the GDP prevalence, with highest GDP prevalence in the North and Southern Africa region shown in red. The top ten countries with the highest GDP are Equatorial Guinea (18918.28), Seychelles (15564.64), Gabon (10772.06), Mauritius (10016.65), Botswana (7123.339), Libya (6573.387), South Africa (6483.855), Angola (5900.53), Algeria (5484.07), and Namibia (5408.234). The countries with lowest GDP are in blue. These are Guinea (539.6168), Liberia (457.86), Madagascar (449.40), Gambia (441.29), Niger (427.37), Niger (427.37), Cote d'Ivoire (422.37), Central Africa (358.54), Burundi (286.00) and Malawi (255.04).

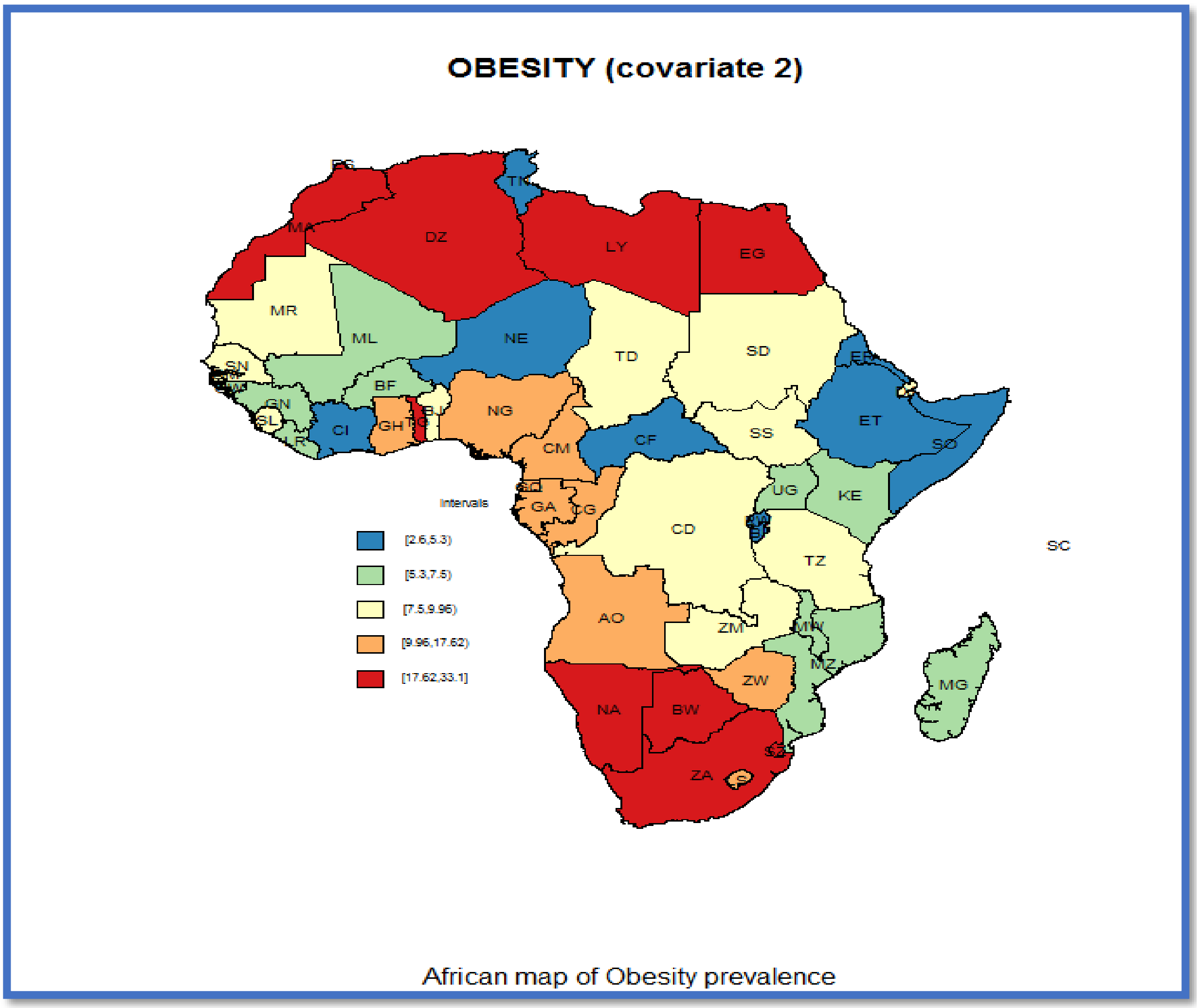


Figure 4.4. Spatial Distribution of African countries' Obesity Prevalence

From Figure 4.4 which shows the Obesity prevalence (covariate 2), the highest prevalence of obesity is seen within the North and Southern Africa region. The top eight highest countries are Libya (33.1), Egypt (28.9), Togo (27.1), South Africa (26.8), Seychelles (26.3), Algeria (24.8), Botswana (22.4) and Morocco (22.3). The countries with the lowest obesity prevalence are Tunisia (4.9), Somalia (4.6), Congo Republic (4.4), Niger (4.3), Eritrea (4.1), Rwanda (4.0), Ethiopia (4.0) and Burundi (2.6).

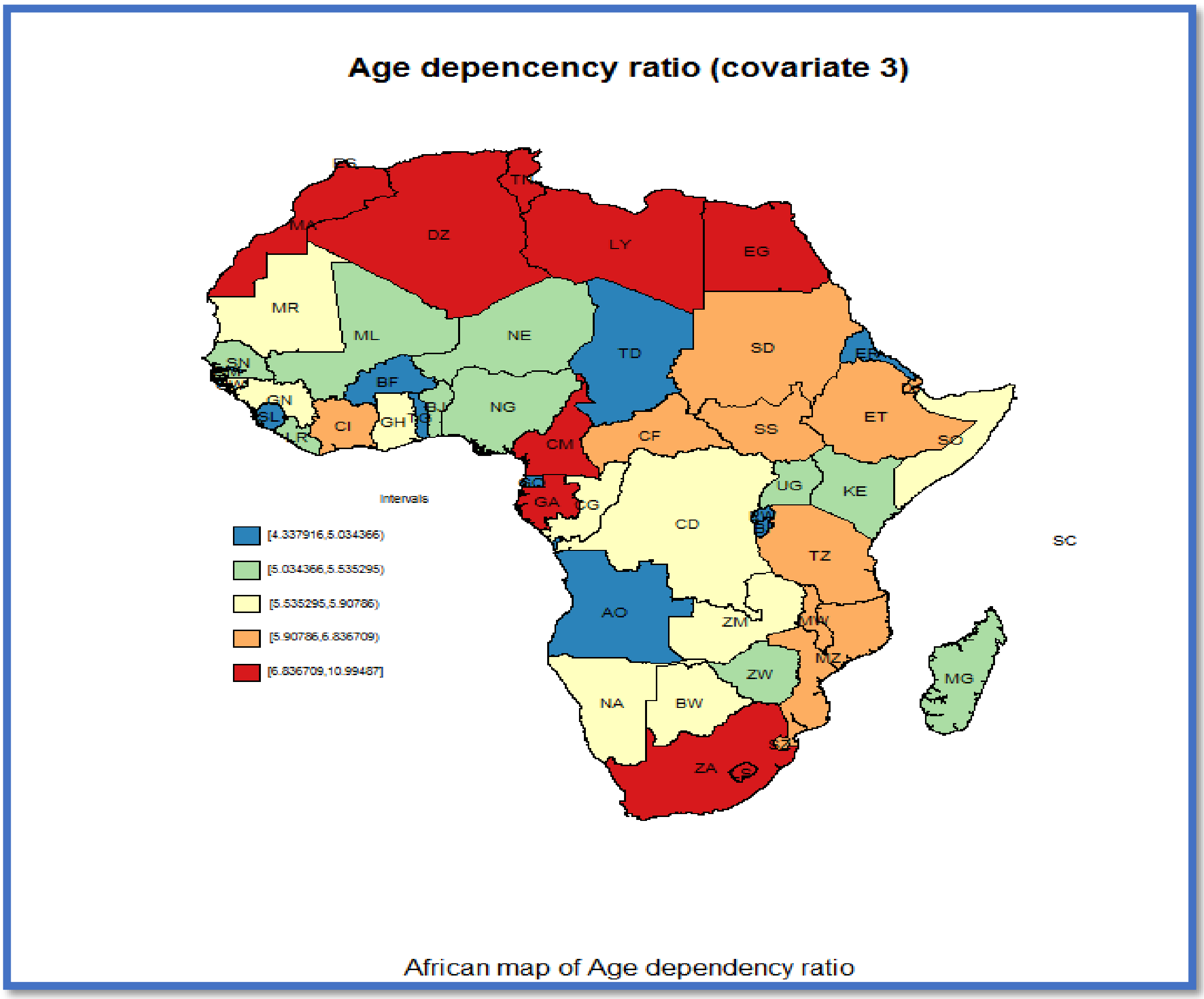


Figure 4.5. Spatial distribution of African countries' age dependency ratio

Figure 4.5 shows age dependency ratio, showing North Africa region having the highest countries with highest prevalence rate. The eight countries with the highest age dependency ratio are Mauritius (13.44), Tunisia (10.99), Seychelles (9.87), Morocco (9.26), Algeria (9.06), Gabon (8.83), Egypt (8.47) and South Africa (7.66). Those with the lowest age dependency ratio are Eritrea (4.33), Gambia (4.49) Burkina Faso (4.60), Angola (4.62), Burundi (6.69), Sierra Leone (4.86), Comoros (4.91), Chad (4.92), Rwanda (4.97), and Equatorial Guinea (4.99).

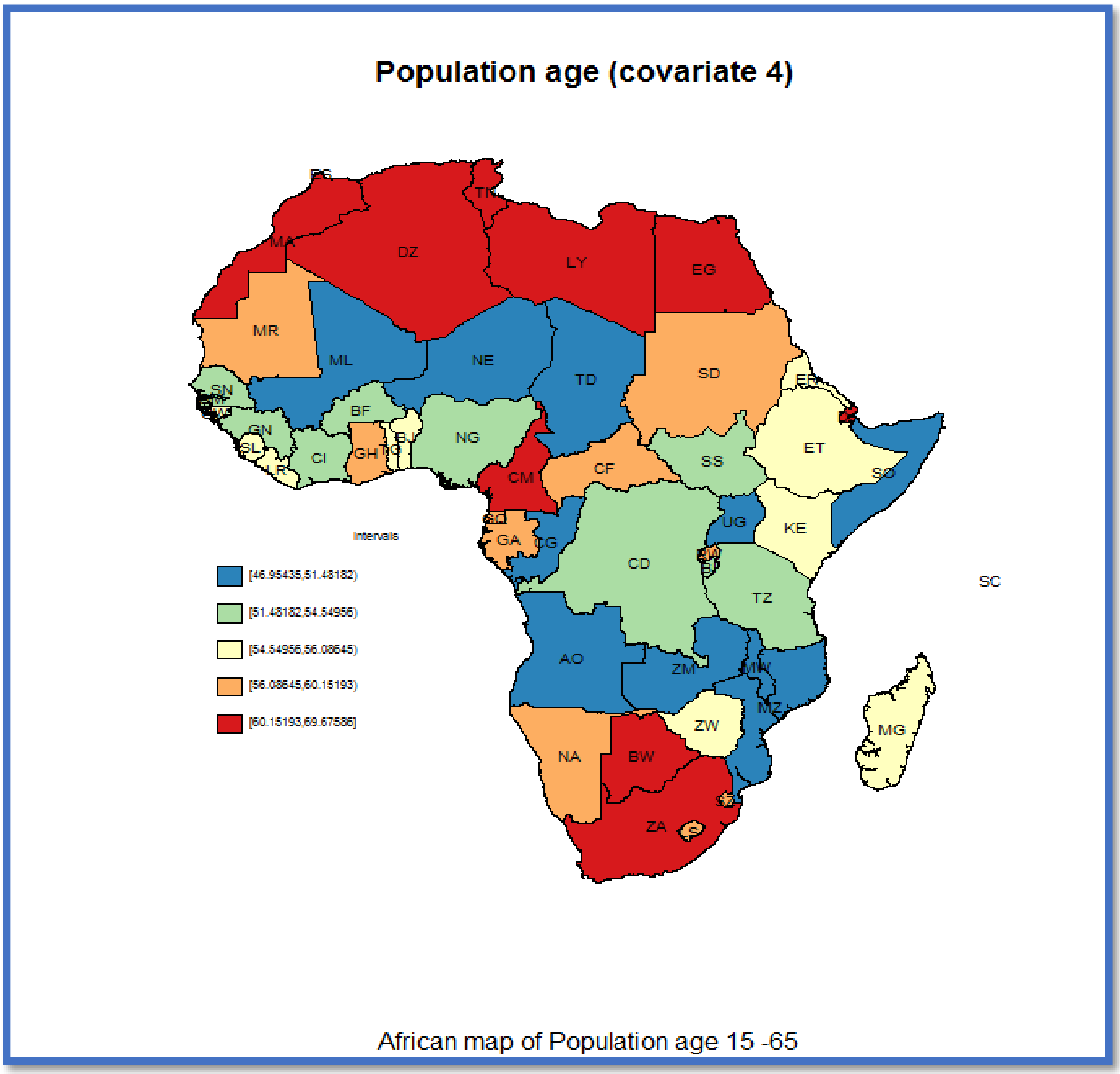


Figure 4.6. Spatial distribution of African countries' population age (15-65)

Figure 4.6 shows the distribution of Population age 15-65 in the continent. Countries with high population age are: Mauritius (71%), Seychelles (70%), Tunisia (70%), Morocco (67%), Cabo Verde (66%), South Africa (66%), Libya (66%), Algeria (66%), Botswana (64%), Djibouti (63%) and Egypt (61%). The countries with the lowest population age are Niger (47%), Uganda (49%), Chad (50%), Mali (50%), Angola (50%) and Somalia (50%).

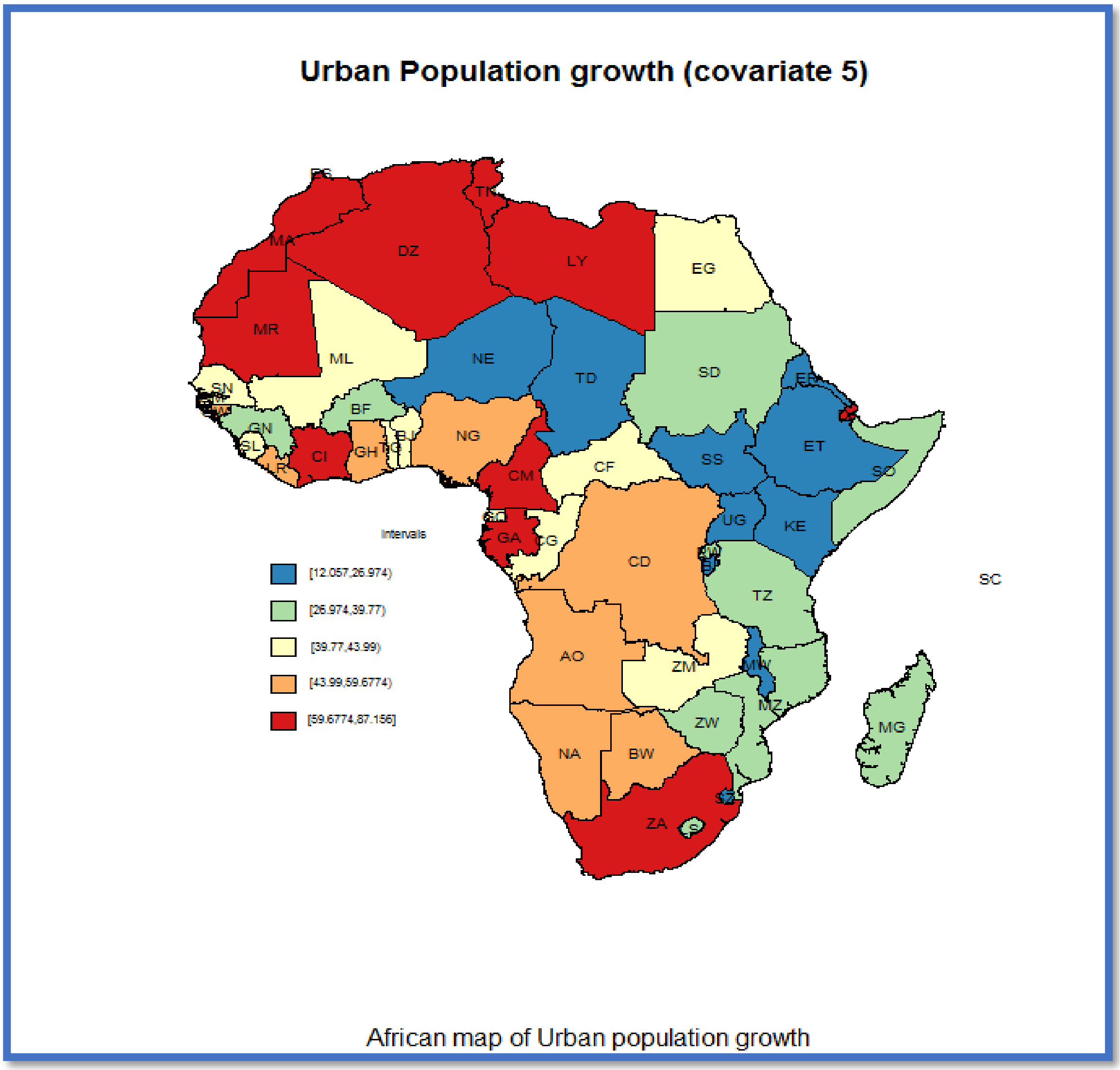


Figure 4.7. Spatial distribution of African countries' urban population growth

Figure 4.7 shows the distribution of urban population growth on the continent. Countries with the highest urban population growth are: Gabon (87.2 %), Libya (78.55%), Djibouti (77.3%), Algeria (70.7%), Tunisia (66.8%), Cabo Verde (65.5%), Congo Republic (65.3%), Sao Tome Principe (65.1%), South Africa (64.8%), Morocco (60.2%). Countries with lowest urban population growth are Burundi (12.1%), Uganda (16.1%), Malawi (16.3%), Niger 18.4%), South Sudan (18.8%), and Ethiopia (19.4%). Generally, highest distribution of urban population growth is seen more at the North Africa and central Africa region, with East Africa region having the lower distribution.

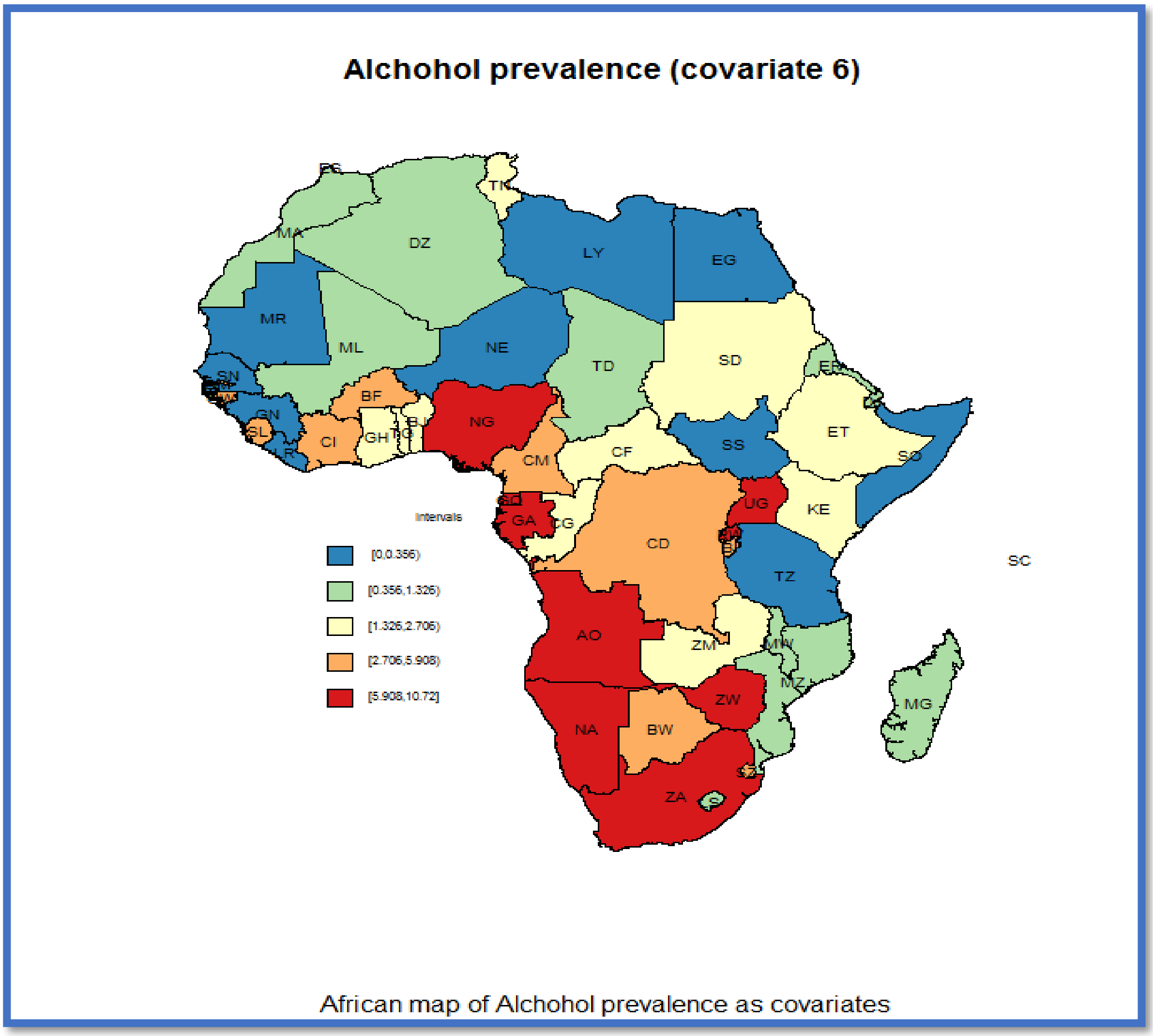


Figure 4.8. Spatial distribution of African countries' alcohol consumption prevalence

Figure 4.8 shows the distribution of Alcohol prevalence. The nine countries with highest alcohol consumption rate are Equatorial Guinea (10.7 %), Uganda (10.7%), Seychelles (9.7%), Gabon (8.9%), Rwanda (8.3%) Angola (8.3%), Nigeria (8.3%), Namibia (7.8%) and South Africa (7.4%). Countries with lowest prevalence of less than 0.2% are Niger, Mauritania, Senegal, Guinea Bissau, Guinea, Liberia, Egypt, South Sudan, Somalia, and Tanzania. The general overview shows that Southern Africa, Central Africa, and East Africa regions have a cluster of highest prevalence, while West and East Africa regions have most of the lowest prevalence of alcohol.

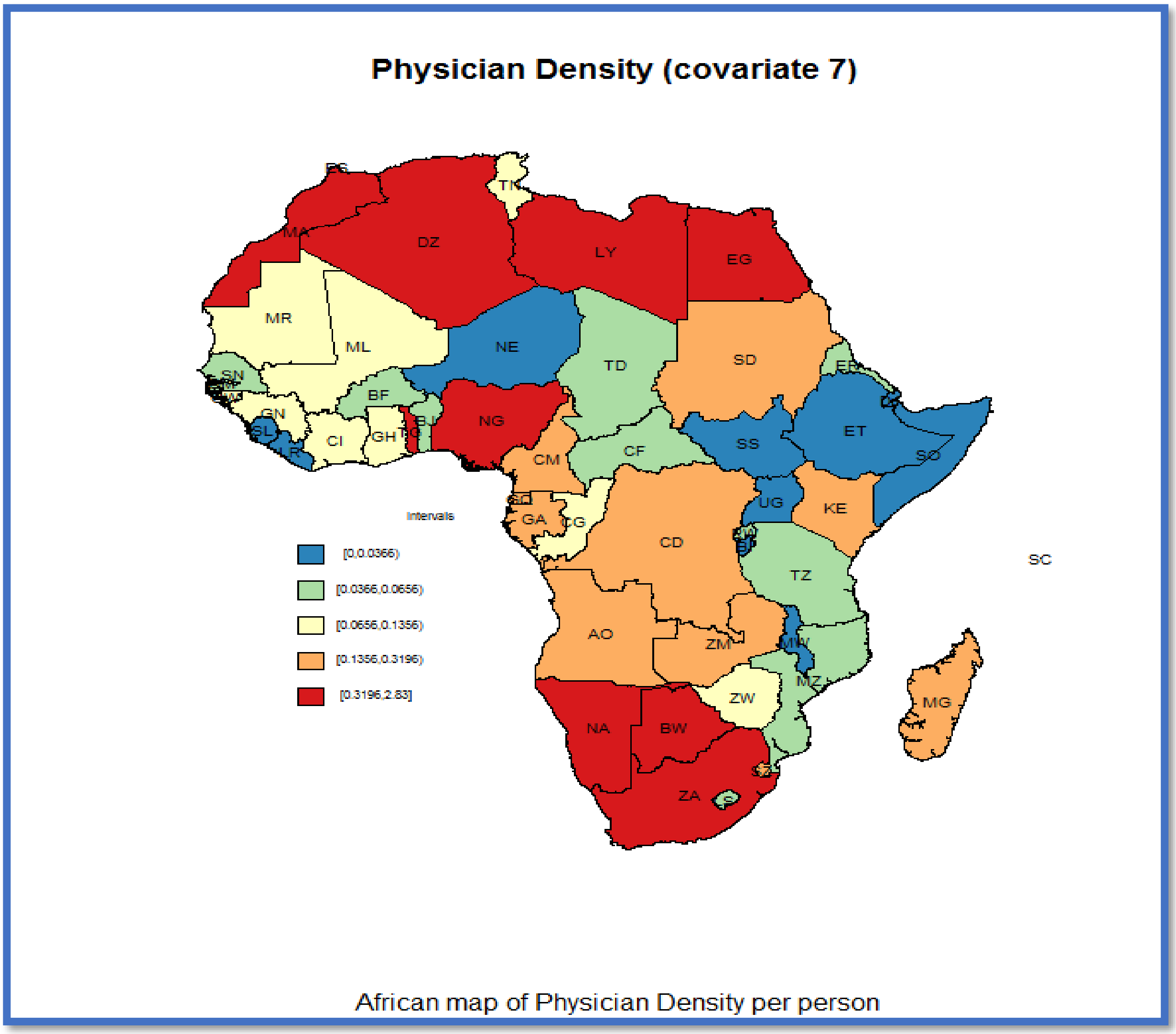


Figure 4.9. Spatial distribution of African countries' physician density per 1000 person

Figure 4.9 shows the distribution of the physician density per 1000 person on the continent. The countries with highest rate are Egypt (2.8%), Libya (1.9%), Togo (1.2%), Algeria (1.2%), Seychelles (1.1%), Mauritius (1.1%), South Africa (0.8%), Morocco (0.6%), Sao Tome Principe (0.5%), Nigeria (0.4%), Botswana (0.4%), Namibia (0.4%), both Cabo Verde and Equatorial Guinea having 0.3%. The countries with the lowest rates of less than 0.036% are South Sudan, Uganda, Ethiopia, Somalia, Djibouti, Burundi, Malawi, Niger, Chad Liberia and Sierra Leon. North, West and Southern region shows the highest number of countries with high rate of physician density while the East region shows the lowest.

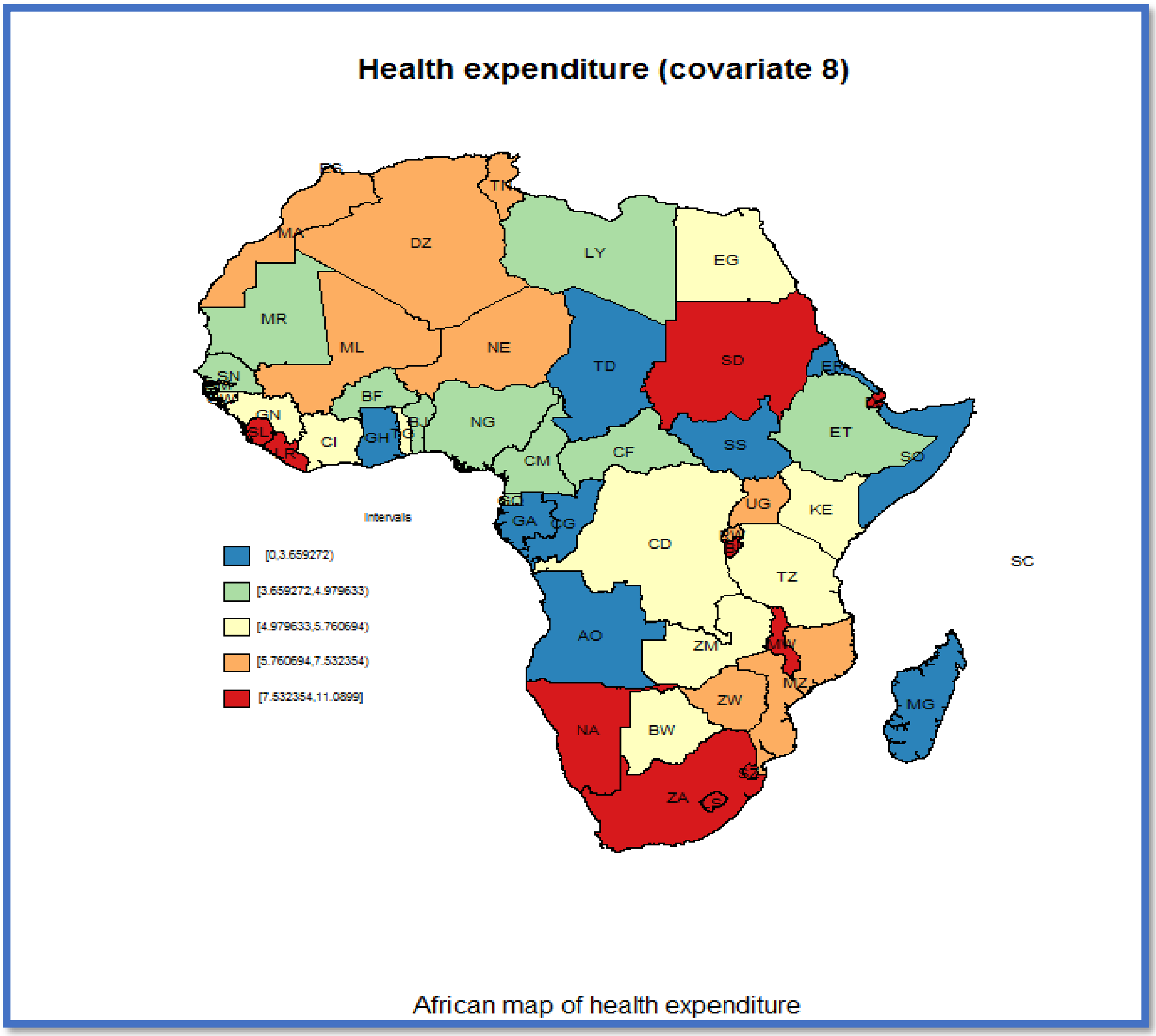


Figure 4.10. Spatial distribution of African countries' health expenditure

Figure 4.10 shows the distribution of the health expenditure in the continent. Countries with highest rate spent on health are Sierra Leone (11.1 %), Lesotho (10.6%), Djibouti (10.6%), Liberia (10.0%), Malawi (9.6%), Swaziland (9.3%), Namibia (8.9%), South Africa (8.9%), Sudan (8.4%), Sao Tome Principe (8.4%), Burundi (7.5%) and Rwanda (7.5%), while countries with lowest rate of health expenditure are Ghana (3.6%), Gabon (3.4%), Seychelles (3.4%), Eritrea (3.3%), Angola (3.3%), Madagascar (3.0%) South Sudan (2.7%) and Congo DRC (2.6%). General overview shows that Southern region has the highest rate of health expenditure.

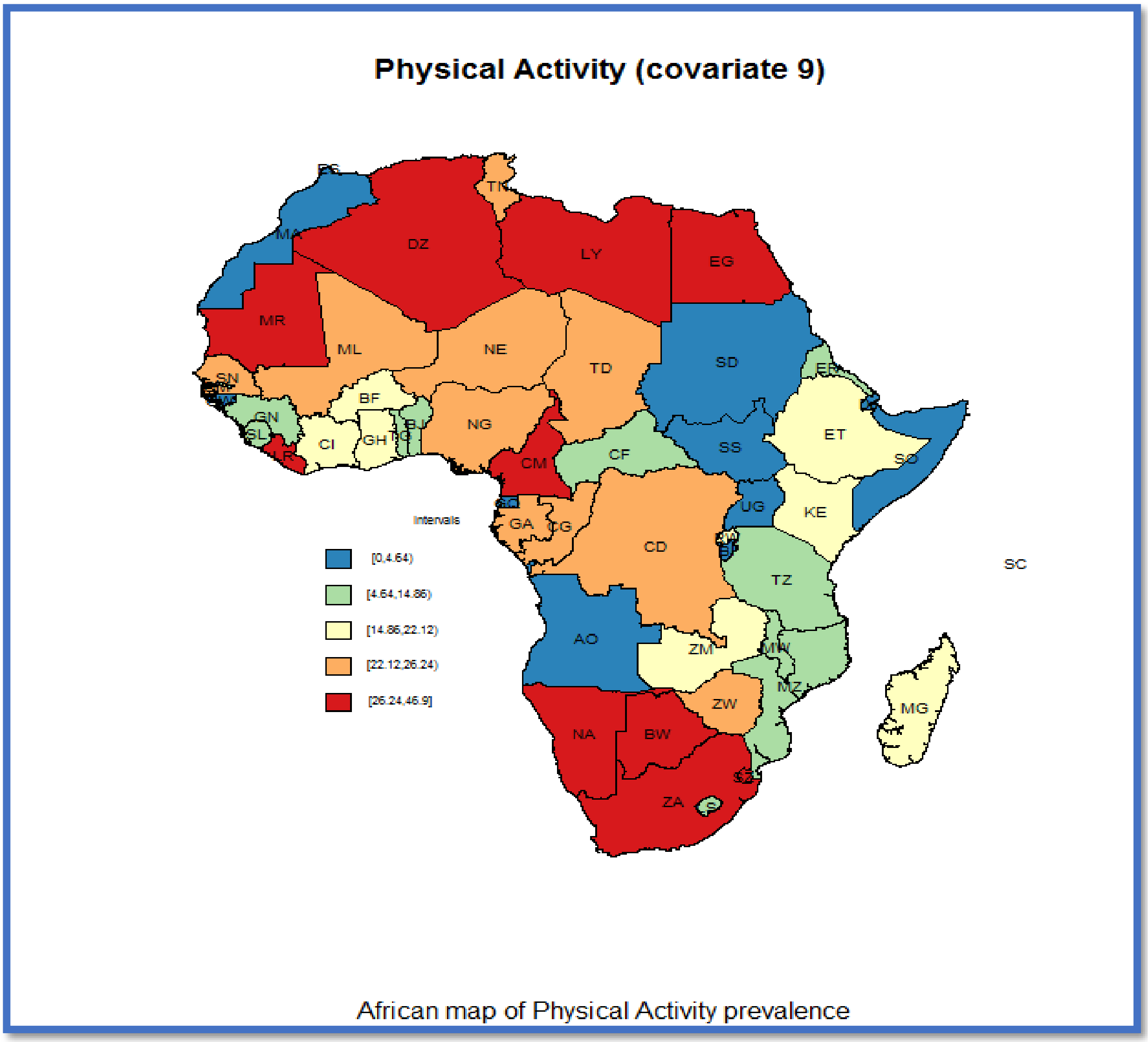


Figure 4.11. Spatial distribution of African countries' physical activity prevalence

Figure 4.11 shows the distribution of physical activity prevalence rate (covariate 9). Countries with highest physical activity rate are South Africa (46.9 %), Mauritania (45.1%), Libya (38.0%), Swaziland (36.8%), Algeria (34.4%), Egypt (32.3%), Namibia (31.8%), Cabo Verde (30.7%), Liberia (27.5%), Botswana (27.2%), Gabon (26.0%) and Cote D'Ivoire (26.0%), while countries with lowest physical activity rate (between 0% – 4.6%) are Morocco, Sudan, South Sudan, Uganda, Burundi, Somalia, Equatorial Guinea, Angola, Djibouti and Guinean Bissau. The highest prevalence is seen mostly within the North and Southern region, while the Central and West region has the lowest prevalence rate.

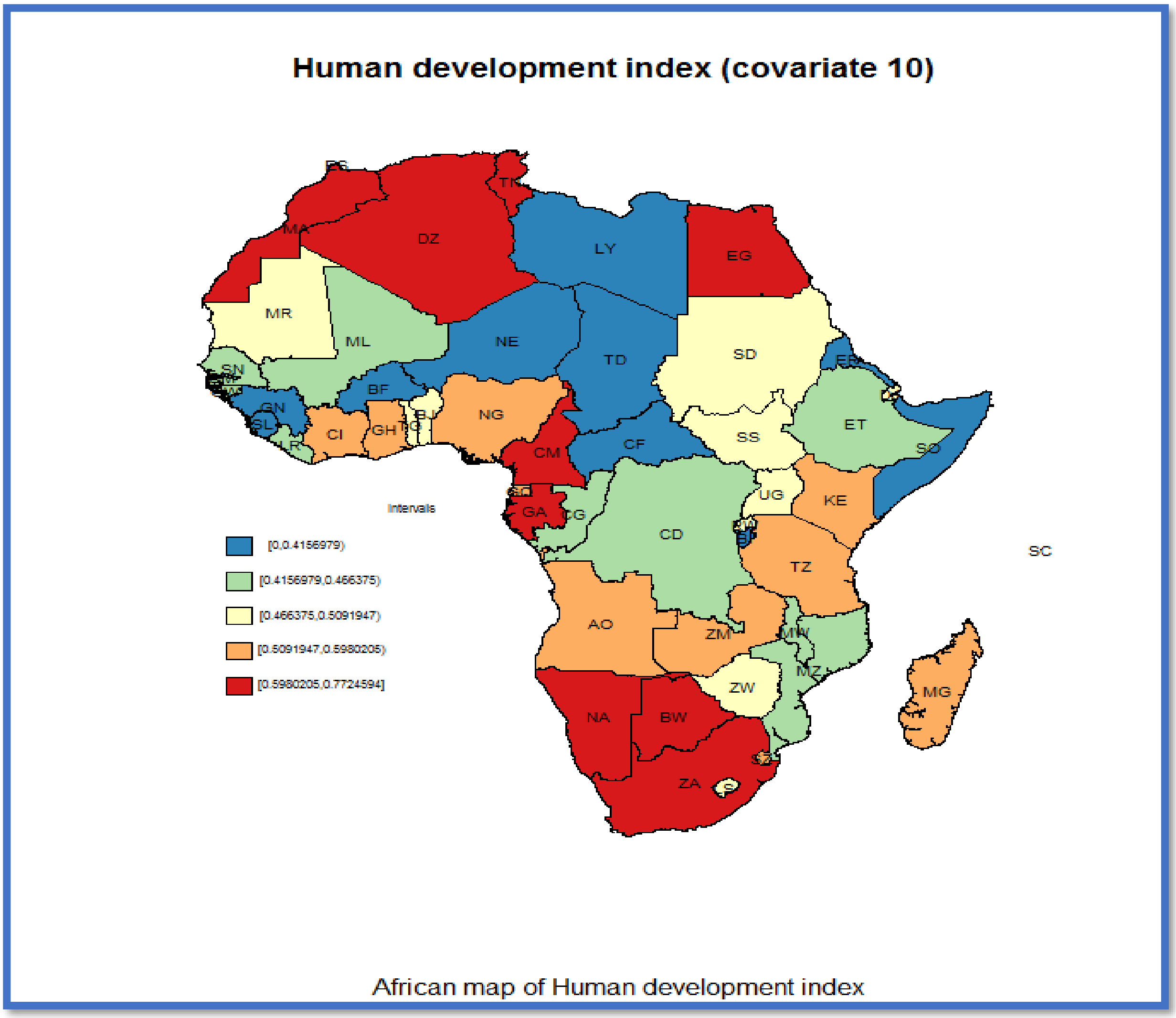


Figure 4.12. Spatial distribution of African countries' HDI

Figure 4.12 shows the distribution of the HDI (Covariate 10) in the continent. Countries with the highest are Mauritius (0.78%), Seychelles (0.78%), Algeria (0.74%), Tunisia (0.72%), Botswana (0.69%), Egypt (0.69%), Gabon (0.68%), South Africa (0.67%), Cabo Verde (0.65%), Morocco (0.63%), Namibia (0.63%) and Congo Republic (0.59%), and countries with lowest rates are Sierra Leone (0.41%), Guinea (0.41%), Burkina Faso (0.40%), Burundi (0.39%), Chad (0.39%), Eritrea (0.39%), Central Africa Republic (0.35%), Niger (0.35%), Somali (0.00%), Libya (0.00%). The general overview shows that Northern, Southern and few parts of Central Africa region have the highest distribution of HDI, while the West Africa region has the lowest rate on the continent.

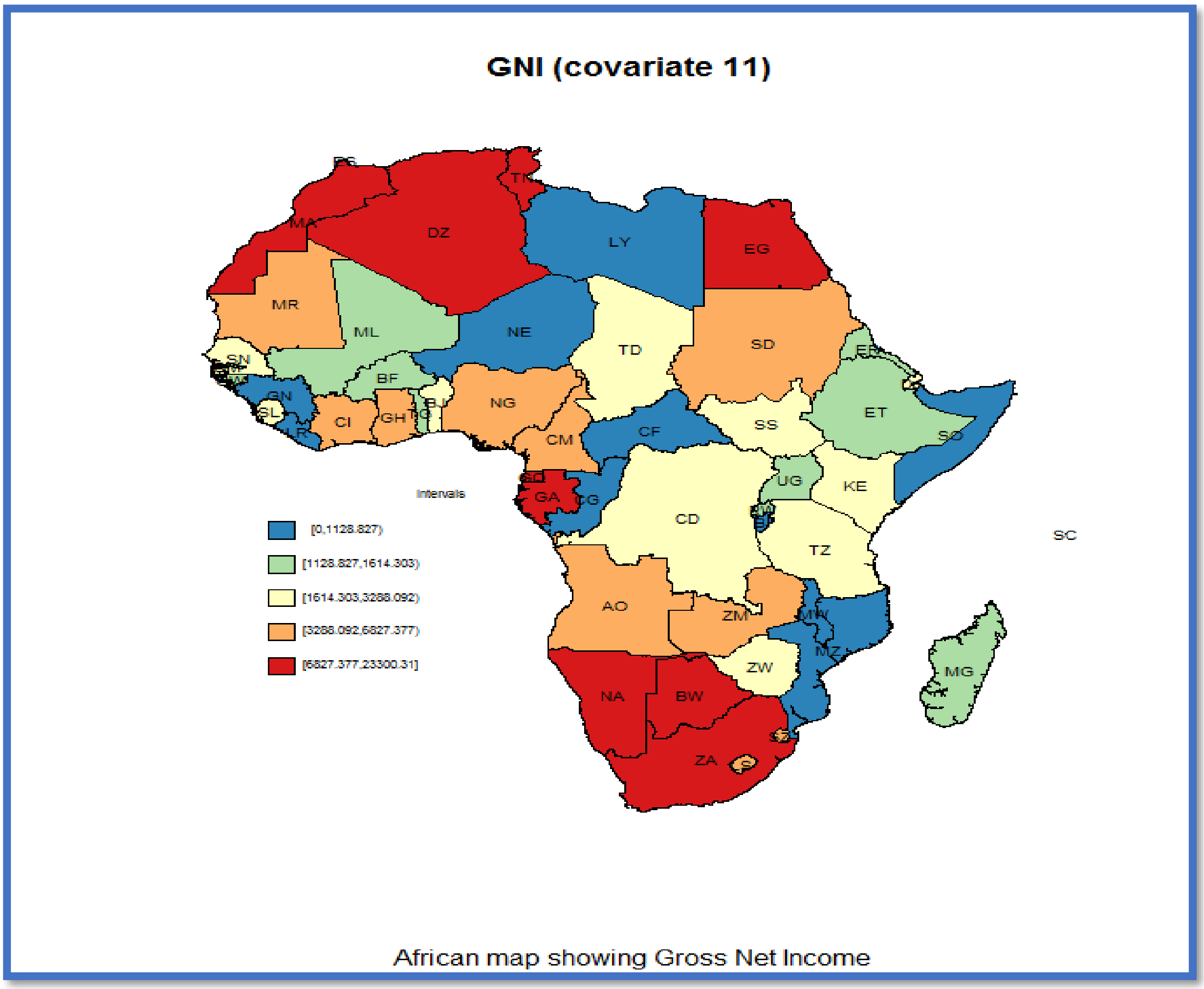


Figure 4.13. Spatial distribution of African countries' GNI

Figure 4.13 shows the distribution of GNI across the continent. Countries with the highest GNI are Seychelles (USD2330.31), Equatorial Guinea (USD 21 055.96), Mauritius (USD 17 469.78), Botswana (USD 16 646.23), Gabon (USD 16 366.93), Tunisia (USD 14 910.92), Algeria (USD 13 054.31), South Africa (USD 12 122.32), Egypt (USD 10 512.42), Namibia (USD 9 417.80), Morocco (USD 6 850.14). Countries with the lowest GNI are Mozambique (USD 1 123.44), Guinea (USD 1 095.76), Niger (USD 908.34), Liberia (USD804.95), Burundi (USD 758.18), Malawi(USD747.33), Congo DRC (USD680.47), Central Africa Republic (USD580.73), Somalia (USD0.00) and Libya (USD0.00). The general overview shows that the Northern, Southern, and Central regions have the highest distribution, while the West and Central Africa region shows the lowest distribution.

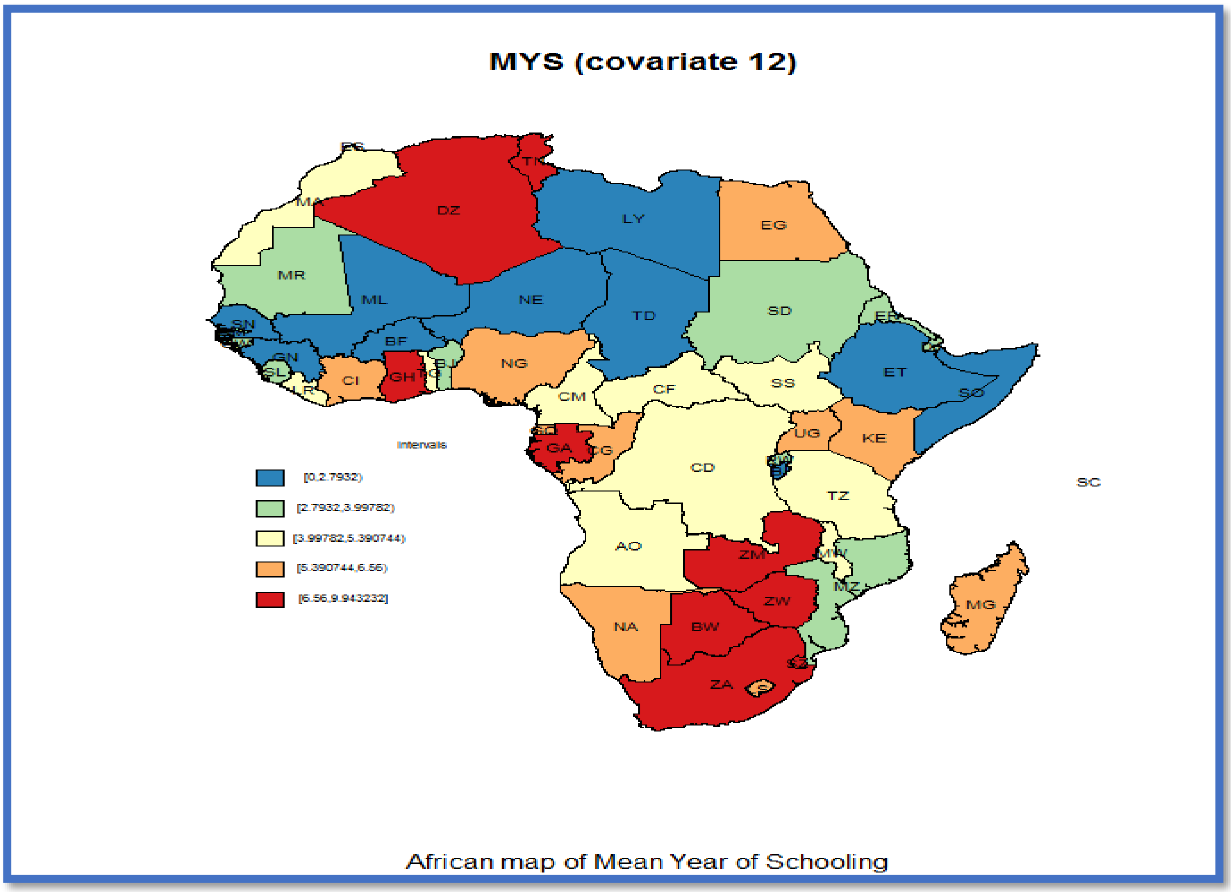


Figure 4.14. Mean Year of Schooling in Africa

Figure 4.14 shows the distribution of the Mean Year of Schooling (MYS) on the continent. MYS is defined as the average number of completed years of education of a country's population aged 25 years and older, excluding years spent repeating individual grades. MYS is an indicator of human capital available in an economy and countries and is comparable across populations. Education and literacy are key drivers of economic development, with also a strong correlation with health and wellbeing. A uniform pattern is seen across the continent. Countries with the highest rate are South Africa (9.94), Seychelles (9.41), Botswana (8.87), Mauritius (8.54), Gabon (7.79), Algeria (7.61), Tunisia (7.31), Zimbabwe (7.25), Swaziland (7.12), Ghana (6.99), Zambia (6.60). The countries with the lowest rate are Burundi (2.67), Senegal (2.48), Guinea (2.42), Ethiopia (2.41), Mali (2.05), Chad (1.93), Niger (1.45), Burkina Faso (1.37), Somalia (0.00), Libya (0.00). The general overview shows thee Southern Africa region to have the highest distribution while the West and East Africa region has the most cluster of lowest rate.

4.1.5. Normality Tests

4.1.5.1. Shapiro–Wilk Normality Test

Normality test was carried out using the Shapiro Wilk test, a non-parametric test to ascertain if the sample is from a normally distribution population. The result is shown in Table 4.2, at alpha significance level 0.05, from the *p*-values, we can conclude that the diabetes prevalence are not from a normally distributed population.

Table 4.2: Shapiro–Wilk normality test of diabetes prevalence across Africa

Country/Region	W (Shapiro test)	P-value	Median prevalence
Africa	0.76763	0.00000007472	4.05
West Africa	0.55696	0.00000631	2.20
Central Africa	0.79303	0.02409	6.45
East Africa	0.79956	0.01432	3.85
South Africa	0.61411	0.00008691	4.10
North Africa	0.73731	0.009332	8.90

However, since the test is biased by sample size, the test may be significant from a normal distribution in any large samples. That is why we have Kolmogorov Smirnov test (n>50)

A Q-Q plot can be used to verify the test. As an additional check of the normality of the variables visually, a Q-Q plot is a visual plot of the observed value versus the expected normal value. Figure 4.15 shows the Normal Q-Q plot of diabetes prevalence. The results hence show that diabetes is not normally distributed.

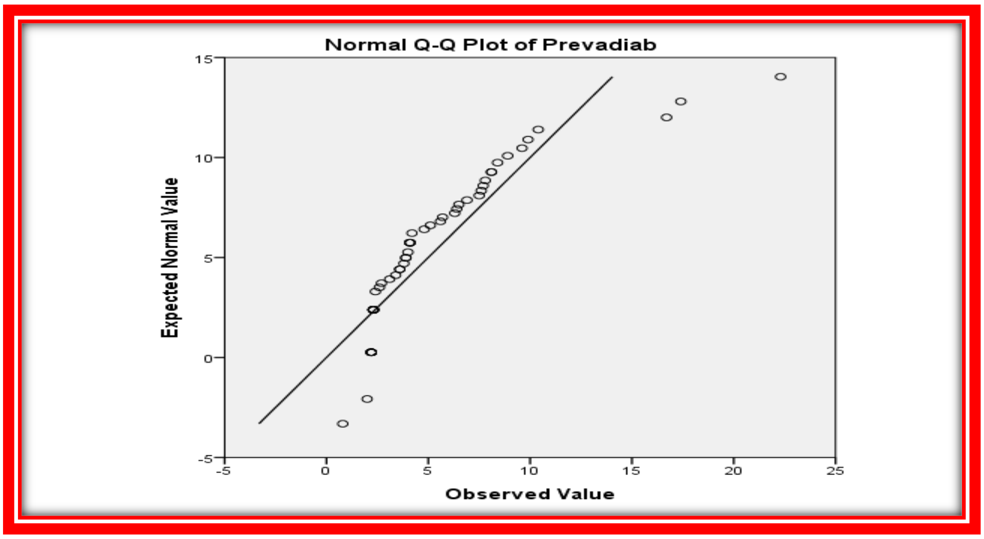


Figure 4.15. Normal Q-Q plot of diabetes prevalence.

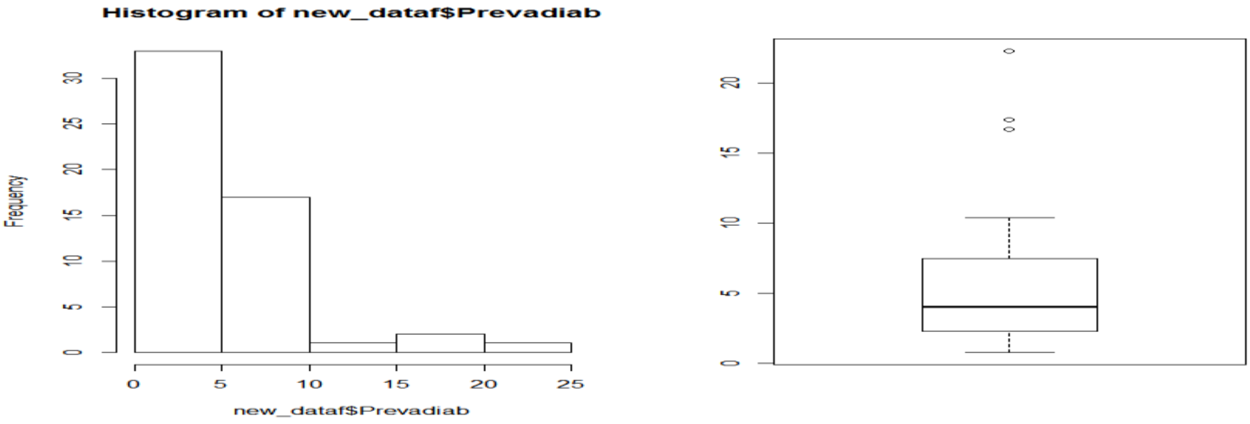


Figure 4.16. Bar, Box, and Histogram of diabetes prevalence.

4.1.5.2. Kruskal -Wallis Test by Ranks

This is a non-parametric method for testing whether samples originate from the same distribution and is mostly used for comparing more than two independent samples of equal or different sizes. Thus, in performing the comparison between multiple groups, we cannot run an ANOVA test for multiple comparisons if the group does not follow a normal distribution. Instead, the Kruskal-Wallis test is used.

The result gave a Kruskal Wallis value of 35.448 with df = 4 and the *p*-value of 0.000000379. In comparing the value of median prevalence for the region in Table 4.2 with the Kruskal-Wallis test, we then conclude that there is a significant difference among the country regions. This means that the prevalence level in the West Africa region can be different from that in the South Africa region. Kruskal Wallis test function is then used to compare the independent diabetes prevalence.

We can then conclude that diabetes is independently distributed across the continent.

4.1.5.3. Spearman Rho Correlation Test

Spearman's rho correlation test was used to measure statistical dependency of the data set. The Table 4.3 below show the result of the spearman correlation test of diabetes and the risk factors. If the p-value < $\alpha = 0.05$, then we say there is a significant relationship between the variables. We also look at if the rho value is positive (+) or negative (-). Lastly, when

- r = 0.1 to 0.29, the strength of correlation is said to be small
- r = 0.30 to 0.49, the strength of correlation is said to be medium
- r = 0.50 to 1, the strength of correlation is said to be large

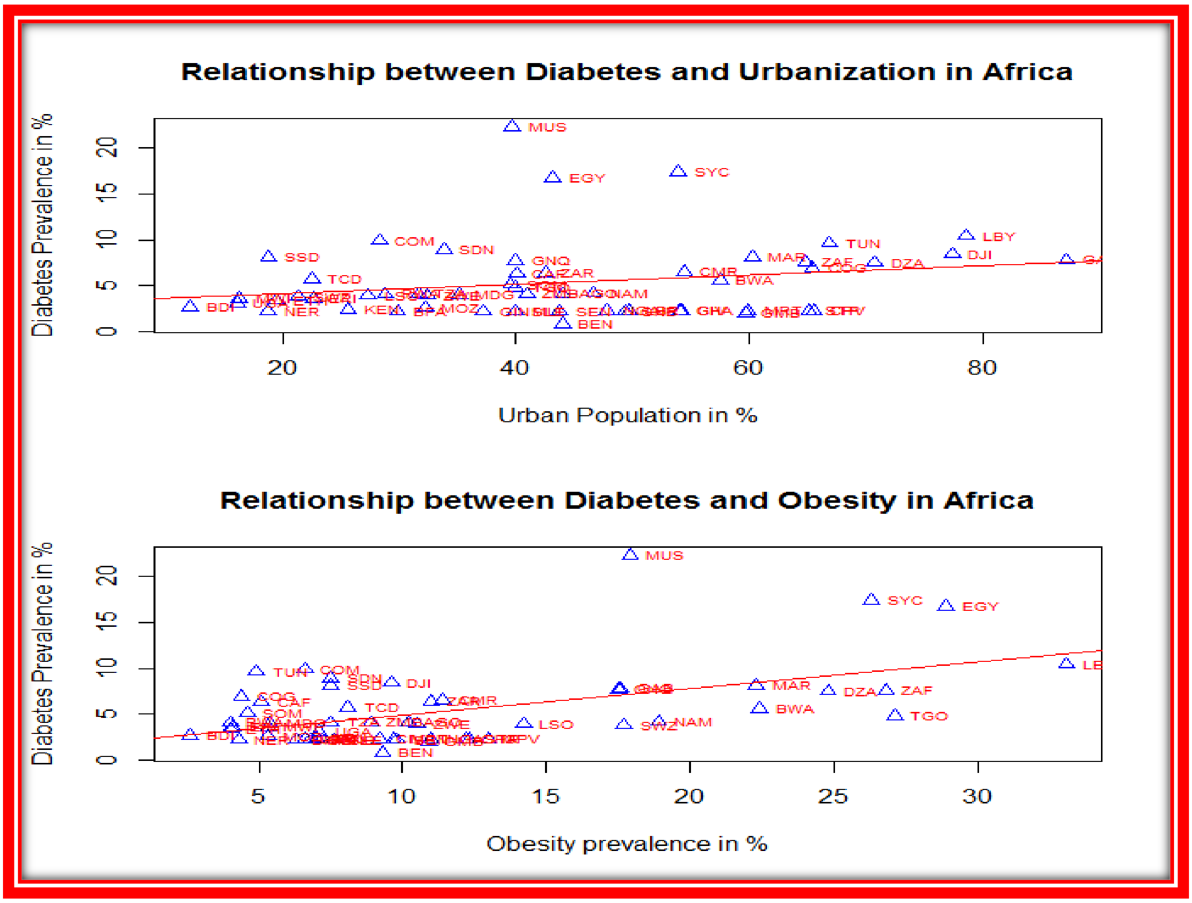
Table 4.3: Spearman Rho correlation test of diabetes and risk factors

Variables	Rho (r)	p-value	Interpretation
Obesity	0.335	0.013	Significant, positive medium strength correlation
Health expenditure	-0.132	0.340	Non-significant, negative, small strength correlation
GDP	0.573	0.0000058	Significant, positive, large strength correlation
Alcohol	-0.015	0.915	Non-significant, negative, small correlation
Population age	0.504	0.0001	Significant, positive, large strength correlation
Age dependency	0.534	0.000032	Significant, positive, large strength correlation
Urban population	0.190	0.169	Non-significant, positive, small strength correlation
Physical activity	0.050	0.718	Non-significant, positive, small strength correlation
Physical density	0.389	0.0036	Significant, positive, medium
HDI	0.439	0.00090	Significant, positive, medium strength correlation
MYSC	0.470	0.00034	Significant, positive, medium strength correlation
GNI	0.389	0.0036	Significant, positive, medium strength correlation

The results show that obesity, physician density, HDI, MYSC, and GNI are significantly correlated to diabetes with positive medium strength correlations. GDP, population age, and age dependency ratio are significant and positive with larger strength correlations with diabetes. Health expenditure and alcohol are not significantly correlated with diabetes. Urban population growth and physical activity show a nonsignificant correlation as well.

4.1.5.4. Relationship Between Diabetes and Risk Factors

The scatter plot is the best way to assess linearity between two numeric variables. From a scatter plot, the strength, direction and the form of the relationship can be identified. The analysis was conducted in R using library. Figure 4.17 to 4.20 shows the scatter plots for diabetes prevalence (dependent variable) and its association with the risk factors (independent variables). This is a linear representation of the results showing the relationship between diabetes and the risk factors as well as country particularities.



Relationship between Diabetes and Urbanization in Africa

Relationship between Diabetes and Obesity in Africa

Figure 4.17. Relationship between diabetes prevalence with Obesity and Urbanisation

The results in the Figure 4.17 show that countries such as Egypt and Seychelles have high levels of diabetes prevalence as well as high percentage of obesity, while countries such as South Arica and Togo have high obesity prevalence, but low diabetes.

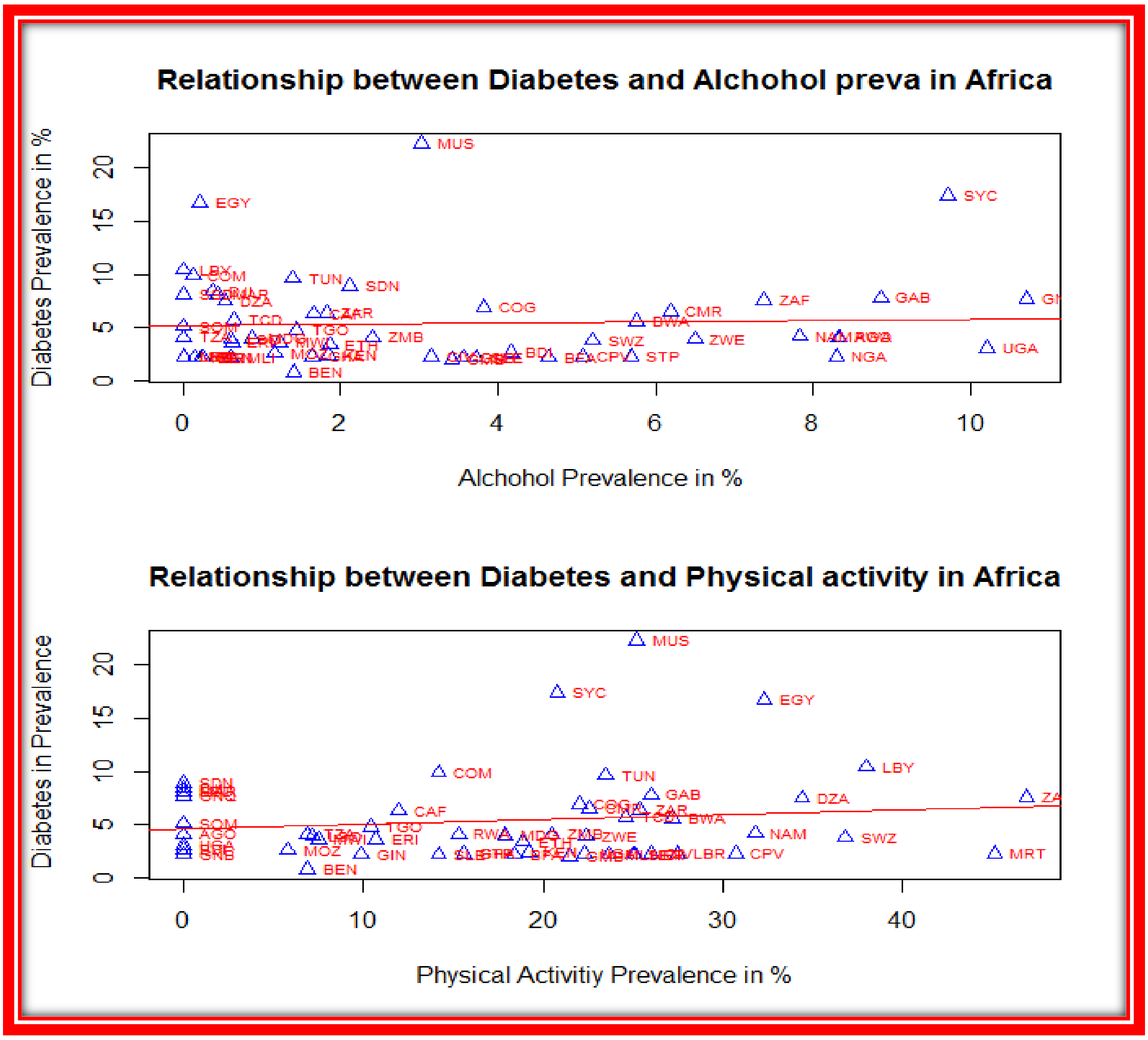


Figure 4.18. Relationship between diabetes prevalence with Alcohol consumption prevalence and Physical Activity in Africa

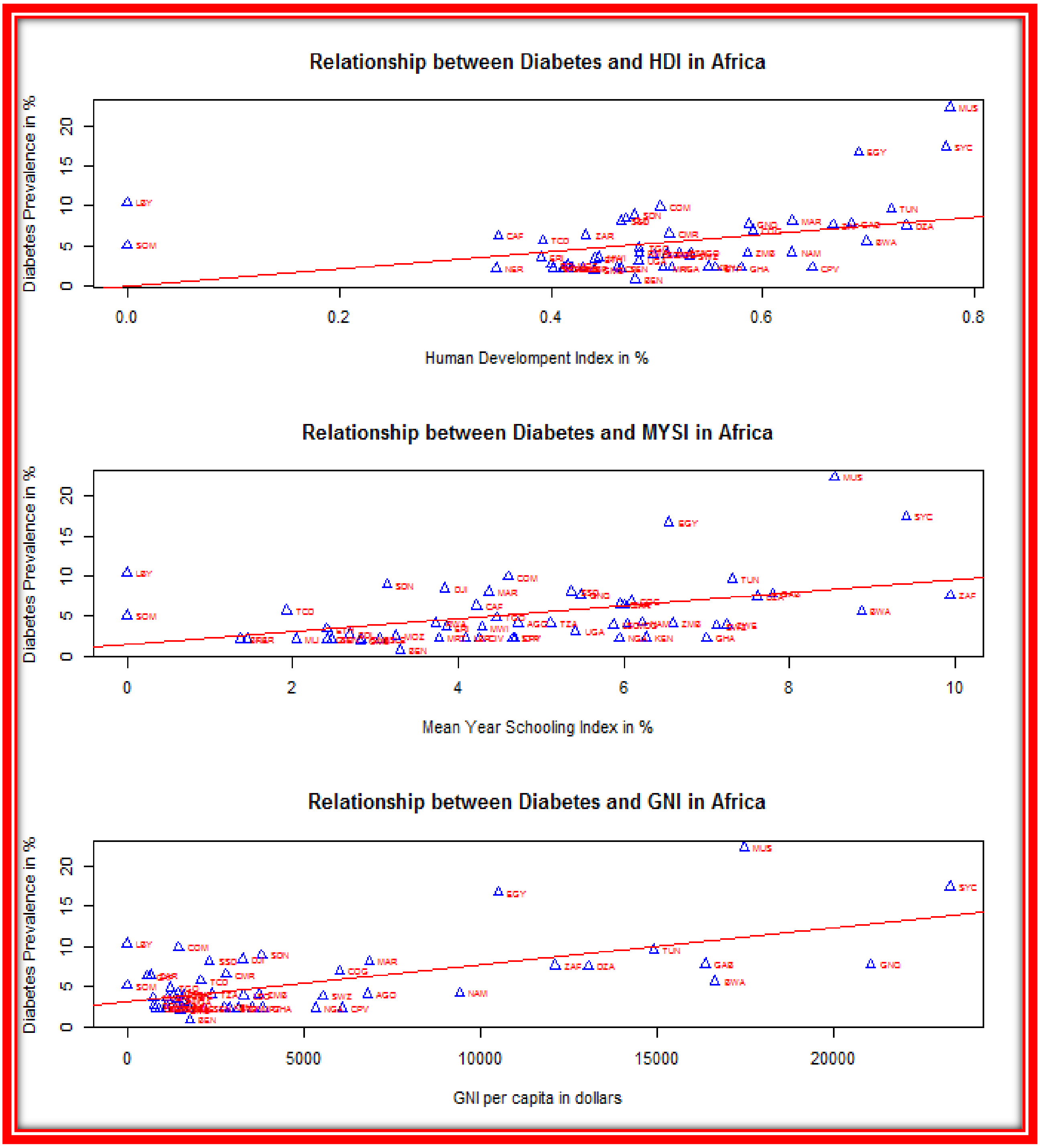


Figure 4.19. Relationship between diabetes prevalence with HDI, GNI, and MYSI

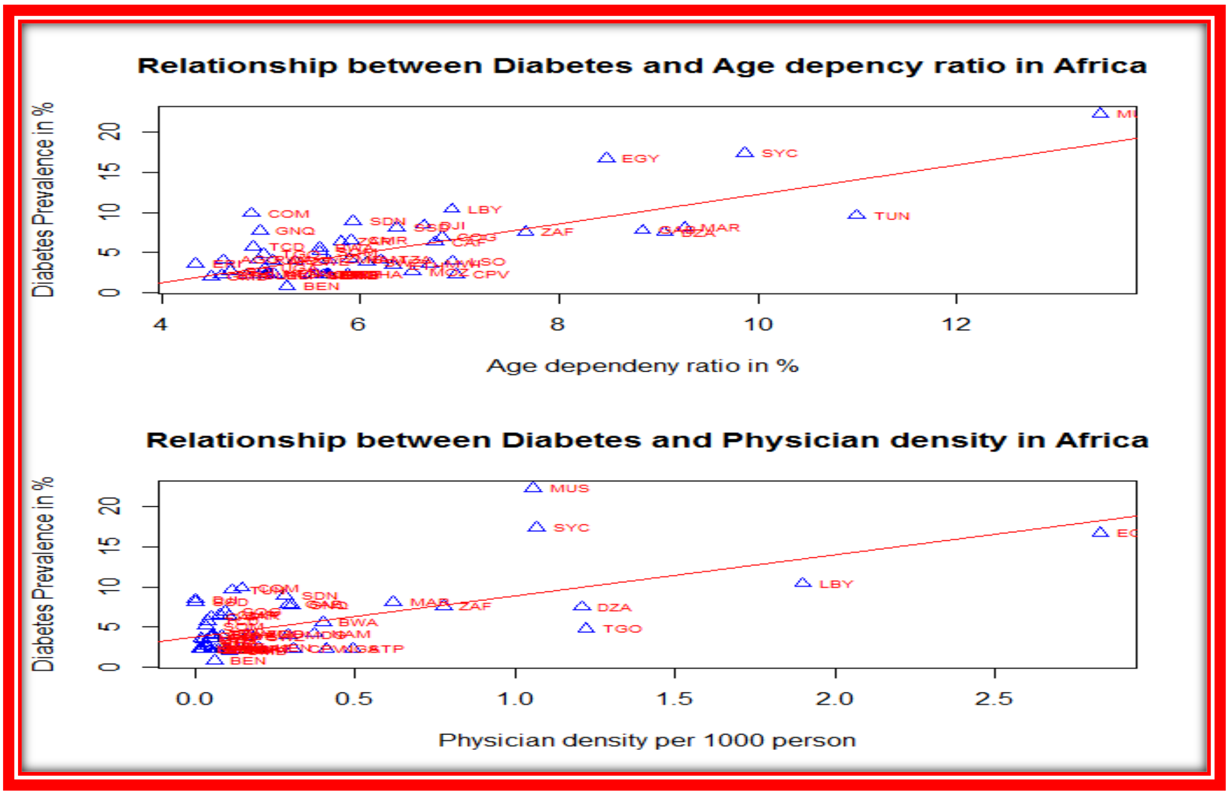


Figure 4.20. Relationship between diabetes prevalence with Age dependency ratio y and Physician density

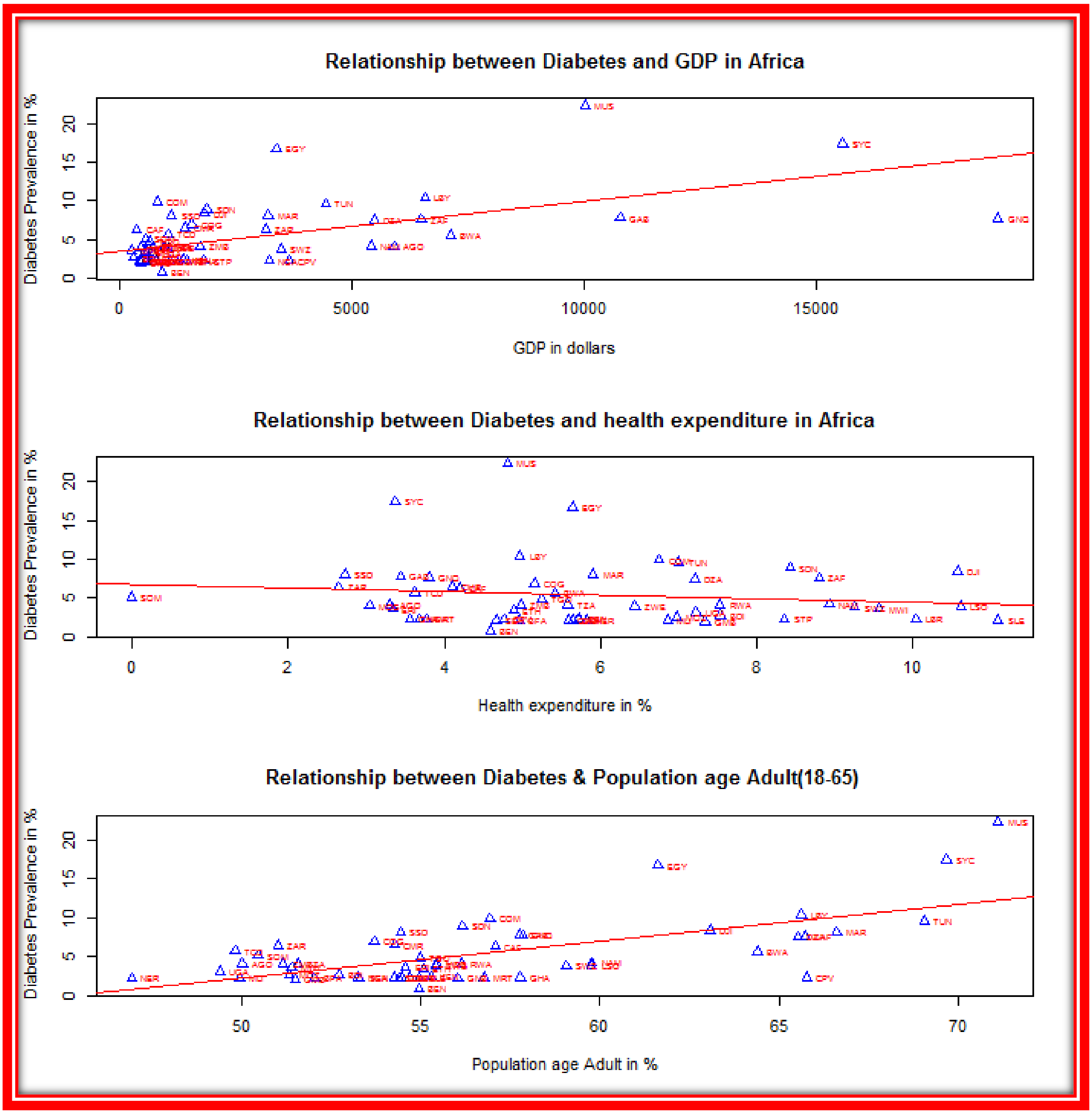


Figure 4.21. Relationship between diabetes prevalence with GDP, Health expenditure, and Population age

Figures 4.17 – 4.21 visually illustrate the relationship between the significant risk factors at a descriptive level and diabetes prevalence, with a linear association observed in all the variables. It also visually shows the outlier countries within the relationship between each covariate and diabetes prevalence. Countries that have high diabetes prevalence are identified both at the continent at large and within the regions. Three countries are identified which have a diabetes prevalence rate above 15%, and they are Egypt (16.7%), Seychelles (17.4%), and Mauritius (22.3%). Within the region, East Africa shows Comoros and Djibouti as outliers, as their prevalence rate was above 8%, 9.9%, and 8.4% respectively, and the highest in the region. In the West Africa region, Togo with a 4.8% diabetes prevalence (a little above 4%) and Benin 0.8% below the mean prevalence of 2 in the region are both outliers compared to other countries in the region. In the Central Africa region, Sao Tome and Principe is an outlier and its diabetes prevalence are below 3%. In the North Africa region, Egypt is the only outlier with a prevalence rate of 16.7%, and the highest in the region. In the South Africa region, Mauritius, Seychelles, and South Africa are outliers with diabetes prevalence rates of 22.4%, 17.4%, and 7.6% respectively, and the highest in the region.

Upon visual inspection, the relationship appears to be linear, with all variables having negative direction and looks moderately strong, except alcohol and health expenditure. The strength of the relationship was quantified using spearman rho correlation test.

4.2. Clustering of Diabetes prevalence in Africa

Cluster analysis was performed and Tables 4.4 and 4.5 show the results of K-means and Ward’s cluster analysis using R, while Figures 4.22 and 4.23 show the dendrogram result of Ward’s and Centroid methods of cluster analysis.

4.2.1. K-Means Results

Table 4.4: Summary of clustering of Diabetes in Africa using K - mean clustering techniques

Countries in clusters	n	Mean	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Cluster 1: Algeria, Comoros, Djoubti, Gabon, Libya, Morocco, South Africa, South Sudan, Sudan, Tunisia, Equatorial Guinea.	11	6.97	2.30	4.13	6.70	7.95	16.70
Cluster 2: Angola, Botswana, Cameroon, Central Africa, Chad, Congo DRC, Congo, Eritrea, Ethiopia, Lesotho, Madagascar, Malawi, Namibia, Rwanda, Somalia, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe	21	3.58	0.80	2.20	3.25	4.10	9.90
Cluster 3: Benin, Burkina-Faso, Burundi, Carbo Verde, Cote D'Ivoire, Gambia, Ghana, Guinea, Guinea Bissau, Kenya, Liberia, Mali, Nigeria, Sao Tome Principe, Senegal, Serial Leone	19	10.74	5.60	7.60	7.80	10.40	22.30
Cluster 4: Egypt and Seychelles	2	4.25	2.20	2.30	2.40	5.70	8.90
Cluster 5: Mauritius	1	12.55	7.70	10.12	12.55	14.98	17.40

The diabetes prevalence rates differ significantly between the clusters (Kruskal Wallis test: *p*=0.01). From Table 4.4. Cluster 1 has a total of eleven (11) countries, with mean of 6.97, median of 6.70, minimum value of 2.30, and maximum value of 16.70. Cluster 2 has a total of twenty-one (21) countries with mean of 3.58, median of 3.25, and minimum and maximum values of 0.80 and 9.90 respectively. Cluster 3 has nineteen (19) countries with mean of 10.75, median of7.80, and minimum and maximum values of 5.60 and 22.30 respectively. Cluster 4 comprises of two (2) countries (Egypt and Seychelles) with mean of 4.25, median of 2.40, 1st and 3rd quartiles of 2.30 and 5.70 respectively. The minimum and maximum value in this cluster is 2.30 and 5.70 respectively. Cluster 5 has one country (Mauritius) with mean and Median of 12.55, 1st and 3rd quartiles are 10.12 and 14.98, respectively, and the minimum and maximum values are 7.70 and 17.40. The highest mean diabetes prevalence rate is found in Cluster 5 followed by Cluster 3. The lowest mean is found in Cluster 2 (3.58).

4.2.2. Ward Clustering Result

Table 4.5: Summary of clustering of diabetes in Africa using Ward clustering technique

Countries in same cluster	N	Mean	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Cluster 1: Algeria, Comoros, Djoubti, Equatorial Guinea, Gabon, Libya, Morocco, South Africa, South Sudan, Sudan and Tunisia	11	7.00	4.10	4.90	7.5	8.6	10.4
Cluster 2: Angola, Eritrea, Ethiopia, Lesotho, Madagascar, Malawi, Namibia, Rwanda, Swaziland, Tanzania, Zambia, Zimbabwe	12	3.81	0.80	2.30	3.10	4.10	9.90
Cluster 3: Benin, Burkina-Faso, Burundi, Cabo Verde, Cote D'Ivoire. Gambia, the, Ghana, Guinea, Guinea Bissau, Kenya, Liberia, Mali, Mauritania, Mozambique, Niger, Nigeria, Sao Tome Principe, Senegal, Sierra Leone, Uganda	20	7.38	2.30	4.48	6.70	7.80	16.70
Cluster 4: Botswana, Cameroon, Central Africa, Chad, Congo DRC, Congo Rep, Somalia, Togo	8	12.55	7.70	10.12	12.55	14.98	17.40
Cluster 5: Egypt, Mauritius and Seychelles	3	15.05	7.80	11.42	15.05	18.68	22.30

Table 4.5 above using the ward techniques showed that diabetes prevalence rate differs significantly between the clusters (Kruskal Wallis test: *p*=0.01).

Cluster 1 has eleven countries with mean of 7.0, median of 7.5, 1st and 3rd quartiles of 4.90 and 8.6 respectively, and minimum and maximum values of 4.10 and 10.4 respectively. Cluster 2 comprises of twelve countries with mean of 3.81, median = 3.10, 1st and 3rd quartiles of 2.30 and 4.10 respectively, and minimum and maximum values of 0.80 and 9.90 respectively. Cluster 3 have twenty countries with mean of 7.38, median 6.70, 1st and 3rd quartiles are 4.48 and 7.80 respectively, while the minimum and maximum values are 2.0 and 16.70 respectively. Cluster 4 has eight countries with mean and median of 12.55. The 1st and 3rd Quartiles are 10.12 and 14.95 respectively, and the minimum and maximum values are 7.70 and 17.40, respectively. Cluster 5 has three countries (Egypt, Seychelles, and Mauritius) with mean and median = 15.05, minimum and maximum value of 7.80 and 22.30 respectively, and the 1st and 3rd quartiles of 11.42 and 18.86, respectively. The highest mean diabetes prevalence rate is found in Cluster 5 followed by Cluster 4 which are characterised with developed countries, high GDP and urbanisation. The lowest mean is found in Cluster 2, which is characterised with developing countries.

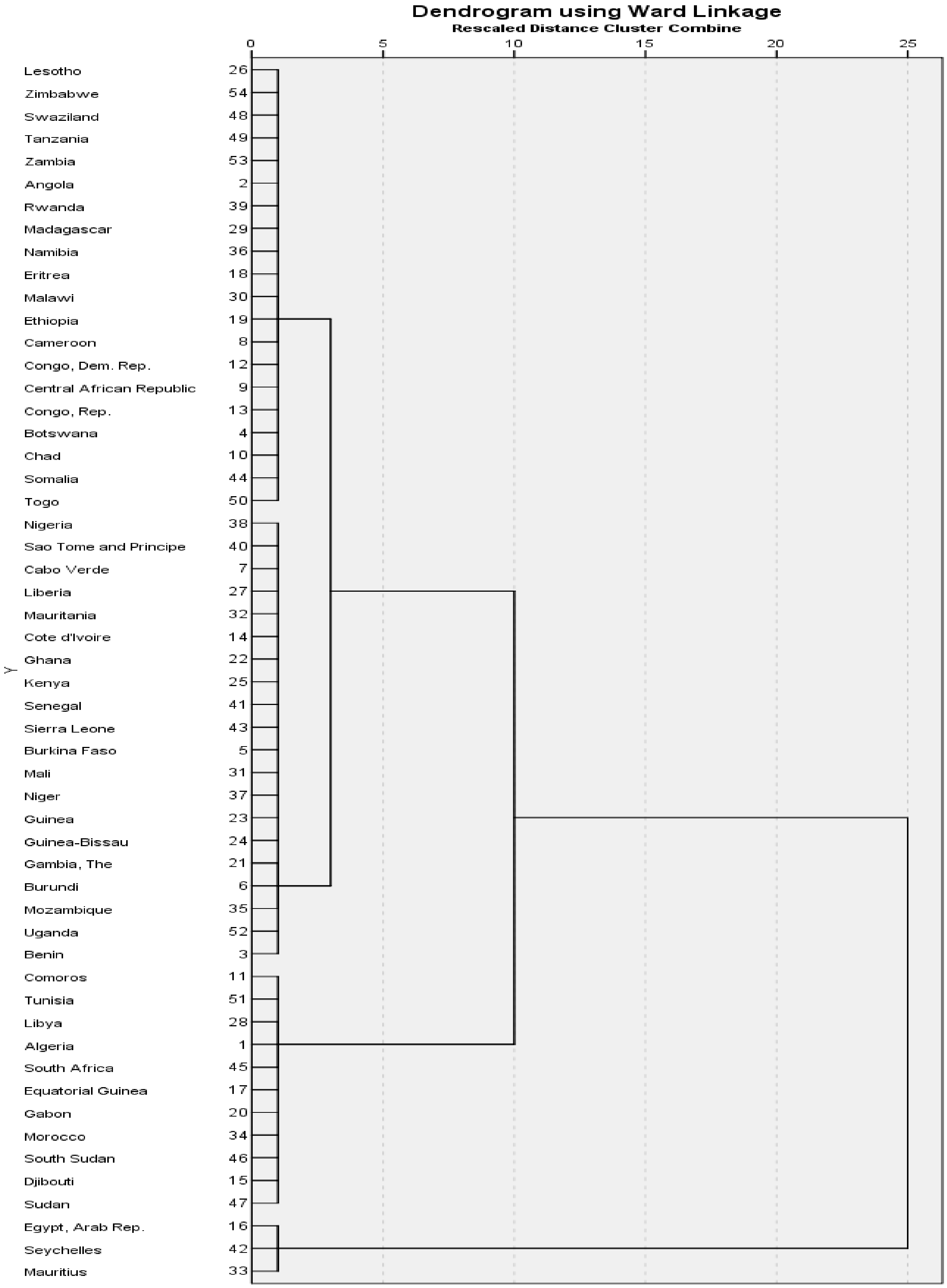


Figure 4.22. Dendrogram showing cluster of countries using Ward technique

The dendrogram in Figure 4.22 shows Ward's method of cluster analysis. The Ward's dendrogram result shows five clusters with cluster 1 having eleven countries, these are: Algeria, Comoros, Djoubti, Equatorial Guinea, Gabon, Libya, Morocco, South Africa, South Sudan, Sudan and Tunisia. Cluster 2 has twelve countries, which are Angola, Eritrea, Ethiopia, Lesotho, Madagascar, Malawi, Namibia, Rwanda, Swaziland, Tanzania, Zambia, and Zimbabwe. Cluster 3 is made up of twenty countries: Benin, Burkina Faso, Burundi, Cabo Verde, Cote dvoire, Gambia, Ghana, Guinea, Guinea Bissau, Kenya, Liberia, Mali, Mauritania, Mozambique, Niger, Nigeria, Sao Tome Principe, Senegal, Serial Leone and Uganda. Cluster 4 has eight countries. These are Togo, Somalia, Congo, Congo DRC, Chad, Central Africa Republic, Cameroon and Botswana. Cluster 5 has three countries which are Egypt, Mauritius and Seychelles.

4.2.3. Centroid Linkage Result

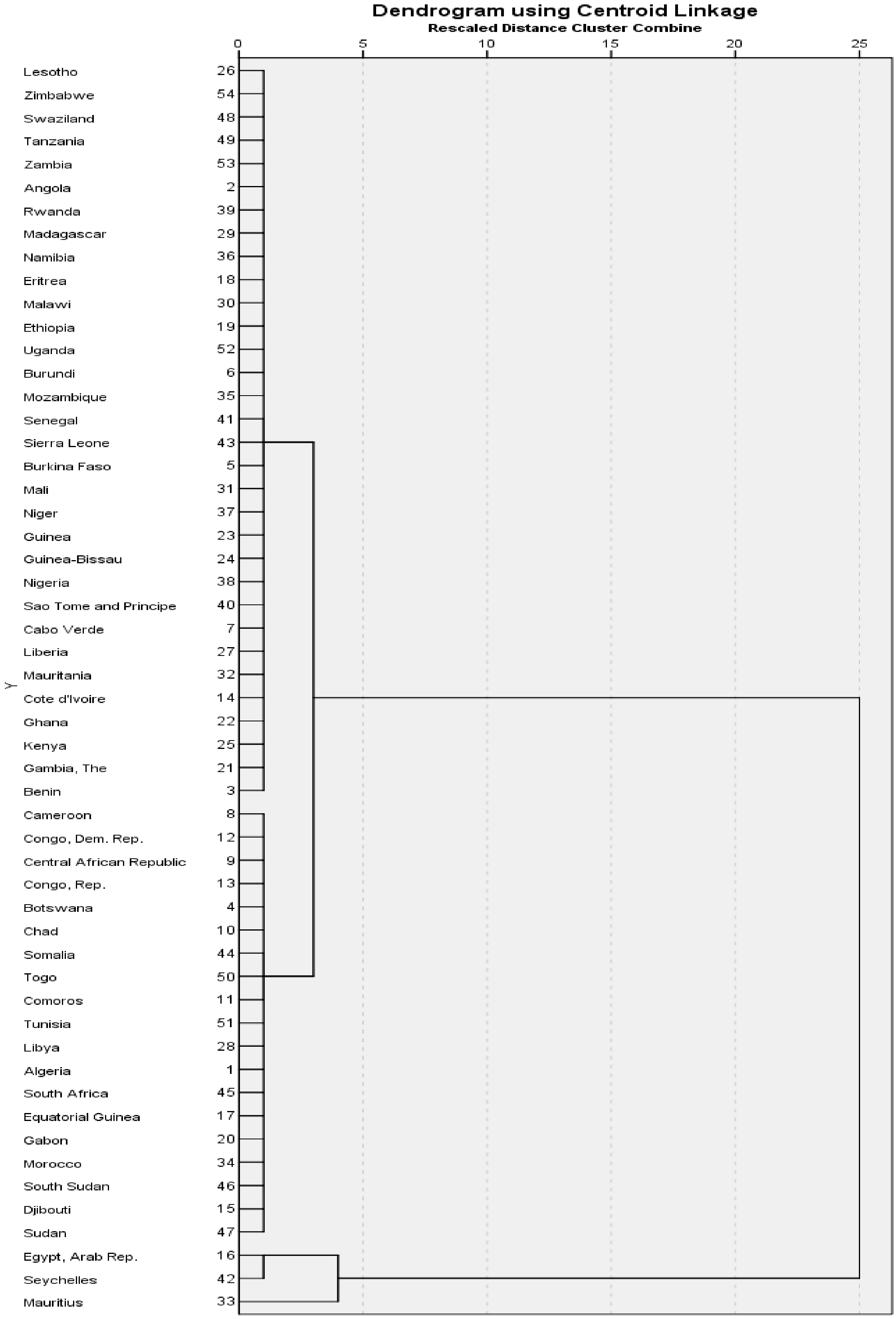


Figure 4.23. Dendrogram showing cluster of countries using Centroid technique

Using the centroid method of cluster analysis, Figure 4.23 above shows the dendrogram result. Five clusters were created, with Cluster 1 having eleven countries, which are Algeria, Comoros, Djibouti, Equatorial Guinea, Gabon, Libya, Morocco, South Africa, South Sudan, Sudan and Tunisia. Cluster 2 has the largest number of countries at thirty-two. Cluster 3 has a total of eight countries, which are, Botswana, Cameroon, Central Africa Republic, Chad, Congo DRC, Congo Republic, Somalia and Togo. Cluster 4 has two countries: Egypt and Seychelles, and Cluster 5 has only Mauritius.

Table 4.6: Overall grouping of countries with similar pattern of diabetes in Africa according to clustering analysis

CLUSTERS	COUNTRIES
Cluster 1 (11 Countries)	Algeria, Comoros, Djoubti, Equatorial Guinea, Gabon, Libya, Morocco, South Africa, South Sudan, Sudan and Tunisia
Cluster 2 (13 Countries)	Angola, Eritrea, Ethiopia, Lesotho, Madagascar, Malawi, Namibia, Rwanda, Swaziland, Tanzania, Uganda, Zambia and Zimbabwe
Cluster 3 (19 Countries)	Benin, Burkina-Faso, Burundi, Cabo Verde, Cote D'Ivoire. Gambia, the, Ghana, Guinea, Guinea Bissau, Kenya, Liberia, Mali, Mauritania, Mozambique, Niger, Nigeria, Sao Tome Principe, Senegal and Sierra Leone
Cluster 4 (8)	Botswana, Cameroon, Central Africa, Chad, Congo DRC, Congo Rep, Somalia and Togo
Cluster 5 (3)	Egypt, Seychelles, Mauritius

Table 4.6 describe the overall grouping of countries with similar pattern of diabetes as derived from the three clustering methods used.

Table 4.7: Overall grouping of countries with similar pattern of socioeconomic risk factors in Africa

Risk variable	Countries
Highest GDP	Equatorial Guinea, Seychelles, Gabon, Mauritania, Botswana, Libya, South Africa, Angola, Algeria, Namibia, Tunisia, Cabo Verde, Swaziland, Egypt, Nigeria, Morocco and Congo DRC
High population age (60 above)	Mauritania, Seychelles, Tunisia, Morocco, Cabo Verve, South Africa, Libya, Algeria, Botswana, Djoubti and Egypt
High Physician density	Egypt, Libya, Togo, Nigeria, Seychelles, Mauritania and South Africa, Morocco.
High Urbanization	Gabon, Libya, Djoubti, Algeria, Tunisia, Cabo Verde, Congo Rep, Sao Tome Principe, South Africa, Morocco, Mauritania and Gambia

Table 4.7 described the overall grouping of countries with similar pattern of socio economic risk factors, having identified four consistently significant risk factors (GDP, population age, physician density and urbanization) from the regression analysis in section 4.4.

4.2.4. Conclusion

In comparing the combination of three cluster analysis methods (k-means, ward, and centroid), it was found that there is similarity in the countries that are members of Cluster 4 in the Centroid technique and in Ward's technique. Likewise, Cluster 2 is similar for both methods. Egypt, Seychelles, Mauritius are characterised as a developed, rich country which is seen in there GDP, GNI. Clusters 2 are also characterised with countries that are developing; this may explain why diabetes is low in these countries.

At the end of this analysis, five groups were identified based on consistency of countries, though some countries were difficult to group.

4.3. Spatial Analysis Results

In this section, we consider already existing grouping of Africa countries into regions to check for similarities in terms of diabetes prevalence.

4.3.1. Global Moran Indices (GMI) and Moran Indices (MI)

This tool measures spatial autocorrelation (feature similarity) based on both feature locations and feature values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I Index value and both a Z score and p-value evaluating the significance of that index. In general, a Moran's Index value near +1.0 indicates clustering while an index value near -1.0 indicates dispersion

Table 4.8 shows the summary of Moran tests for the entire Africa continent and the five regions. The results of the Moran I indices provide the expected, observed and the p-value.

Table 4.8: Global Moran's Indices and spatial patterns of diabetes prevalence in Africa for 2015 (under assumption of normality)

Region	Number of Countries used in calculation	P-value	Standard Deviation (SD)	Expected value (E)	Observed value (O) (Z score)	Spatial Autocorrelation Pattern
Africa continent (combined)	54	0.000902	0.0270	-0.0188	0.0867	Clustered

Table 4.8 shows that there is an evidence of spatial autocorrelation of diabetes prevalence across Africa and in some regions. This implies that the diabetes prevalence pattern is consistent across the Africa continent. Furthermore, there is a significant spatial clustering pattern in diabetes prevalence on the African continent with *p*-value = 0.000902 and *Z*-score = 0.0867. We found significant spatial clustering and dispersion at region-level diabetes prevalence on the African continent.

A further analysis of Moran indices to strengthen the earlier result was carried out. Using the spatial weights, spatial distance was used under randomisation. In this case, some countries that do not share spatial polygons were taken out, and the number of observations were reduced for the African continent and some regions. 49 countries were observed instead of 54, taking out Seychelles, Comoros, Madagascar, Cape Verde, and Sao Tome Principe because of their locations, since they were considered Medium Island countries. Table 4.9 shows the summary table of Moran Statistics under randomisation.

Table 4.9: Summary table of Moran statistics under randomisation

Regions	No observations	Moran's I statistics	Expected	Variance	SD	P-value	Spatial auto Correlation pattern
Overall	49	0.035	0.086	0.011	1.604	0.0342	Positive spatial autocorrelation
Southern Africa	9	0.027	-0.111	0.005	1.920	0.027	Positive Spatial autocorrelation
West Africa	15	-0.139	-0.071	0.019	-0.494	0.621**	Negative spatial autocorrelation
North Africa	7	0.076	-0.167	0.033	1.337	0.181**	Positive spatial correlation
Central Africa	7	-0.309	-0.167	0.049	-0.646	0.518**	Negative spatial correlation
East Africa	9	-0.009	-0.125	0.024	0.755	0.450**	Negative spatial correlation

**Not significant

In the case of Moran’s I, a large positive value is an evidence of positive spatial correlation, while a large negative value is evidence of negative spatial correlation. A positive spatial autocorrelation occurs when Moran's I is close to +1 and a negative spatial autocorrelation is said to occur when Moran I is near -1. We found that the Southern region on the continent has positive spatial autocorrelation with Moran's I of 0.027. This implies that the countries at this regions level are clustered together. The Z score of 0.027 for the Southern region indicates that there is a less than 1% likelihood that the clustered pattern could be the result of random choice. The results show no significant pattern for other regions. This implies that dissimilar countries with patterns of diabetes are next to each other and there is little influence.

To compare the results of the original Moran’s I, Monte Carlo simulation was done to account for non-normality of the data. Table 4.10 presents the summary statistics of the moran.mc test.

Table 4.10: **Summary table of Monte Carlo (MCMC) across the region**

	Statistics	**Observed rank**	***p*-value**
Central Africa	-0.309	320	0.68**
East Africa	-0.009	780	0.22**
North Africa	0.076	897	0.103**
West Africa	-0.139	289	0.711**
Southern Africa	0.027	800	0.20**

**Not significant

Moran statistics of Tables 4.9 and 4.10 show a slightly dissimilar result in the p-value, where the MCMC simulation shows a non-significant result in all the regions as against the summary result under randomisation. However, there is a similarity in the value of the Moran I in table 4.9 (under randomisation) and the p-values in Table 4.10.

4.3.2. Methodological steps in model building

A strong negative correlation was observed among the variables, note that correlation does not imply causation, but indicates whether a mutual relationship, causal or not, exist between variables. If the relationship is non-linear, common approach is linear modelling is to transform both the response and the predictor variables to coerce the relationship to one that is more linear. Common transformations include natural and base ten logarithmic, square root and inverse transformations. Therefore, the relationship was strengthened using a square root transformation.

The first step in the model building process is to map the dependent variable and explore spatial heterogeneity. If the dependent variable is not clustered, there is no need to build a spatially explicit model. Without clustering, the global model will be similar to the local model (Fotheringham et., al. 2002). Having seen that diabetes is clustered across the continent from the Moran I statistics in Table 4.8 and 4.9.

Initial data exploration and model specification using OLS was completed using R 3.2 software. Three factors motivated the decision to first specify the OLS model: 1) we wished to identify variables significantly correlated with the dependent variable before specifying the regression model; 2) the GWR software used for spatial analysis does not provide a variance inflation factor (VIF) to assess multicollinearity; and 3) the GWR software does not enable the researcher to extract regression residuals to assess spatial autocorrelation for the global model.

In the OLS regression we included only variables significantly correlated with the dependent variable, diabetes prevalence. Residuals from the global OLS model were mapped and analysed for spatial autocorrelation using Moran's *I*. The same set of variables was then used to specify a GWR model using the GWR4 software (http://geodacenter.asu.edu/gwr). While conducting GWR, we used the adaptive kernel, which was produced using the bi-square weighting function. The adaptive kernel uses varying spatial areas but a fixed number of observations for each estimation, a method most appropriate when the distribution of observations varies across space. Finally, a process that minimizes the Akaike Information Criteria (AIC) was used to determine the best kernel size.

The residuals of GWR models are assumed to be normally distributed; a further assumption is that they are not spatially autocorrelated or clustered across space. Such clustering suggests that the local model underestimates or overestimates diabetes prevalence in particular areas.

Table 4.11: Results from Ordinary Least Square Model of Diabetes prevalence in Africa (2015)

Variables	Parameter estimate	Standard error	t-value	Pr > |t|
Intercept	-0.581	2.084	-0.279	0.782
Obesity	-0.189	0.143	-1.326	0.192
Health expenditure	-0.011	0.147	-0.077	0.939
GDP	0.012	0.007	1.73	0.091
Population age	0.164	0.332	0.496	0.623
Urban Population	-0.093	0.059	-1.564	0.125
Physician density	0.920	0.362	2.542	0.015
MYSC	0.463	0.265	1.75	0.088
Age dependency ratio	0.987	0.349	2.830	0.007
Alcohol consumption	-0.142	0.091	-1.551	0.129
Physical activity	-0.060	0.035	-1.705	0.096
HDI	-1.498	1.214	-1.234	0.242
GNI	-0.0004	0.0066	-0.064	0.949

```
Significant codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4.11 shows that there is statistical significant association of diabetes with country GDP, Age dependency ratio, Physical inactivity and physician density.

Result of OLS also shows that residual standard error is 0.4594, multiple R-squared is 0.725, and the adjusted R-square of 0.6445. The OLR model was significant ($F_{12,41}$ = 9.007, P < .001). The model explained 64.4% of the variance in the continent diabetes prevalence.

As a result of the limitation of OLS, the global GWR model in Table 4.12 produced coefficient of estimates for each variable, the minimum and maximum coefficient, standard error and mean, the t-value and the confidence intervals. The change in both magnitude direction of the coefficients suggests spatial non-stationarity of the relationship between the predictors and diabetes prevalence. The R-squared is 0.3283 with the adjusted R-squared of 0.110009. The model was significant ($F_{12,41}$ = 4.741, P < .001). The model explained 11.0% of the variance in the continent diabetes prevalence.

However, there is a need to account for the local variability which is the strength of geographic weighted regression in the case of a spatial clustering in the continent. Therefore, bandwidth and geographic ranges of x coordinates (minimum -25.17, maximum 200.00 and range 227.17) and a Y coordinates (minimum (-29.58, maximum 34.11 and range 63.69) was selected during analysis. The model summary of the local GWR gives R-square of 1.0000 and adjusted R-squared of 1.0000, the local GWR result is shown in Table 4.13.

Table 4.12: Results of Global Geographical Weighted Regression model of diabetes prevalence in Africa 2015

Variables	Parameter estimate	Standard error	t-value	Lower Quartile	Median	Upper Quartile
Intercept	5.511964	0.646	8.529156	5.511964	5.511964	5.511964
Obesity	-0.5559	1.173	-0.474	-0.555929	-0.555929	-0.555929
Health expenditure	-0.059	0.927	0.064	0.058908	0.058908	0.058908
GDP	-0.529	2.115	-0.250	-0.529274	-0.529274	-0.529274
Population age	-0.829	2.373	-0.349	-0.828562	-0.828562	-0.828562
Urban Population	-0.660	0.787	-0.838	-0.659873	-0.659873	-0.659873
Physician density	1.909	2.481	0.769	1.908624	1.908624	1.908624
MYSC	-0.506	1.493	-.339	-0.505818	-0.505818	-0.505818
Age dependency ratio	-0.718	3.421	-0.210	-0.717885	-0.717885	-0.717885
Alcohol consumption	0.757	0.875	0.864	0.756968	0.756968	0.756968
Physical activity	-0.693	0.789	-0.878	-0.692921	-0.692921	-0.692921
HDI	-1.244	1.493	-0.339	-1.244072	-1.244072	-1.244072
GNI	1.888	2.387	0.791	1.887590	1.887590	1.887590

The adjusted R^2 for the local GWR model is 1.00; the adjusted R^2 in the OLS model was 0.64. Explicitly, the global OLS R^2 of 0.64 masks a wide distribution of local associations between the predictors and diabetes prevalence. The Confidence interval (CI) for all the variables shows that they are significant.

Without GWR, we would have been unable to estimate local models. The local GWR account for 100% of variation a spatial variation that would have been missed with the OLS model alone. Residuals for the GWR model, although significant, were less spatially autocorrelated than residuals for the OLS model (Moran's *I* = 0.01; *z* = 3.74; *P* < .001). Compared with OLS, the local GWR model greatly improved model fit. We looked at the corrected Akaike Information Criterion (AICc) which is a measure of model performance and help in comparing the OLS and GWR models, model with lower AICc is said to provide a better fit. From the results output OLS model AICc is 344.718595 while the GWR model AICc is 341.508827 with a model improvement of 3.209768.

Comparing the GWR AICc value to the OLS AICc value is one way to assess the benefits of moving from global model to a local regression model GWR. The local GWR model explained more variance in diabetes prevalence and reduced the AICc ($\Delta R^2 = 0.36$; ΔAICc = 2.45). Adaptive kernel bandwidth was used, which accounted for differences in the size of countries and therefore the distance of influence.

In conclusion, the AICc is usually preferred as a means of comparing models in GWR because the effective number of degree of freedom is a function of the bandwidth so the adjustment is usually marked in comparison to a global model like OLS.

Table 4.13 Results of the local Geographic Weighted Regression model of diabetes in Africa 2015

Variables	Mean	Standard deviation	Lower Quartile	Median	Upper Quartile
Intercept	5.409891	0.101124	5.511964	5.511964	5.511964
Obesity	-0.545634	0.010199	-0.659873	-0.659873	-0.659873
Urban population growth	-0.647853	0.012106	-0.659873	-0.659873	-0.659873
Age dependency ratio	-0.704591	0.013170	-0.717885	-0.717885	-0.717885
Physician density	1.873279	0.035016	1.908624	1.908624	1.908624
HDI	-1.221034	0.022824	-1.244072	-1.244072	-1.244072
GNI	1.852635	0.034630	1.887590	1.887590	1.887590

The final model in Table 4.13, is the output of the improved model and this indicate a perfect fit, while other variables not in the table became a fixed term and then obesity, urban population growth, age dependency ratio, physician density, HDI, and GNI became the final parameters that remain in the model, the output result gave the mean, standard deviation and the confidence interval. Obesity, urban population growth, age dependency ratio, physician density, HDI and GNI were each significantly associated with diabetes prevalence, relationships that were spatially nonstationary across Africa. The variation in parameter estimates from GWR suggests the need to apply this spatial analysis tool to other diabetes studies that have been restricted to global models. A result like this is expected when regressing a strongly trended series like diabetes.

In conclusion the OLS with adjusted R square of 0.11 (11.0% variability) shows a lesser spatial variability and shows that all variables are significant , whereas the global GWR shows an adjusted R square of 0.64 (64% variability) more variability with all variables significant, local GWR improved the model fit accounting for 100% variability with six significant variables (obesity, urban population growth, Age dependency ratio, physician density, HDI and GNI). This shows that there is a relationship between these variables and diabetes in Africa , with consideration on the non-stationarity of the variables.

4.4. Local Indicator of Spatial Association (LISA)

The Lisa maps in Figure 4.23 to 4.27 display the spatial distribution of diabetes prevalence in 2015 for Africa and its respective regions, also known as the hot spots.

The LISA analysis yielded five categories of spatial units. These categories were defined as “high-high” (HH), “low-low” (LL), “high-low” (HL), “low-high” (LH), and “not significant” (NS). The high-high and low-low locations (positive local spatial autocorrelation) are referred to as spatial clusters, while the high-low and low-high locations (negative local spatial autocorrelation) are referred to as spatial outliers.

Similarly, HH clustering denotes countries with high diabetes prevalence and are associated with neighbouring countries with high diabetes prevalence rates.

- LH clustering denotes countries with low diabetes prevalence which are associated with neighbouring countries with high diabetes prevalence rates.
- LL clustering denotes countries with low diabetes prevalence which are associated with neighbouring countries with low diabetes prevalence rates.
- HL clustering denotes countries with high diabetes prevalence which are associated with neighbouring countries with low diabetes prevalence rates.

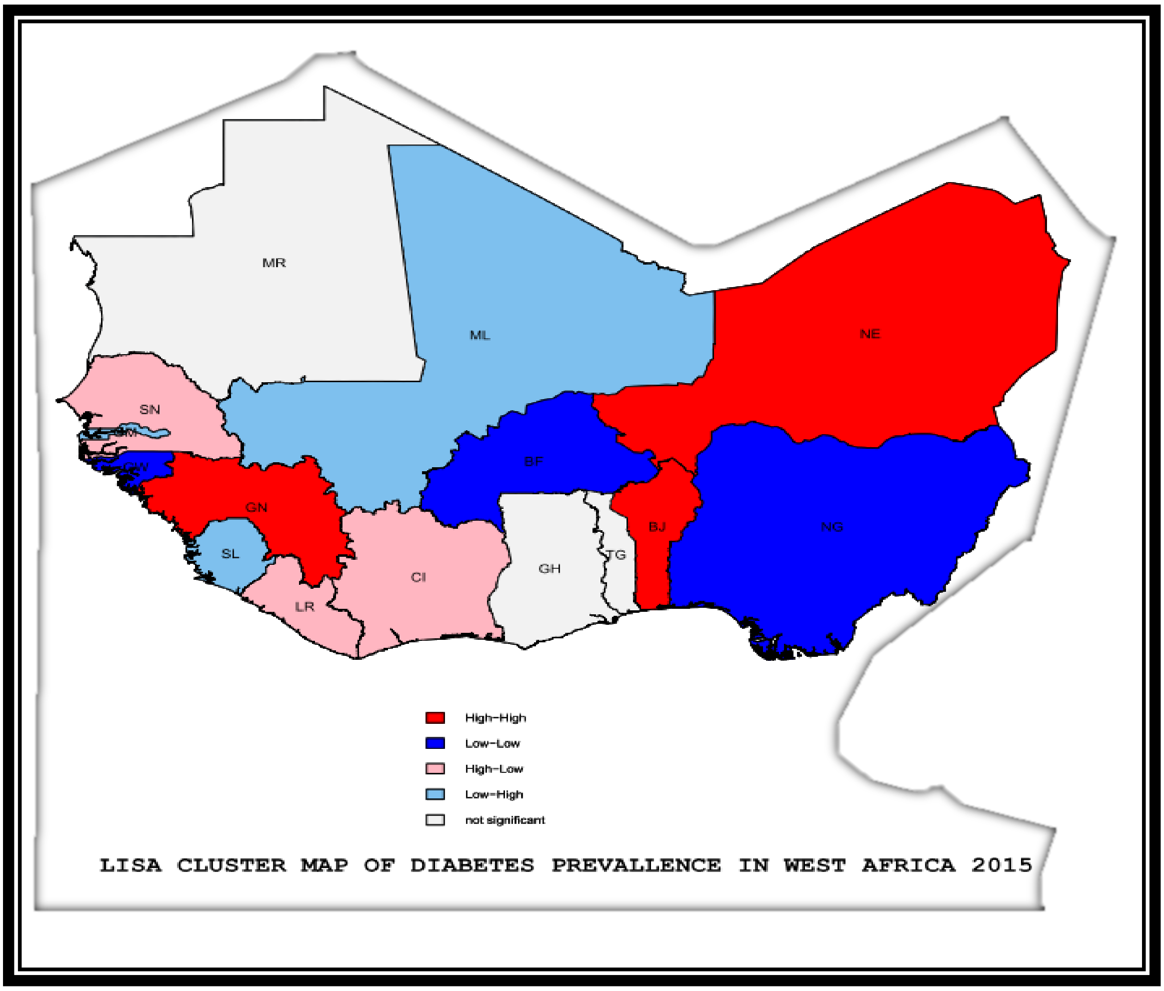


Figure 4.24. LISA Cluster map of diabetes prevalence in West Africa region

From the LISA map in Figure 4.24, within the West Africa region, countries such as Senegal, Cote D'Ivoire and Liberia (in the HL category) as well as Mali and Serial-Leone (in the LH category) are spatial outliers with negative spatial autocorrelation. Therefore, diabetes prevalence was significantly dispersed in the region. The region has the highest number of outliers.

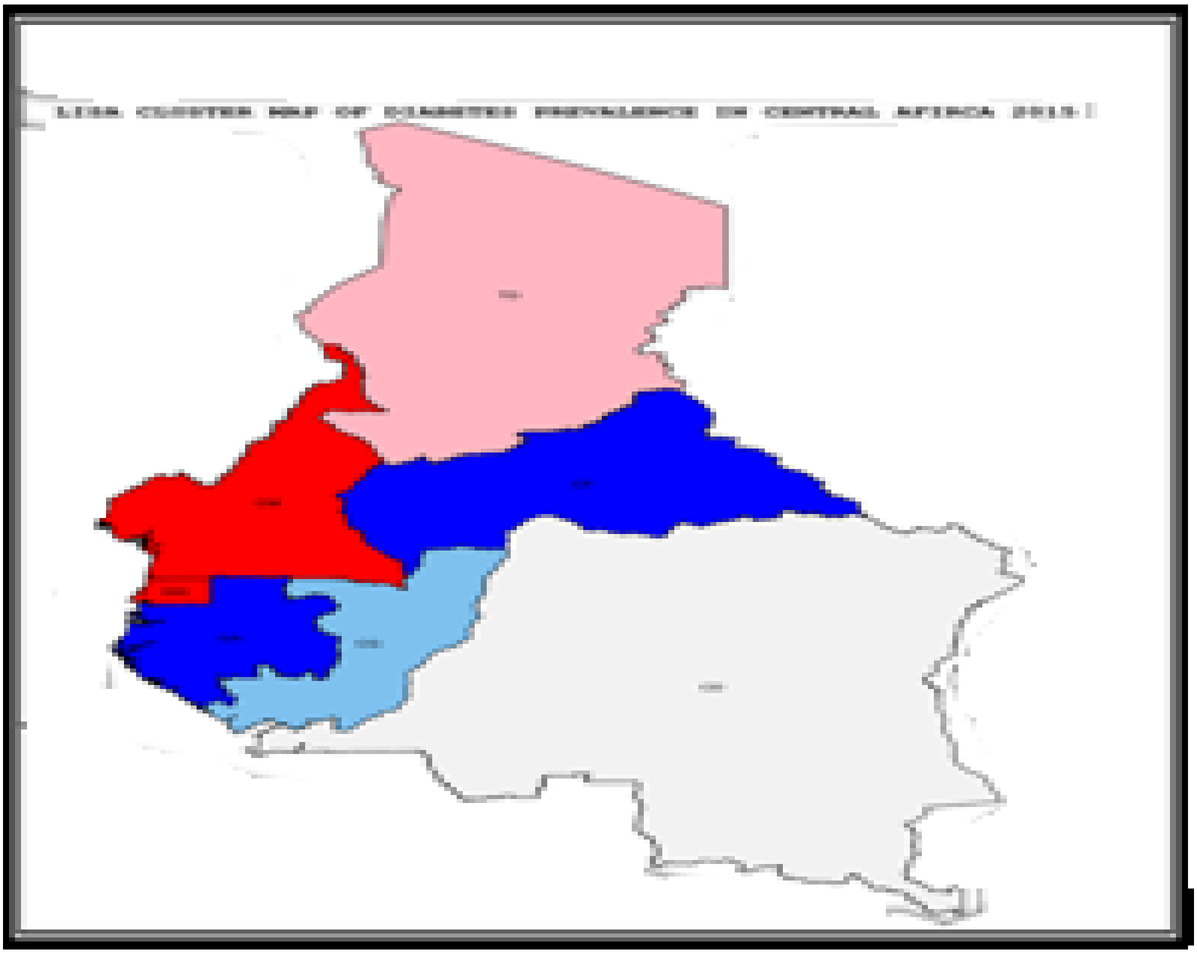

LISA CLUSTER MAP OF DIABETES PREVALENCE IN CENTRAL AFRICA 2015

Figure 4.25. LISA cluster map of diabetes in Central Africa region

The LISA map in Figure 4.25 shows that from the Central Africa region, countries such as Chad (in the HL category) as well as Congo DRC (in the LH category) are spatial outliers with negative spatial autocorrelation. Therefore, diabetes prevalence was significantly dispersed in the region.

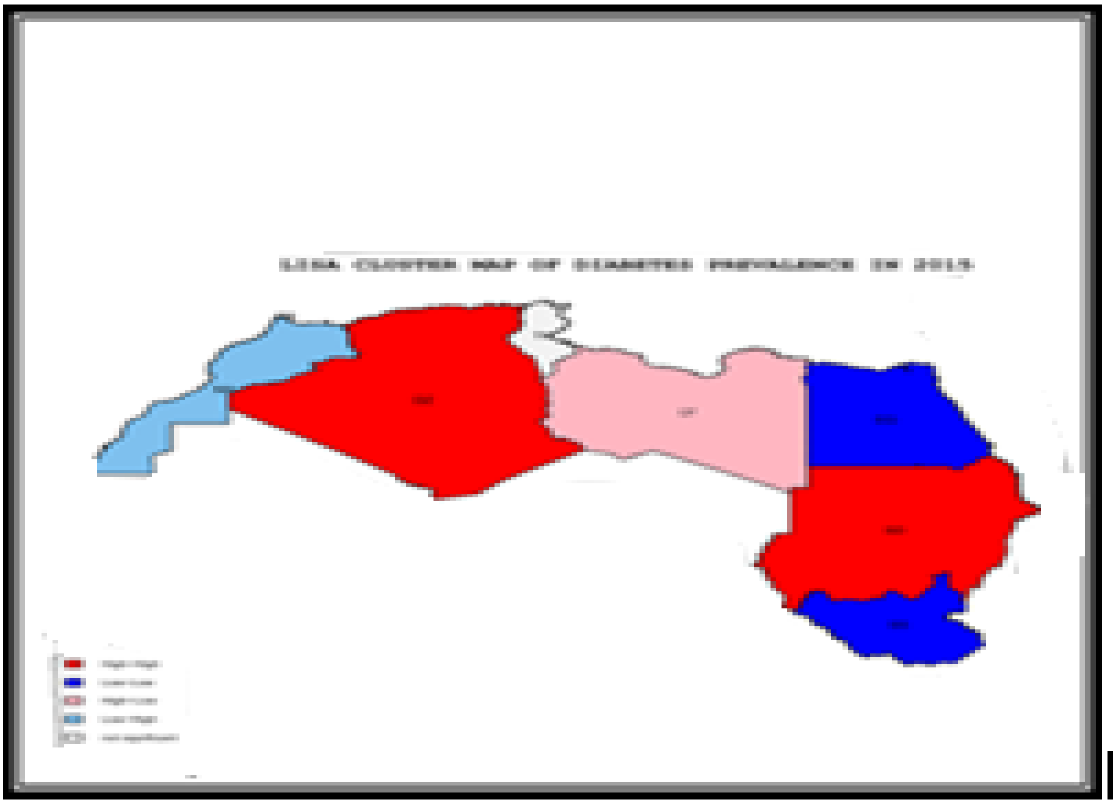


Figure 4.26. LISA cluster map of diabetes in North Africa region

The LISA map in Figure 4.26 shows that from the North Africa region, Libya in the HL category, as well as Morocco in LH category are spatial outliers with negative spatial autocorrelation. Therefore, for diabetes in the North Region, two HH areas and two LL spots were detected, the largest HH areas being in Algeria and Sudan, while the two smaller hotspots were found in Egypt and South Sudan. In addition, diabetes prevalence was significantly dispersed in the region.

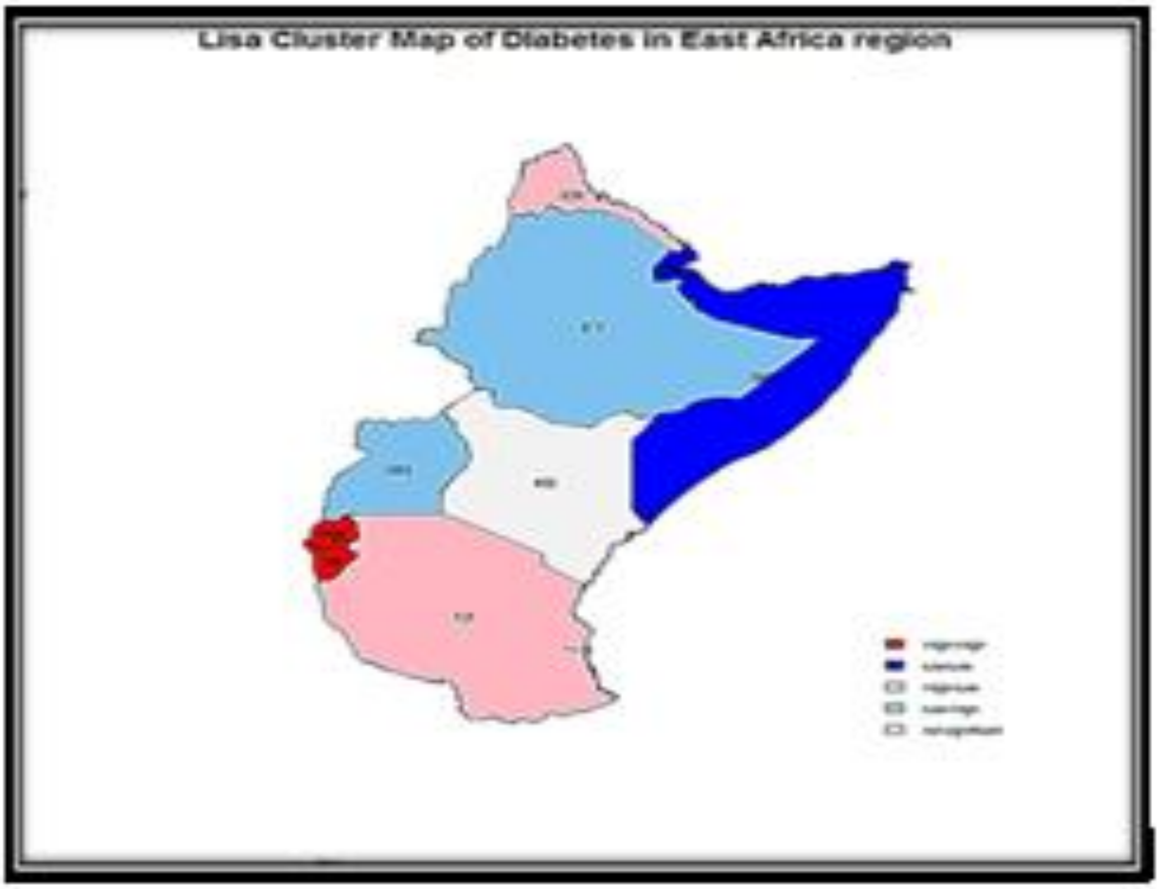


Figure 4.27. LISA cluster map of diabetes in East Africa region

The LISA map in Figure 4.27 shows that from the East Africa region, Eritrea and Tanzania (in the HL category), as well as Ethiopia and Uganda, are spatial outliers with negative spatial autocorrelation. The region shows a non-significant association with diabetes. Diabetes prevalence was significantly clustered in the region.

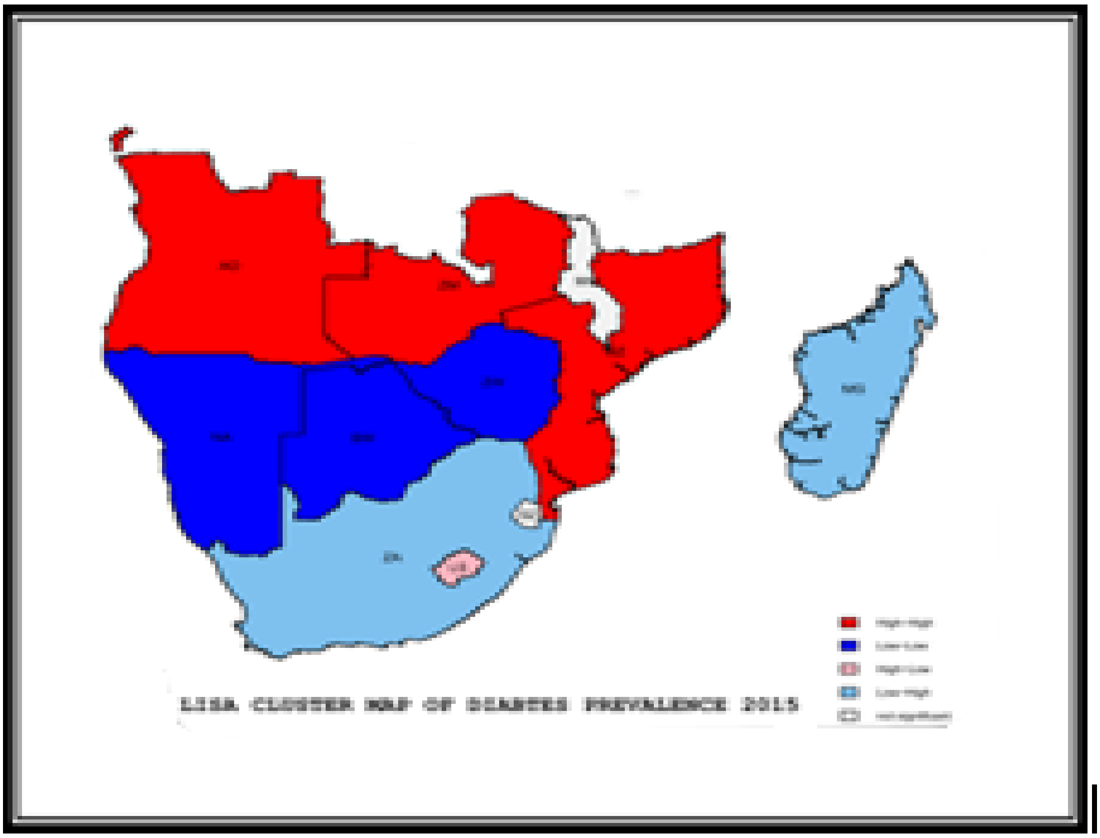


Figure 4.28. LISA cluster map of diabetes in the Southern Africa region

The LISA map in Figure 4.28 shows that from the Southern Africa region, countries such as South Africa and Madagascar (in the HL category) and Lesotho (LH category) are spatial outliers with positive spatial autocorrelation. The region shows a significant association with diabetes. Diabetes prevalence was significantly clustered in the region. Other countries in this region are Angola, Mozambique, and Zambia in High-High categories with Botswana, Namibia and Zimbabwe in the with Low-Low categories.

The spatial clusters shown on the LISA cluster map only referred to the core of the cluster. The cluster is classified as such when the value at a location (either high or low) is more like its neighbours (as summarized by the weighted average of the neighbouring values and its spatial lag) than would be the case under spatial randomness.

In the North Region, two HH areas and two LL spots were detected, the largest HH areas being in Algeria and Sudan, while two smaller hotspots were found in Egypt and South Sudan.

4.4.1. Conclusion

The results of the LISA cluster analysis and scatter plots for the five regions show that the following countries have high diabetes prevalence alongside neighbouring countries also with high diabetes prevalence rates (they are denoted with red and in the high-high category (HH)): Benin, Guinea, Guinea Bissau, Cameroon, Equatorial Guinea, Rwanda, Angola, Mozambique, Zambia, Sudan, Algeria, Burundi.

Likewise, the countries which have high diabetes prevalence and have neighbouring countries with low diabetes prevalence rates (HL) are Cote D'Ivoire, Liberia, Senegal, Chad, Eritrea, Tanzania, Lesotho, Libya, and Swaziland. They are denoted with pink. The countries with low diabetes prevalence which have neighbouring countries also with low diabetes prevalence rates (in the low-low category (LL)) are denoted in royal blue and are Burkina Faso, Nigeria, Guinea Bissau, Central Africa Republic, Gabon, Somalia, Djibouti, Namibia, Botswana, and Zimbabwe.

Similarly, countries with low diabetes prevalence, but which have neighbouring countries with high diabetes prevalence rates (in the LH category) are denoted in light blue, and are South Africa, Mali, Sierra Leone, Congo republic, Congo DRC, Ethiopia, Uganda, and Madagascar.

From the overall analysis, we saw the importance of LISA in identifying the countries from the map which have contributed strongly to the overall trend of the positive autocorrelation in the GMI and Moran I. Countries with high diabetes surrounded by low diabetes countries are: from Southern Africa region are: South Africa and Madagascar, from East Africa region are: Eritrea and Tanzania, from North Africa region: Libya, from Central Africa region: Chad, from West Africa region are: Senegal, Cote d'Ivoire and Liberia. These countries need urgent action**.** Likewise, countries with low diabetes surrounded by high diabetes countries are: from Southern Africa region: Lesotho, from East Africa region: Ethiopia and Uganda, from North Africa region: Morocco, from central Africa region: Congo DRC and from West Africa region: Mali. These countries can be used as an example for their neighbours.

In conclusion the LISA map of the African continent has helped to identified (16) sixteen outlier countries such as South Africa, Mali, Serial-Leon, Congo DRC, Ethiopia, Uganda, Madagascar, Cote ‘Ivoire, Liberia, Senegal, Chad, Eritrea, Tanzania, Lesotho, Libya, Morocco. These countries are contributing strongly to the global diabetes prevalence in Africa based on the pattern of the prevalence observed from the analysis.

4.5. Results of Classical Statistics

4.5.1. Simple Linear Regression Result

The result of the linear regression for the data is indicated in Table 4.14 below.

Table 4.14: Parameter estimate and overall fit for Simple linear regression (All variables)

Variables	DF	Parameter estimate	Standard error	t-value	Pr > |t|	95% confidence limits	
						Low	High
Intercept	1	-3.962	4.701	-0.84	0.404	-13.455	5.531
Obesity	1	-0.20039	0.09611	-2.09	0.043*	-0.394	-0.006
Health expenditure	1	-0.049	0.141	-0.35	0.730	-0.334	0.236
GDP	1	0.0004	0.0003	1.73	0.091	-0.00007	0.0009
Population age	1	0.101	0.101	0.99	0.326	-0.104	0.306
Urban Population	1	-0.030	0.021	-1.44	0.1551	-0.073	0.012
Physician density	1	4.622	1.121	4.12	0.0002*	2.358	6.885
MYSC	1	0.463	0.265	1.75	0.088	-0.072	0.999
Age dependency ratio	1	1.061	0.308	3.44	0.001*	0.438	1.684
Alcohol consumption	1	-0.107	0.135	-0.79	0.435	-0.380	0.166
Physical activity	1	-0.036	0.028	-1.31	0.198	-0.092	0.020
HDI	1	-5.574	4.693	-1.19	0.242	-15.052	3.903
GNI	1	-0.00003	0.0002	-0.17	0.869	-0.0005	0.0004

```
Residuals:
   Min        1Q       Median     3Q      Max
-2.9483    -1.3793    -0.1717   0.6411   5.6512
```

Residual standard error: 2.085 on 41 degrees of freedom
Multiple R-squared: 0.8044, Adjusted R-squared: 0.7472
F-statistic: 14.05 on 12 and 41 DF, p-value: 6.266e-11 < 0.0001
SS Model =732.73 SSE = 178.16 SST = 910.89 Root Mean = 2.08
Coefficient variation = 38.86 Dependent mean = 5.36

* significant variables

The result shows that the sum of squares (SS) associated with three sources of variance are the sum of square model (SS Model), sum of square of residual (SSResidual), also known as sum of square error, and sum of square total (SSTotal). The SSModel is given by $SSTModel = \Sigma\left(\hat{Y} - \bar{Y}\right)^2 = 732.73$, SSError by $SSError = \Sigma(Y - \hat{Y})^2 = 178.16,$ and SSTotal (the total variability around the mean) by $\Sigma(Y - \bar{Y})^2 = 910.89.$ One of the important results is the F statistics, used to answer the question "Do the independent variables reliably predict the dependent variable?" With the value of $F_{(12,41)} = MSModel/MSResidual = 14.05$, and a p-value of 0.0001 (less than $\alpha = 0.05$), it can be inferred that the independent variables included in the model reliably predict the dependent variable. Therefore, this group of variables can be used to reliably predict diabetes prevalence.

Also, in checking the overall model fit, it is important to compute the R-square. The R-square is the proportion of variance in the dependent variable which can be predicted from the independent variables. Our findings show that R-square = 0.8044, which indicates that 80.4% of the variance in diabetes prevalence can be predicted from all the independent variables. However, this is an overall measure of strength of association, and does not reflect the extent to which any independent variable is associated with the dependent variable. To determine the extent to which each of the independent variables is associated with the dependent variable, we look at the coefficients of the parameter estimates table. This means that for every increase in obesity prevalence by one percent, we expect diabetes prevalence to decrease by -0.20039 on average, holding all other variables constant. Similarly, for every increase in health expenditure by one percent, we expect diabetes prevalence to decrease by -0.049 on average, holding all other variables constant. Also, for every increase in population age by one percent, we expect diabetes prevalence to increase by 0.101 on average, holding all other variables constant. Similarly, for every increase in age dependency ratio by one percent, we expect diabetes prevalence to increase by 1.061 on average, holding all other variables constant. We expect diabetes prevalence to be higher by 4.622 for every increase in physician density per 1000 person and by 0.463 for every increase in Mean year of school age.

Findings from this analysis show that obesity (p-value = 0.04), age dependency ratio (p-value = 0.0013), physician density (p-value = 0.0002), and GDP are the significant predictors of diabetes.

4.5.1.1. Diagnostic Plots Results for Linear Regression

There is a need to check if the model works well for the intended data. After fitting the model, residuals could reveal some unexplained pattern in the data by the fitted model.

Residuals could show how poorly a model represents data. With this knowledge, it is not easy to check if linear assumptions are met. This is then further revealed by performing a diagnostic plot. A diagnostic plot is a plot used in checking for heteroscedastic, normality, and influential observations in the model data.

In R, there is a built-in diagnostic plot for linear regression which by default gives a four-diagnostic plot which shows residuals in four separate ways. The “plot” command in R gives the result in Figure 4.29. The plots are individually explained in the following subsections.

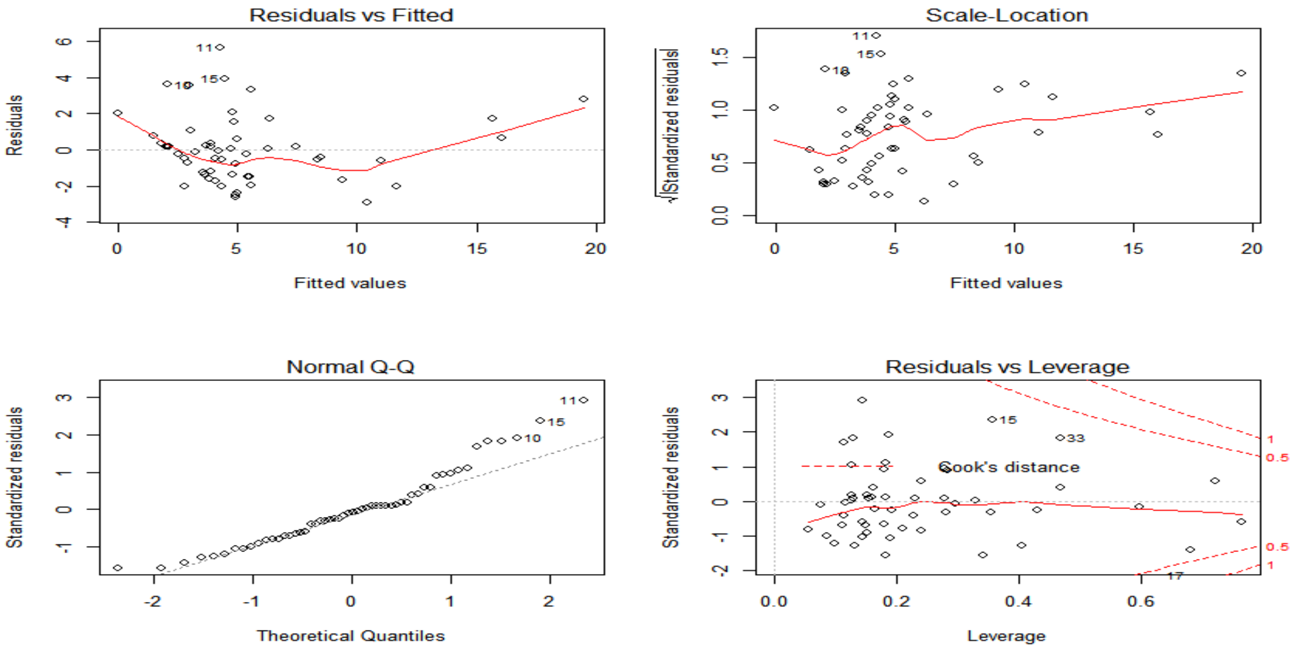


Figure 4.29. Diagnostic plot of linear regression

4.5.1.1.1. Residual vs Fitted

This plot shows if residuals have a non-linear pattern, that is, verifying the assumptions of a linear model. It is possible to have a non-linear relationship between predictor variables and an outcome variable, which could show up in the plot if the model does not capture the non-linear relationship. Note that an equally-spread residual around the horizontal line without a distinct pattern is an indication of a non-linear relationship.

The plot indicates dependency between the residual and fitted values. Thus, the plot suggests that there is heteroscedastic linear model.

4.5.1.1.2. Normal Q-Q Plot (Quantile-Quantile Plot)

The Q-Q plot shows if residuals are normally distributed. It answers the question whether residuals close a straight line well or deviate severely. A good plot will show residuals lined well on the straight dashed line.

Similarly, often, we are concerned about what is happening at the tails of our distribution. To get a better resolution of this, a Q-Q plot can be computed. A Q-Q plot is useful in many ways. It is used to compare distribution shapes (skewness, location). Also, it can be used to compare two samples of data as a non-parametric approach. It is also used to compare a dataset to a theoretical model, assessing goodness of fit.

The above Q-Q plot in Figure 4.29 compares randomly generated, independent standardized residual data on the vertical axis to a standard normal population on the horizontal axis. The plot suggests that the data are right-skewed with observations 11, 10, and 15 as outliers.

4.5.1.1.3. Scale-Location plot

This is also called spread-location plot. This plot shows if there is an equal spread of residuals along the ranges of predictors. It is used in assessing the assumption of equal variance (homoscedasticity). A horizontal line with equally (randomly) spread points shows a good model.

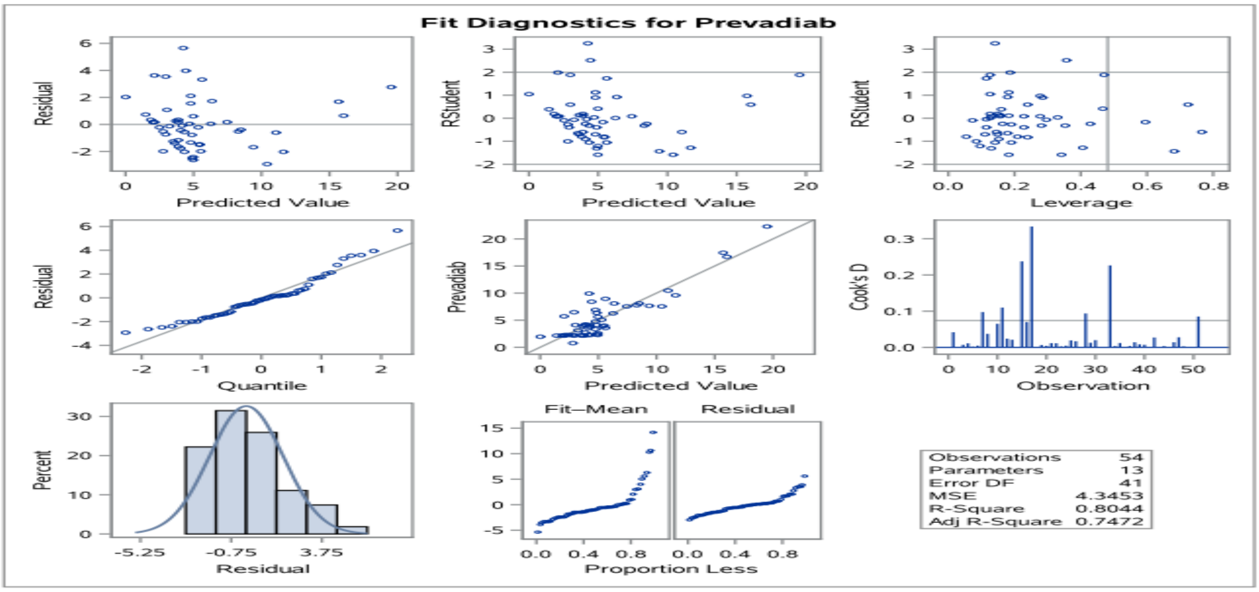
From the plot above, the residual is not randomly spread, the residual begins to spread wider along the x axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is no longer horizontal and shows a steep angle and curve. This suggests that there is homoscedasticity and over dispersion of the location on the study area.

4.5.1.1.4. Residual vs Leverage

This plot helps to determine influential cases (i.e. subjects), if any. In linear regression, not all outliers are influential. Data may have extreme values and might thus not be influential to determine a regression line. This means that either the outliers are included or excluded from the analysis, the results wouldn't be much different. On the other hand, extreme cases could be very influential if within a reasonable range of the values. However, if these outliers are excluded from the analysis, there could be another extreme case against the regression line which can alter the result. Unlike the other plots, here, patterns are not relevant. We look out for outlaying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line.

We look for cases that are outside of a dashed line, known as Cook's distance. Cases that are outside the Cook's distance have high Cook's distance scores and are influential to the regression results. However, the regression results will be altered if those cases are excluded.

From our results, one case is far beyond the Cook's distance lines. Also, the plot identifies outliers at case #1. Some countries were removed, bringing the variables to eight from twelve.



Observations	54
Parameters	13
Error DF	41
MSE	4.3453
R-Square	0.8044
Adj R-Square	0.7472

Figure 4.30. Diagnostic plot of linear regression

4.5.2. Result of Linear Regression (6 variables)

Six variables were removed because of collinearity among the variables. Collinearity test was estimated using the Spearman Rho correlation coefficient test of the dependent variable and the initial twelve independent variables, as shown in Table 4.3. The result of the fitted linear regression model with outlier variables removed is indicated in the Table 4.15 and Figure 4.31 below. This is denoted as model 2.

Table 4.15: Parameter estimate and overall fit for linear regression

Variables	DF	Parameter estimate	Standard error	t-value	Pr > |t|	95% confidence limits	
						low	High
Intercept	1	-0.113	4.691	-2.417	0.019	-20.267	-2.915
Health expenditure	1	-0.239	0.166	-1.447	0.155	-0.459	0.165
GDP	1	0.0002	0.0001	1.904	0.063	-0.459	0.165
Population age	1	0.325	0.0995	3.266	0.002 *	0.152	0.520
Urban Population	1	-0.056	0.024	-2.313	0.025*	-0.082	0.011
Physician density	1	2.947	0.832	3.545	0.0009*	3.384	8.159
MYSC	1	0.115	0.198	0.580	0.564	-0.162	0.578

Residual standard error: 2.609 on 47 degrees of freedom
Multiple R-squared: 0.6488, Adjusted R-squared: 0.6039
F-statistic: 14.47 on 6 and 47 DF, *p*-value: 0.000000002.878

SSModel=732.73 SSE = 178.16 SST = 910.89 Root Mean = 2.08
Coefficient variation = 44.69 Dependent mean = 5.36

Residuals:

```
  Min      1Q       Median      3Q         Max
-5.2540   -1.5957     -0.3406     1.6479       7.3674
```

*significant variables

The result shows that the F statistics = 14.47 with a *p*-value < 0.001, this indicates a good fit of the overall model. A residual standard error of 2.609 can also be observed. The R^2= 0.6488, which indicates that 64.88% of the variance in diabetes prevalence can be predicted from all the independent variables. The adjusted R^2 = 0.6039. A coefficient of variation of 44.691 can be for the model also aims to describe the model fit in terms of the relative sizes of the square residuals and outcome value. This indicates that the overall measure of dispersion is high.

Findings from this analysis shows that population age, with *p*-value = 0.002, urban population growth (*p*-value = 0.025), and physician density (*p*-value = 0.0009) are the significant predictors of diabetes. Therefore, for every increase by one percent of population age, we expect diabetes prevalence to increase by 0.325 on average, holding all other variables constant.

The parameter of estimates shows that for every increase in population age by one percent, we expect diabetes prevalence to increase by 0.325 on average, holding all other variables constant. Also, for every increase in urban population by one percent, we expect diabetes prevalence to decrease by -0.056 on average, holding all other variables constant. Lastly, for every increase in physician density by one percent, we expect diabetes prevalence to increase by 2.947 on average holding all other variables constant.

From the diagnostic plots in Figure 4.31, the residual plot, and the q-q plot, it is can be seen that the data is normally distributed.

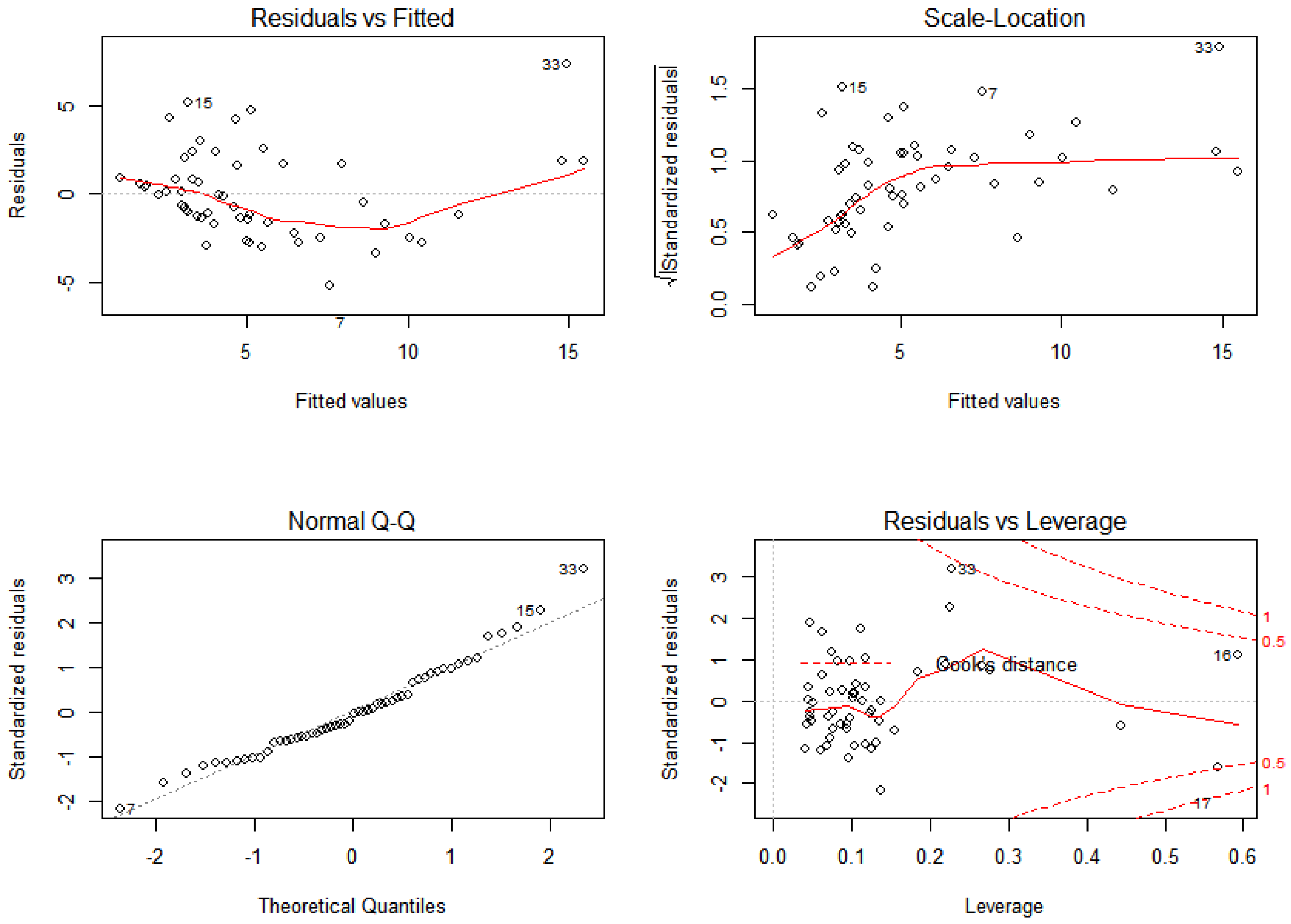


Figure 4.31. Diagnostic plots of the linear regression

4.6. Model Transformation

We estimate the model by transforming the data using the log form of the dependent variable. Generally, there are two reasons for taking the log of a variable in regression, one statistical, and the other substantive.

Statistically, OLS regression assumes that the errors, as estimated by the residuals, are normally distributed. When they are positively skewed (long right tail), taking logs can sometimes help. Sometimes logs are taken of the dependent variables, sometimes of one or all the independent variables, sometimes both dependent variable and independent variables.

This section compares the results of the different transformations implemented, which are:

- Log taken on dependent variable only
- Log taken on dependent variable and all the independent variables
- Fractional probit regression.

4.6.1. Results of Log Linear Regression Model

Table 4.16: Log -linear Regression Dependent variable transformed (Model 2)

Coefficients	Estimates	Standard error	t-value	Pr (> |t|)
Intercept	-0.797	0.877	-0.948	0.368
Health expenditure	-0.033	0.031	-1.081	0.285
GDP	0.00004	0.00002	1.568	0.124
Population age	0.041	0.019	2.219	0.031*
Urbanization	-0.006	0.005	-1.416	0.164
Physician density	0.368	0.155	2.369	0.022*
MYSC	0.036	0.037	0.970	0.337

```
Residuals:
    Min        1Q      Median        3Q        Max
-1.43218      -0.27361   -0.07971    0.22681     0.96483
Residual standard error: 0.4876 on 47 degrees of freedom
Multiple R-squared:  0.5143,    Adjusted R-squared:  0.4523
F-statistic: 8.295 on 6 and 47 DF, p-value: 0.00000380
```

*significant variables

The result of log linear regression with the six variables (model 2) shows that urban population growth and physician density are significant with p-values of 0.031 and 0.022 respectively.

R^2 = 0.5143, which shows that 51.43% of the variance in the diabetes prevalence can be predicted from all the independent variables in the data set. The overall model fit given $F_{(12,41)} = 8.295 > p(value) = 0.0000$, leading us to conclude that the independent variables reliably predicts the dependent variable.

From the parameter of estimates in Table 4.16, it can be seen that for every increase in population age by one unit, we expect diabetes prevalence to increase by 0.041 on average, holding all other variables constant. Lastly, for every increase in physician density by one unit, we expect diabetes prevalence to increase by 0.368 on average, holding all other variables constant.

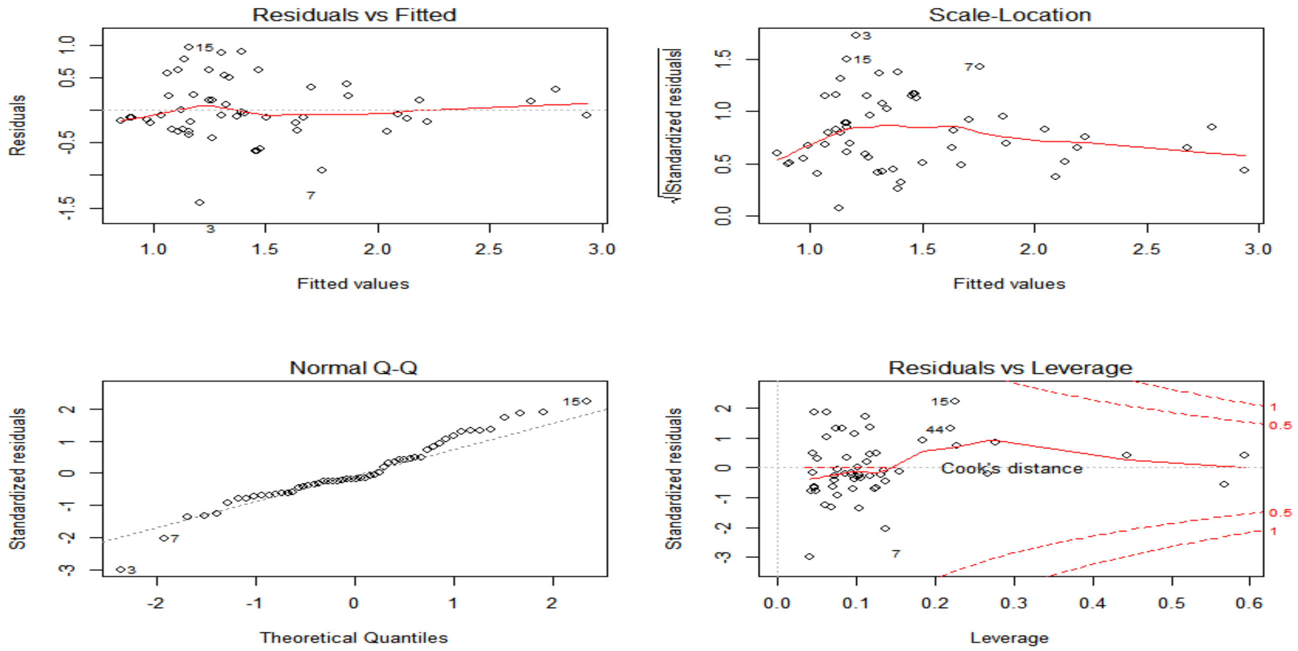


Figure 4.32. Diagnostic plots of the-Log linear regression

The diagnostic, residual, and q-q plots show that the data is normally distributed.

4.6.2. Result of Fractional Probit Model

Table 4.17: Parameter estimates for fractional Probit regression model

Variables	Parameter estimate	Standard error	z- value	Pr > |t|	95% confidence limits	
					Low	High
cons	-3.030	0.360	-8.41	0.000	-3.735	-2.324
Health expenditure	-0.22	0.015	-1.44	0.149	-0.510	0.008
GDP	0.00001	0.000006	2.26	0.024*	0.000001	0.00003
Population age	0.026	0.008	3.45	0.001*	0.011	0.414
Urban Population	-0.004	0.002	-1.61	0.108	-0.008	0.0008
Physician density	0.182	0.039	4.66	0.000*	0.012	0.259
MYSC	0.125	0.012	0.99	0.324	-0.012	0.037
Wald Chi (6) = 169.35 Prob > chi2 = 0.0000 Pseudo R2 = 0.0417 Pseudo likelihood = -10.819						

*significant variables

The fractional probit (also known as fractional response estimator) used the proportion of the diabetes prevalence which was first calculated as the percentage of diabetes prevalence divided by the total population for each country. The fractional response estimator fits models on continuous zero to one data using probit, logit, and heteroskedastic probit. The result of a fractional probit with the six variables (model 3) shows that GDP, population age, and physician density are significant with *p*-values of 0.024, 0.001 and 0.000 respectively.

The parameter of estimates shows that for every increase in GDP proportion, we expect diabetes prevalence to increase by 0.00001 on average, holding all other variables constant. Also, for every increase in the proportion of physician density, we expect diabetes prevalence to increase by 0.183 on average, holding all other variables constant.

4.6.3. Poisson Regression Results

In this section, instead of working with prevalence of diabetes, the dependent variable is transformed into count and a Poisson regression was utilised. The Poisson regression model was estimated in SAS using the PROC GENMOD and the result of the parameter estimate is shown in Table 4.18.

Table 4.18: Parameter estimates of Poisson regression

Parameter	DF	Estimate	Standard error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chi square
Intercept	1	15.2881	0.0020	15.2842	15.2921	5.712E7	< .0001
Health expenditure	1	-0.0789	0.0001	-0.0791	-0.0787	984347	<.0001
GDP	1	-0.0001	0.0000	-0.0001	-0.0001	2098808	< .0001
Population age	1	-0.0367	0.0000	-0.0369	-0.0367	739373	< . 0001
Urbanization	1	-0.0021	0.0000	-0.0021	-0.0002	39139.5	< .0001
Physician density	1	1.0227	0.0001	1.0924	1.0929	5,937E7	< .0001
MYSC	1	0.1975	0.0001	0.1973	0.1977	3606594	< .0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

AIC = 55551557.29 AICC = 55551560.80 BIC = 555515
Deviance = 1293835.13, Scaled Deviance = 1293839.13, Pearson Chi-Square = 1670118.01 with degree of freedom 43

*significant variables

Table 4.18 shows the result of the parameters estimates, using a maximum likelihood method. All the predictors were significant with *p*-value < 0.05. Therefore, we can conclude that for a one-unit increase in physician density, the difference in the logs of expected counts would be expected to increase by 1.0227 units while holding other variables in the model constant.

However, for a one-unit increase in urbanization, the difference in the logs of expected counts would be expected to decrease by 0.0021 units while holding other variables in the model constant.

4.6.4. Negative binomial model result

Table 4.19. Parameter estimates of Negative binomial model

Parameter	**DF**	**Estimate**	**Standard error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > Chi square**
Intercept	1	13.9574	2.774	9.8857	18.0290	45.14	< .0001*
Health expenditure	1	-0.1388	0.0934	-.0.3236	0.0459	2.17	0.1407
GDP	1	-0.0001	0.0001	-0.0003	0.0000	3.79	0.0517*
Population age	1	0.0052	0.0478	-0.0885	0.0989	0.01	0.9130
Urbanization	1	-0.0166	0.0176	-0.0511	0.0179	0.89	0.3463
Physician density	1	1.2683	0.5245	0.2403	2.2962	5.85	0.0156*
MYSC	1	0.1940	0.1128	-0.0270	0.4150	2.96	0.0854
Dispersion	0	1.3729	0.2345	0.9823	1.9189		
AIC = 1478.18 AICC = 1481.69 BIC = 1493.47 deviance = 1.3999, Pearson Chi-Square 1.4128 with 43 degrees of freedom							

*significant variables

The output in Table 4.19 begins with the model information and the criteria for assessing goodness of fit. The number of observations is 50, and the link function is log. Pearson Chi-square is 1.4128, the deviance statistics is 1.3999 with 42 degrees of freedom for both, and with the dispersion ratio as given. The deviance statistics shows that there is over-dispersion, and a non-significant value of Pearson statistics suggested that the model adequately fits the data. A dispersion ratio value close to one is an evidence of over-dispersion in the model. Therefore, the model shows convincing evidence of over-dispersion. The AIC, BIC, and AICC are 1478.18, 1493.47, and 1481.69, respectively. Finally, from the Table 4.16, the algorithm for parameter estimates has converged, implying that a solution was found.

GDP and physician density were seen to be the significant variables in this model with *p*-values of 0.0517 and 0.0156, respectively. The coefficient estimates of physician density =1.2683 shows an increase compared with Poisson regression results), with SE=0.5245, 95% CI of (0.2403, 2.2962), a wider interval as compared to Poisson regression, with *p*-value = 0.0693, and a Wald chi square statistic of 5.85. This means that for a one-unit increase in physician density, the expected log count of the diabetes prevalence will increase by 1.2683. A significant GDP of *p*-value = 0.0517 means that for each unit increase in GDP, the expected log count of the diabetes prevalence will decrease by 0.0001. This shows that with a negative binomial model, physician density per 1000 persons, GDP, and population age are significant and have a positive relationship with diabetes prevalence on the continent.

In addition, there is an estimate of the dispersion coefficient (often called alpha) = 1.3729. A Poisson model is one in which this alpha value is constrained to zero. In this study, the estimated alpha has a 95% CI that does not include zero, which suggested that the negative binomial is more appropriate than Poisson. An estimate greater than zero suggest over-dispersion (variance greater than mean), and an estimate less than zero suggest under-dispersion, which is rare.

With a sign of over-dispersion in Poisson regression, negative binomial is used to relax the assumptions stated earlier and to uses a different probability which allow for more variability in the data.

4.7. Results of Bayesian Analysis of Factors Affecting Diabetes

Bayesian analysis in this section was done in SAS, with procedure PROC REG for the Bayesian linear regression, PROC GENMOD for the Bayesian Poisson-gamma, and Negative binomial regression. PROC GENMOD works with the scale parameter that is related to the exponential family. With PROC GENMOD, the Gamermen algorithm is the default sampling method, and uses a non-informative prior (conjugate and Jeffrey prior).

The aim in Bayesian modelling is to determine the parameter(s) of interest of the posterior distribution of a predictor along with their credible interval. An accurate estimation of the posterior distribution can be difficult, and thus require a considerable amount of computation. Simulation-based methods are one of the most-used for such computations. A repeated simulation process draws samples from a target distribution and uses the collection of samples to empirically approximate the posterior. Somewhat, the credible interval is like confidence interval in classical statistics. The credible interval informs us that the true parameter of interest can be found within the range of the interval 95% of the time. The parameter of interest can be mean, mode or median. Here, the mean was the parameter of interest.

The Bayesian modelling procedure begins with the selection of a prior distribution for the predictor of interest, usually a non-informative prior that is conjugate to the distribution of the model. In the Bayesian context, the conjugate distribution is one where the posterior distribution for a model parameter has the same form as the prior. For example, the gamma distribution is conjugate for the Poisson mean, as is the beta distribution to negative binomial.

Markov chain convergence is an important focus in Bayesian analysis. A non-convergence chain does not explore the parameter space efficiently, and we say that the samples cannot approximately target the distribution accurately. However, inference should not be based upon unconvergence of the Markov chain, as this could produce a misleading result. Convergence terminologies include:

- Convergence: this explains drift in the samples towards a stationary (target) distribution
- Burn-in: This refers to samples at start of the chain that are discarded to minimize their impact on the posterior inference.
- Slow mixing: This is the tendency for high autocorrelation in the samples. A slow-mixing chain does not traverse the parameter space efficiently.
- Thinning: This is the practice of collecting every kth iteration to reproduce autocorrelation. However, thinning a Markov chain can be wasteful as one throws away a $\frac{k-1}{k}$th fraction of all the posterior samples generated. However, it is used to reduce autocorrelation and results in an independent sample that can be used otherwise. Similarly, autocorrelation does not lead to a biased Monte Carlo estimate, but it is simply an indicator of poor sampling efficiency.
- Trace plot: This is a visual plot of the sampled values of a parameter versus iteration number, often the most useful approach.

In general, the Bayesian methods can be more easily implemented for members of the family of generalized linear models (GLM) than for models not based on an exponential distribution (Hilbe, 2011). This is because the GLMs have clearly identified likelihood and deviance function, and many conjugate distributions are members of GLM. This compatibility of GLM and Bayesian methodology has allowed the SAS software to incorporate Bayesian modelling into its GLM procedure.

The results of the output in SAS are explained in the sections below.

4.7.1. Results of Bayesian Linear Model

The section below describes the posterior summaries and MCMC diagnostic convergence plots of the analysis, after running the MCMC algorithm for 1000000 iterations and thinning of 50 with burn-in size of 2000, with an improper normal prior distribution. Statistics of the posterior samples contain the size of the sample, the mean, the standard deviation and the quartiles for each model parameter. The fit statistics is displayed with the deviance information and the effective number of parameters. The interval estimates for the posterior sample table contain the HPD intervals and credible intervals for each model parameter. Since non-informative prior distributions for the regression coefficients were used, the mean and standard deviations of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

Table 4.20: Results of Bayesian Linear Regression

Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
Intercept	20000	-11.3423	4.8003	-14.5540	-11.3280	-8.1607
Health expenditure	20000	-0.2404	0.1696	-0.3542	-0.2399	-0.1269
GDP	20000	0.0002	0.0001*	0.0002	0.0002	0.0003
Population age	20000	0.3249	0.1017	0.2583	0.3254	0.3929
Urbanization	20000	-0.0555	0.0256*	-0.0719	-0.0556	-0.0393
Physician Density	20000	2.9456	0.8492	2.3907	2.9410	3.5129
Mean Year of School age	20000	0.1140	0.2028	-0.0223	0.1152	0.2502
Dispersion	20000	7.1133	1.5294	6.0318	6.9085	7.9515

Posterior Intervals

Parameter	Alpha	Equal-Tailed interval		HPD Interval	
Intercept	0.050	-20.72221	-1.8826	-20.7946	-1.9715
Health expenditure	0.050	-0.5720	0.0941	-0.05547	0.1085
GDP	0.050	-0.00001	0.00050	-0.00000008	0.000502
Population Age	0.050	0.1236	0.5232	0.1225.	0.5215
Urbanization	0.050	-0.1039	-0.0069	-0.1045	-0.0075
Physician density		1.2531	4.6161	1.3087	4.6554
Mean year of school age	0.050	-0.2849	0.5090	-0.2915	0.5003
Dispersion	0.050	4.7469	10.7112	4.4494	10.1647

DIC = 255.445
pD(effective number of parameters) = 8.139

The posterior summary of Bayesian linear regression in Table 4.20 shows that GDP is significant with mean = 0.0002 and standard deviation of 0.0001, which indicates that for every increase in GDP, there is a positive increase in diabetes prevalence by 0.0002. Also, a significant urban population with mean of -0.0555 and standard deviation of 0.0256 is an indication that for every unit in urbanization, there is a decrease in urbanization by 0.0555.

The bottom part of the Table 4.17 shows the posterior intervals, which are:

Equal tail: $100\ \left(\frac{\alpha}{2}\right) th\ and\ 100\ \left(1-\frac{\alpha}{2}\right) th\ percentiles$, and

Highest Posterior Density (HPD): posterior probability is 100 (1 - $100\ \left(1-\frac{\alpha}{2}\right)$)%

HPD interval for GDP is (0.0000008, 0.000508) which means that there is a 95% chance that GDP is in the credible interval of (0.00000008, 0.000508).

4.7.2. MCMC Diagnosis Plots of Diabetes and Risk Factors for Bayesian Linear Model

A convergence diagnostic test, as explained in section 3.7.6 above, was conducted to assess the convergence of the Markov chain before obtaining the poterior summary table, since the aim of MCMC diagnostic plots are to check convergence, autocorrelation, and the efficiency of the function. The SAS function produces, for each variable along with its dispersion and intercept, three plots which are, posterior density at the bottom right, autocorrelation at the bottom left, and trace plot at the top.

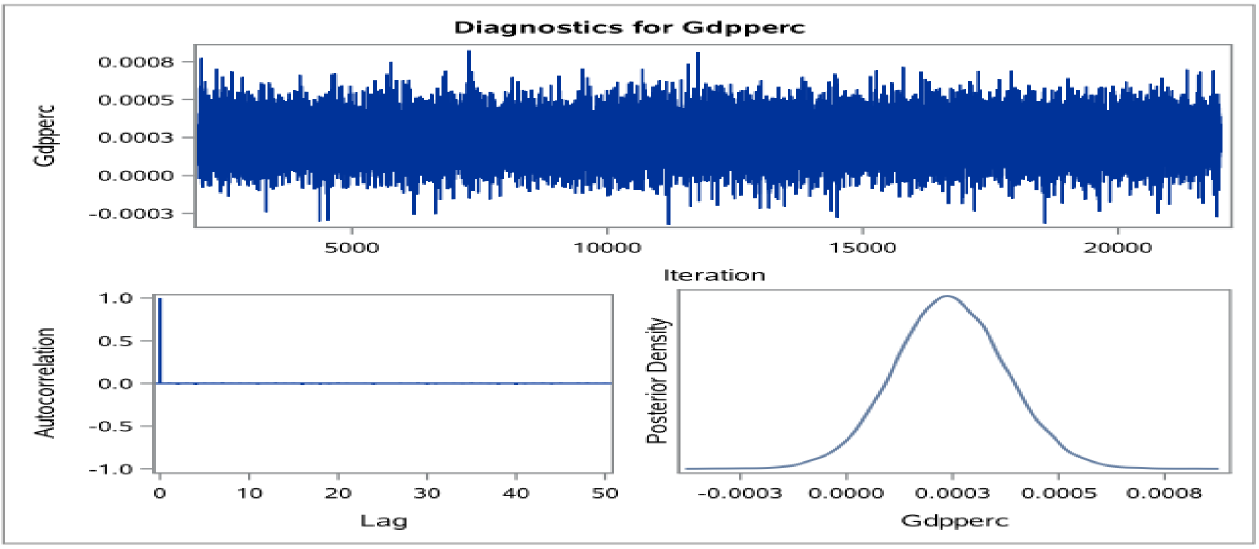


Figure 4.33. Diagnostic plot of GDP in Bayesian linear regression

The top frame shows that MCMC (Gamerman sampling algorithm) convergence steadily reaches the appropriate mean value of the posterior for all the variables. The lower left frame displays the autocorrelation and indicates the approximate independence of the sampling from the posterior distribution. All the variables thin out at around 1 thin. Thinning reduces autocorrelations and allows one to obtain seemingly independent samples. The plots show that there is no autocorrelation in the variables. The lower right displays the kernel density (posterior density) which estimates the marginal posterior distribution of the parameters using the sampling mechanism. The posterior mean is at the top of the distribution, which shows where most of mass rests.

From the diagnostic plots in Figure 4.33 and Appendix C, all the trace plots show a significant upward and downward trend along the iterations with the so called “thick pen” as described by Gelfand et al. (1990). This indicates insignificant deviations from the stationarity. The MCMC algorithm can then be considered to have converged. The density plot also shows a symmetrical distribution.

4.7.3. Results of Bayesian Poisson Model of Diabetes and Predictors

This section described the results of the Poisson model of diabetes and the predictors in SAS using PROC GENMOD, with non-informative prior distributions. By default, the maximum likelihood estimates of the model parameters were computed using a non-informative independent normal prior distribution with zero means and variance. The command produces varied outputs such as initial values for the Markov chain, fit statistics, interval statistics, posterior sample correlation matrix, and MCMC convergence diagnostics plots such as autocorrelations, posterior density, and trace plots.

Table 4.21 gives the descriptive statistics such as mean, standard deviation, and credible intervals of the estimate of the posterior distribution.

Table 4.21: Statistics of the Posterior samples of Bayesian Poisson model

Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
Intercept	50000	15.2881	0.00203	15.2868	15.2881	15.2895
Health expenditure	50000	-0.0789	0.00008	-0.0790	-0.0789	-0.0789
GDP	50000	-0.0001	0.00000	-0.0001	-0.0001	0.0001
Population age	50000	-0.0369	0.00004	-0.0369	-0.0369	-0.0368
Urban population	50000	-0.0021	0.00000	-0.0021	-0.0021	-0.0021
Physician Density	50000	1.0927	0.00014	1.0926	1.0927	1.0927
Mean Year of School age	50000	0.1975	0.00010	0.1975	1.1975	0.1976

Posterior Intervals

Parameter	Alpha	Equal-Tailed interval		HPD Interval	
Intercept	0.050	15.2841	15.2922	15.2841	15.2921
Health expenditure	0.050	-0.0791	-0.0787	-0.0791	-0.0788
GDP	0.050	-0.000010	0.00010	-0.00010	0.00010
Population Age	0.050	-0.0369	-0.0368	-0.0369	-0.0368
Urbanization	0.050	-0.00215	-0.00211	-0.00215	-0.00211
Physician density	0.050	1.0924	1.0929	1.0924	1.0929

DIC = 55635658
pD (effective number of parameters) = 6.985

The posterior distribution was obtained after one million iterations performed gradually while assessing convergence at every stage with burn-in of 50000 and thinning of 50. The posterior summary of Poisson model in Table 4.18 shows that all the variables are significant but indicates that for a one-unit increase in urbanization, the difference in the logs of expected counts of diabetes would be expected to decrease by 0.0021 units while holding other variables in the model constant. However, for a one-unit increase in physician density per 1000 persons, the difference in the logs of expected counts of diabetes would be expected to increase by 1.0927 units, while holding other variables in the model constant.

From the posterior interval result, in the Bayesian paradigm, the interpretation of the credible interval is that interval indicates that there is a 95% probability that the true parameter (for respective intervals) is included in the given interval. For instance, for the risk variable population age, there is a 95% probability that the parameter falls in the interval (-0.0369 -0.0368).

4.7.4. Convergence Diagnosis Results of Bayesian Poisson Model of Diabetes and Covariates

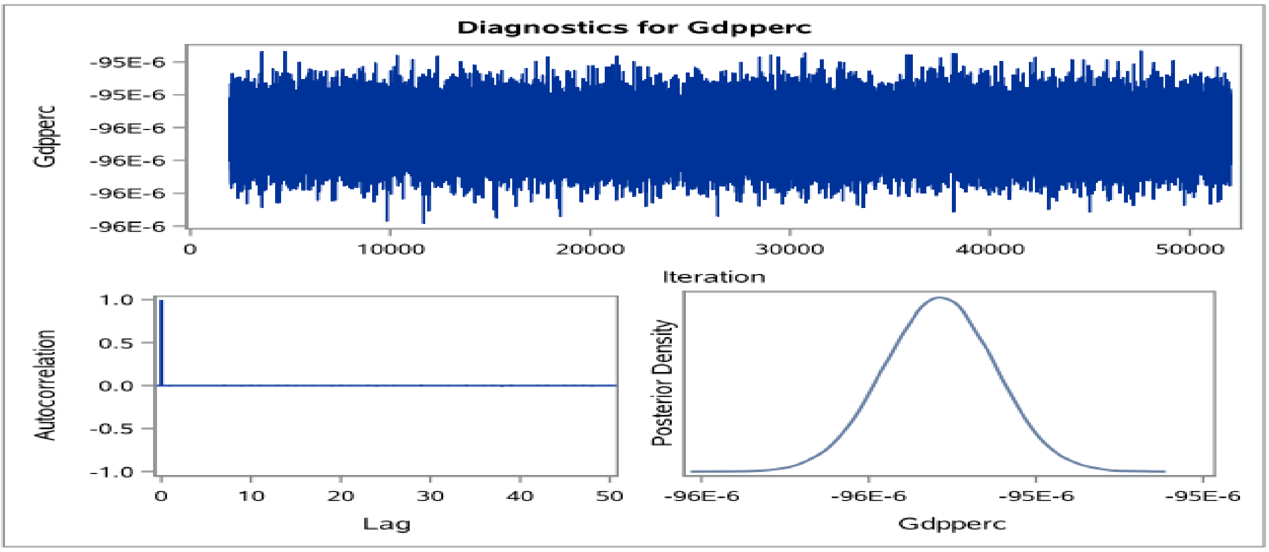


Figure 4.34. Poisson diagnostics plots for health expenditure

The diagnostic plots in figure 4.34 show that there is convergence in all the samples, that is, initial drift in all the parameter toward a stationary distribution. The trace plots show a good-mixing with no autocorrelation and asymmetric posterior density, indications of a good model. Other variable diagnostic plots for the Poisson is shown in Appendix D.

4.7.5. Results of Bayesian Negative Binomial Model of Diabetes and Predictors

Table 4.22: Posterior Summaries Negative Binomial model

Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
Intercept	20000	14.0835	2.1926	12.6105	14.0979	15.5586
Health expenditure	20000	-0.1408	0.1003	-0.2090	-0.1423	-0.0739
GDP	20000	-0.0001	0.00009*	-0.0002	-0.0001	0.00007
Population age	20000	0.0054	0.0502*	-0.0284	0.0047	0.0387
Urban population	20000	-0.0189	0.0184*	-0.0312	-0.0189	-0.0064
Physician Density	20000	1.4146	0.5808	1.0015	1.3446	1.7641
Mean Year of School age	20000	0.1864	0.1218	0.1051	0.1867	0.2677
Dispersion	20000	1.5688	0.2856	1.3632	1.5377	1.7415

Posterior Intervals

Parameter	Alpha	Equal-Tailed interval		HPD Interval	
Intercept	0.050	9.7301	18.3371	9.8197	18.3830
Health expenditure	0.050	-0.3358	0.00582	-0.3340	0.0599
GDP	0.050	-0.00026	0.000079	-0.00027	0.000057
Population Age	0.050	-0.0913	-.1059	0.0977	0.0990
Urbanization	0.050	-0.0551	0.0175	-0.0536	0.0187
Physician density	0.050	0.4638	2.7211	0.3557	2.5668
Mean Year school age	0.050	-0.0541	0.4250	-0.0556	0.4222
Dispersion	0.050	1.0978	2.2184	1.0648	2.1547

DIC = 1477.85

PD (effective number of parameters) = 7.332

(* significant variables)

Table 4.22 shows the posterior summaries of a negative binomial model with GDP, Urbanization, and Population age being significant. The posterior mean of GDP is -0.0001 and SD of 0.00009. Similarly, a population age posterior means of 0.0054 and SD of 0.0502 shows an increase in diabetes prevalence is associated with increase in population age, that is a high probability population age predicting diabetes prevalence. The urbanization posterior mean estimate is 0.0189, with an SD of 0.0184 shows a low prediction on diabetes, indicating that an increase in diabetes prevalence is associated with a decrease of 0.0189 in urbanization. For negative binomial, the dispersion parameter is an important parameter, dispersion mean is given as 1.5688 and SD is 0.2856.

From the posterior interval result, in the Bayesian paradigm, the interpretation of the credible interval is that it indicates that there is a 95% probability that the true parameter (for the respective intervals) is included in the given interval. For instance, for the risk variable GDP, there is a 95% probability that the parameter falls in the interval (-0.00027, 0.00057), for urbanization credible interval (CI) is

(-0.0536, 0.0187) and credible interval for population age is (0.0977 ,0.0990).

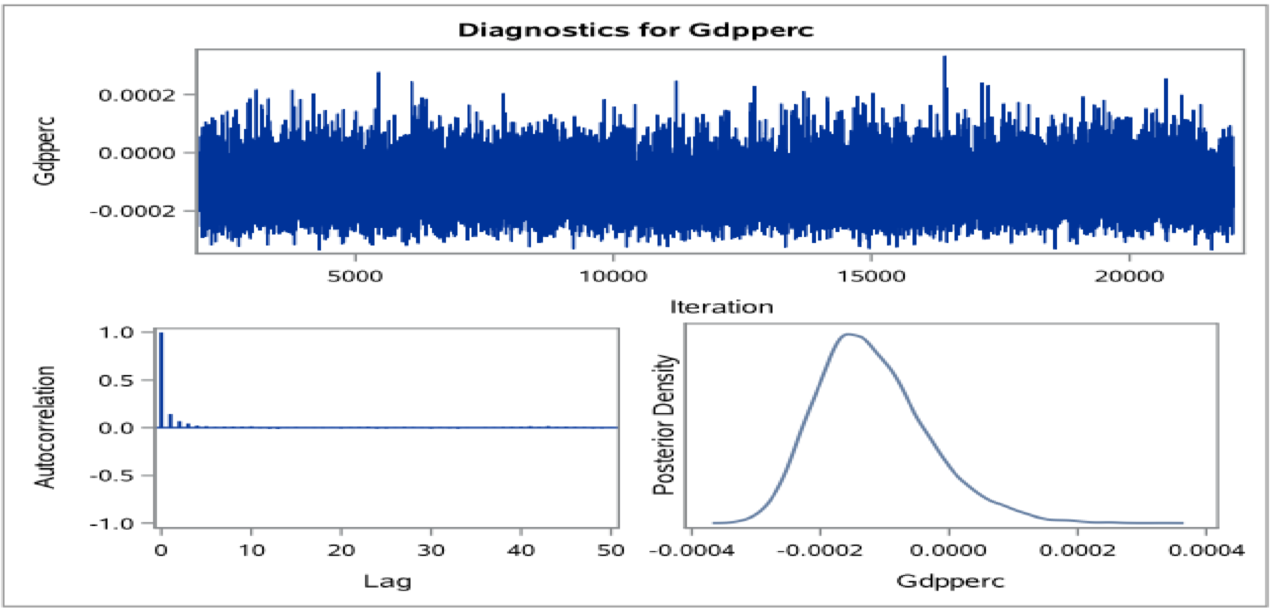


Figure 4.35. Negative binomial diagnostic plots for GDP

Negative diagnostic plot for GDP in figure 4.35 shows a stationary plot and no autocorrelation. Other variables diagnostic plots for the Poisson is shown in Appendix E with similar results.

4.8. **Conclusion**

We computed a linear regression model using all twelve risk variables from the data, but due to linearity and high missing values of some of the variables, the number variables were reduced to six. The six variables used were GDP, health expenditure, population age, urbanisation, physician density, and mean year of school age. The linear regression model revealed that population age, urbanization, and physician density were significant.

A transformation of the data by taking the log of the dependent variable was deemed fit to further assess the normality of the distribution. The log-linear model results show that population age and physician density were significant. Further transformation by using a fractional probit regression model revealed that GDP, population age, and physician density were significant risk factors for diabetes. In modeling count data, a Poisson model was deemed fit, and the Poisson regression model was computed which showed that all the variables were significant. To relax the assumption of the Poisson regression, and due to over-dispersion, a negative binomial model was computed which revealed that GDP and physician density are significant risk factors for diabetes.

A Bayesian linear regression was computed from a GLM perspective, Bayesian Poisson model and Bayesian Negative binomial model for diabetes and the social- economic factors at the population level was computed to compare the classical result. A non-informative prior probability distribution was used in the building of the model. From the Bayesian paradigm, the Bayesian linear regression revealed that GDP and urbanization were significant risk factors of diabetes. We computed the Bayesian Poisson regression, and it also revealed that all the variables are significant. The Bayesian negative binomial model revealed that GDP, Urbanization, and population age are significant socioeconomic risk factors for diabetes.

In general, it was revealed that physician density, GDP, population age, and urbanization were consistently significant in all the models. This implies that countries with high GDP will be associated with high diabetes. This shows the economic impact diabetes imposes on the economy of a nation, the global health care system, and the wider global economy. This burden of diabetes as measured through direct and indirect cost has a negative impact on the nations' GDP. Overall, the population age (15-65) was also significant, implying that this age group contributed significantly to diabetes prevalence on the continent. Also, results showed that as life expectancy increases, diabetes prevalence increases. Likewise, a significant physician density implies that access to health care system, health capability, the number of physician servicing the population play a significant role in diabetes prevalence. Lastly, it was revealed that urbanization was significant, which implies that as countries continue to urbanize, the changing pattern of people's standards of living has a positive impact on diabetes prevalence, that it is Urbanization increases diabetes prevalence in Africa.

Chapter 5

Discussion and Conclusion

5.0. Introduction

This chapter gives a general discussion of the results, the conclusions of the study, its implications, and recommended areas of further work/study.

5.1. Discussion

This study has the main aim of investigating the spatial distribution of diabetes prevalence across Africa and the associated risk factors using both classical and Bayesian approach. The first objective of the study was to determine the relationship between identified factors of diabetes and its prevalence in Africa. The second objective was to examine spatial configurations of diabetes prevalence levels and its underlying socio-economic factors in Africa. The third objective was to compare Bayesian statistics and classical statistics as applied in spatial clustering. However, identification of high diabetes mortality risk countries and the underlying risk factors is generally hampered by lack of reliable data. This study attempts to address this lacuna and mostly succeeds.

The first objective of the study was met by using different techniques such as scatter plot, spearman rho correlation coefficient ordinary least square regression and geographic weighted regression to model the predictor and the response variables to determine the relationship between diabetes prevalence and the associated risk factors.

Country GDP, Age dependency ratio, physical inactivity and physician were each significantly associated with country diabetes prevalence, relationships that were spatially nonstationary across the Africa continents. The variation in parameter estimates from GWR suggests the need to apply this spatial analysis tool to other diabetes studies that have been restricted to global models. In the global OLS model, 32.2% of county-level diabetes prevalence was explained by population age, alcohol consumption, health expenditure, country GDP, the mean year of school age as well as physical inactivity of the population. However, at an individual country level, the explanatory percentage increased to 100%, and the individual county-level models were significantly clustered. This clustering suggests that local contexts, policies, programs, and built environment attributes are associated with diabetes prevalence and that the amplitude of such contexts, policies, programs, and environments varies across the nation. The relationships among physician density, physical inactivity, and diabetes are not new (Green et., al. 2003, Hu et., al. 1999). Others found these relationships have a spatial component (Hu et., al. 2004). Except for recent work by Siordia and colleagues (2012), there has been no investigation into the non-stationarity of these relationships. Similarly, the strong association between physical activities and diabetes is consistent with findings of other studies (Hu et., al. 2004), but it has not been investigated for spatial heterogeneity or nonstationary. That there is a significant association between obesity, GDP, physical inactivity, age dependency ratio, physician density and diabetes and that this relationship has a spatial but nonstationary association highlights the need for local, context-specific diabetes prevention programs. Using GWR, public health researchers and practitioners can gain a nuanced understanding of health-related issues and respond to the notion that "all health is local" (Tobler 1970, Gebreab and Diez Roux, 2012). In doing so, they can provide clarity for designing and funding context-specific public health programs and policies, especially for national programs that have local reach, including those of the Centre for Disease and Control (CDC), IDF, UN and WHO. Our analyses could also be used by local public health departments and offices for resources such as MIYO (Make It Your own - http://www.miyoworks.org/) to tailor messages and materials for their target audiences. The use of GWR is a key advancement in public health research and practice because many health behaviours and outcomes vary spatially (e.g., obesity) (Kohl et., al. 2013).

Shedding light on spatial variations can provide new insights into well-established relationships. The methodology of GWR needs to be expanded to additional public health efforts to understand the impact of environment and place on health and how these relationships may vary across space. For diabetes prevalence, we presented an initial step in this direction, but much work remains before we understand why these variations exist and why alcohol consumption, population age, health expenditure, GDP, and physical inactivity have little explanatory effect in some regions but explain up to 100% of diabetes prevalence in other regions.

To achieving the second objective of this study, by examining the spatial configuration of diabetes in Africa, different techniques were implemented for comparison purposes. Moran I indices both under randomization and specific distance measures were used to determine the spatial configuration. LISA and cluster analysis were also implemented in determining which countries have similar patterns of diabetes prevalence. Firstly, it was revealed that the disease is heterogeneous across the Africa continent as stated by (Schwitzgebel, 2014; Tuomi et al., 2014). Secondly, diabetes was seen to be spatially correlated and clustered across the continent. By further stratifying the data into regions, in the West, Central and North regions, diabetes was seen to be dispersed, while clustered in the East and South regions. Under randomization, the West, Central, and East regions showed a negative spatial correlation, while the South and North regions showed positive spatial correlation of diabetes. This could be because of the elevated level of risk factors of diabetes in these regions. For example, countries like South Africa, Botswana, Gabon, Libya, Mauritius, Equatorial Guinea, and Seychelles are seen to have high GDP, obesity, alcohol consumption, physician density, urban population age, urban population growth, health expenditure and age dependency, which shows an enormous impact on the countries high diabetes prevalence rate. Likewise, countries with low risk factor rates show low diabetes prevalence rates. In general, the North and South regions have high diabetes prevalence and high-risk factors of diabetes. This shows a relative linear association between the risk factors and diabetes. Also, many countries in these regions were seen to have high GDP, high population age, high physician density, and high urbanization.

The study likewise gives a big picture of countries that are referred to as hot spots, as well as low spots in terms of diabetes prevalence, considering the factor of neighboring countries. That is, a country can be classified as having a high prevalence depending on both her risk factors as well as those of her neighboring countries. The results of LISA cluster map analysis and scatter plots for the five regions show that the following countries have high diabetes prevalence with neighbouring countries which also have high diabetes prevalence (in the high-high category): Benin, Guinea, Guinea Bissau, Cameroon, Equatorial Guinea, Rwanda, Angola, Mozambique, Zambia, Sudan, Algeria, and Burundi. This High-High clusters identifies swath of the countries in which diabetes is high, interventions to prevent complications from diabetes might be needed in these areas.

Similarly, the countries with high diabetes prevalence with neighboring countries with low diabetes prevalence rates are Cote dʾIvoire, Liberia, Senegal, Chad, Eritrea, Tanzania, Lesotho, Libya, Swaziland. These High- low clusters categories serve to remind policy makers that a country may need intervention even when its neighbour needs are less severe.

The countries with low diabetes prevalence with neighboring countries with low diabetes prevalence rates (in the low-low category) are Burkina Faso, Nigeria, Guinea Bissau, Central Africa Republic, Gabon, Somalia, Djibouti, Namibia, Botswana, and Zimbabwe. These low-low clusters identified broad areas of the country where diabetes is less common; perhaps there are lessons from these areas that can be applied elsewhere, such as in countries in H-H and H-L.

Similarly, countries with low diabetes prevalence and having neighboring countries with high diabetes prevalence rates (in the low-high category) are South Africa, Mali, Sierra Leone, Congo Republic, Congo DRC, Ethiopia, Uganda, and Madagascar. The low –high clusters identifies pockets with regions in which effort to prevent diabetes have perhaps been successful or have lowered risk factors for diabetes. All this knowledge along with information on the influence of associated socio-economic risk factors moves us toward better preventing diabetes and its complications.

The third objective of this population-based study was met by modelling diabetes prevalence in Africa using both classical and Bayesian statistical methods. These techniques used are simple linear regression, log linear regression, fractional probit, Poisson regression, negative binomial, Bayesian linear model, Bayesian Poisson regression model, and Bayesian Negative-binomial model. This is to ascertain the extent to which diabetes is related to various risk factor variables. Because of over-dispersion of the study population and since the variables of the study are in the form of continuous count data, GLM using a Poisson distribution and negative binomial was deemed an appropriate technique for analyzing the data. In examining the impact of the explanatory risk factor variables of diabetes, the classical approach revealed that GDP, Population age, physician density, and urbanization are significant. On the contrary, there was no significant difference in the results from classical and Bayesian techniques, due to the use of a non-informative prior. As revealed in the literature (Mbanya et. al, 2010; Hu, 2011; Bruro & Landi, 2011; Levitt, 2008; Gill et. al., 2009, Chen et. al., 2012; Mensah et. al, 2014; Hall et. al., 2011; Motala, 2008; Hu, 2011; Bruno and Landi, 2011; Levitt, 2008; Ramachandran et. al., 2007; Hopkins et. al, 2010), this study shows that health care access, population age, urbanization, and GDP have an impact on the increase of diabetes in Africa. We show that countries with high GDP, high obesity, high MYSC, high physician density, high population age, and high urbanization also have high diabetes prevalence rates.

Lastly, the study found that obesity, physical activity, alcohol consumption, physician density, GDP, GNI, Population age, urban population growth, and health expenditure are key risk factors which are associated with and have an impact on diabetes prevalence in Africa. Countries with elevated risk factor rates also tended to have high diabetes prevalence. Egypt, Seychelles, Mauritius, Libya, Comoros, Tunisia, Sudan, Djibouti, South Sudan, Gabon, Equatorial Guineas, Algeria and South Africa were identified as the top countries with high diabetes prevalence. The high prevalence in these countries can be attributed to their high GDPs, physician densities, population ages, and population growth factors. The countries with low prevalence rate on the continent are Benin (0.8%), Gambia, Burkina Faso, Niger, Mali, Senegal, Sierra Leon, Guinea, Guinea Bissau, Cabo Verde, Ghana, Liberia, Mauritania, Nigeria, Sao Tome and Principe, and Kenya, all having less than 2.3% diabetes prevalence rates. This can be attributed to their relatively low rates of the risk factors GDP, Urban population growth and physician density.

Our study has several limitations. First, the diabetes prevalence estimates were modelled from survey data, and we did not account for the survey sampling uncertainty or the biases and limitations of the survey. Second, we did not consider changes over time and therefore do not know how rapidly the diabetes prevalence in our spatial clusters could change. Third, we did not account for the movement of people between countries in our estimates of prevalence of diabetes. It is not known what percentage of the residents developed diabetes while residing in another county. Fourth, the Moran I for the regions was not included due to small sample size, as the number of countries in those regions were included in the data and not local counties or municipalities level.

5.2. Conclusion

This study of diabetes prevalence in Africa has been modeled using Bayesian GLM and classical statistics to ascertain the extent to which diabetes prevalence is related to various explanatory risk variables.

One of the objectives of this study was to examine the spatial configuration distribution of diabetes across Africa. This was achieved using GMI, Moran’s I, and LISA map analyses.

From the LISA results, outlier countries were identified, that is, countries with high diabetes prevalence, but surrounded by countries with low diabetes (South Africa, Mali, Sierra Leone, Congo Republic, Congo DRC, Ethiopia, Uganda, Madagascar). Countries with low diabetes but surrounded with countries with high diabetes (Cote d'Ivoire, Liberia, Senegal, Chad, Eritrea, Tanzania, Lesotho, Libya, Swaziland) were also identified. This pattern should be given a close attention as it has implications on the spread of diabetes on the continent. As mentioned in the literature review (Motala, 2002; Ogurtsova et. al, 2015; Peer et. al, 2014), it has been discovered that a good percentage of these countries are high-income countries and the most populous ones. However, risk factors are not limited to high income or populous countries as Non-communicable disease (NCD) burden is common to developing countries. As positioned by the International Diabetes Federation (IDF), in efforts towards halting the rise of diabetes, a global network for diabetes awareness at the country and continental levels can be created and facilitated through joint policies and collaborative efforts (Day, 2016). The result of LISA and cluster analysis from this study could be used as a tool or instrument in mapping the joint policies and collaborative effort in the fight against diabetes instead of individual country fighting the disease in silos. Likewise, countries in this region need to improve on non-communicable disease surveillance and monitor policies can be formed with best evidence. Efficient and continuous program for prevention and awareness need to be put in place.

In conclusion the use of cluster analysis and LISA map helps in strengthening the results, due to the heterogeneity of the population and diabetes in Africa, the cluster analysis brings together the homogenous groups of countries in the same clusters, that is countries with similar pattern in terms of the level of their prevalence and risk factors can come together to form joint policies. On the similar pattern, the hot spots countries from LISA maps can be further identified and re- grouped from their various clusters and then given more attention and urgent action in order to reduce diabetes burden in the continent.

To achieve one of the study's objectives of comparing classical and Bayesian statistics, we fitted several models such as linear regression, Poisson, and negative binomial from the classical perspective, and their respective Bayesian models from the Bayesian perspective. The Bayesian linear, Poisson, and negative binomial models were fitted using the MCMC estimation method. The regression analysis showed that GDP, physician density, urbanization, and population age were significant across the models, with population age and physician density having positive significance in all the regression analyses. The differences between the classical and Bayesian approaches were observed, as well as differences in the mean and the parameter of estimate. We can conclude that the use of Bayesian techniques did not make a significant difference in the result due to the use of a non-informative prior. Using information from previous years can help strengthen the use of Bayesian techniques.

Overall, the highest prevalence is seen within the North, Central, East and South regions, while the West Africa region having has the prevalence. The high prevalence rate within the North, Central, East, and South regions could be attributed to increasing urbanization, demographic changes, and increases in population within the older age groups with the associated changes in levels of risk factors such as obesity, physical inactivity, alcohol consumption, and tobacco smoking. Countries in Sub-Saharan Africa, especially South Africa, are currently known to be undergoing a wide range of epidemiological transitions, with an increasing burden of non-communicable disease (Azevedo and Alla, 2008; Mbanya et al., 2010; Sobngwi et al., 2012) However, the determinant factors for low diabetes prevalence in the West Africa region could be the low rates of obesity, physical inactivity, and urbanization in most of the countries as compared with the other regions. Another contributing factor of low diabetes prevalence in this region could be the climate condition (Adeghate et al., 2006; Fisch et al., 1987; Vandenheede et al., 2012; Aikins, 2005; Tishkoff and Williams, 2002).

In determinng the relationship between diabetes and the socio economic risk factors, OLS and GWR was compared and the strenght of GWR was established. The improved model from local GWR shows that Obesity,Urban population growth, age dependency ratio, physician density , HDI and GNI are significant variables. However,there are also limitations to our findings. The local R^2s accounted for 100% of country-level diabetes prevalence. The residual parameters was not tested for spatial correlation because it was not part of the study objective. A further analysis could have been done by mapping the residuals to visualised the variables that are significants and then identify which country they are most significant. The primary strength of this study is the use of GWR in the analysis of the spatial distribution and correlates of diabetes prevalence. Siordia and colleagues (7) introduced the concept of spatial.

5.3. Implications and Recommendations

In an attempt to reduce the number of people diagnosed or living with diabetes in Africa, efforts should be directed towards having joint policies towards the implementation, control, and planning of diabetes awareness and care in Africa either at the regional level or among countries in the same clusters or hotspots countries. In other words, a collaborative initiative is recommended between countries and key international and national diabetics organizations to improve diabetes care and awareness in Africa and worldwide.

Strategic actions should be taken in terms of meaningful population-based intervention policies and projects, management of urbanisation, and programs to fight diabetes in Africa both at the national and at the international level such as the NCD alliance strategy plan by United Nation for NCDs.

The identified hotspot countries can be used as a form of NCDs alliance strategy to curb this prevalence. Explicit or implicit measures should be instituted by governments to influence population size, growth, distribution, and composition, as well as the health care management of diabetes. Likewise, risk factors that can be controlled such as physician density and urban population growth can be used as instruments in reducing diabetes in the Africa.

Special attention should be given to countries in the Southern and Northern regions of Africa like South Africa, Botswana, Mauritius, Namibia, Libya, Egypt, Seychelles, and Egypt where the diabetes prevalence is high and the significant risk factors such as GDP, population age, urban population growth and physician density are also high.

5.4. Further Study

In the analysis, this study puts into consideration the way data was extracted. Further work can therefore be done on Bayesian spatial models with applications to other disease in Africa or specific countries (e.g. South Africa). One of the limitations to this study was the unavailability of diabetes prevalence data for previous years. A trend of the disease and the risks could have been investigated, leading to a strengthening of the Bayesian technique. Therefore, an area of further work could be the acquisition of the said data and conducting of such investigations, based on which stronger Bayesian analyses can be done. Finally, further work can be done on the contribution of ethnicity to the spread of diabetes in Africa.

Lastly, with the identification of the hot spots countries, an intervention project could be considered with the government of these countries to curb the mortality of diabetes. A cohort study on each of these hot spots countries on the individual level can be done to further identify the hotspots states, district, province or town where diabetes is prevalence. This will assist in the fight against the increase of diabetes and a systematic reduction can be realised as against the projected figures for 2020.

Bibliography

ABRAHAM, T.M., PENCINA, KM., PENCINA, M.J and FOX, C.S (2015). Trends in Diabetes incidence: The Framingham Heart Study. American Diabetes Association: Diabetes Care 38(3): 482-487.

ADA, (2014). Diagnosis and classification of diabetes mellitus. American Diabetes Association: Diabetes care, Issue 37, S81-S90.

ADEGHATE, E., SCHATTNER, P. & DUNN, E. (2006). An update on the etiology and epidemiology of diabetes mellitus. *Annals of the New York Academy of Sciences,* 1084**,** 1-29.

AIKINS, A. D.-G. (2005). Healer shopping in Africa: New evidence from rural-urban qualitative study of Ghanaian diabetes experiences. The *BMJ,* 331**,** 737.

ALEXANDER, N., MOYEED, R. & STANDER, J. (2000). Spatial modelling of individual-level parasite counts using the negative binomial distribution. Biostatistics, 1, 453-463.

ALWAN, A. (2011). *Global status report on noncommunicable diseases (2010)*, World Health Organization.

ANDERSON, R. J., FREEDLAND, K. E., CLOUSE, R. E. & LUSTMAN, P. J. (2001). The prevalence of comorbid depression in adults with diabetes. American Diabetes Association: Diabetes care, 24, 1069-1078.

ANSELIN, L, SRIDHARAM, S., GHOLSTON, S. Using exploratory spatial data analysis to leverage social indicator databases. The discovery of interesting patterns. Social Indicators 2007; 82(2): 287-09.

ANSELIN, L., SYABRIL, L., KHO, Y. (2006). GeoData: An introduction to spatial data analysis. Anal Journal of Geography 2006; 38(1): 5-22.

ARSLAN, O., CEPNI, M. & ETILER, N. (2013). Spatial analysis of perinatal mortality rates with geographic information systems in Kocaeli, Turkey. Journal of public health. 10.1016/j.puhe.2012.12.009*,* 127**,** 369-379.

ASSAH, F. K., EKELUND, U., BRAGE, S., MBANYA, J. C. & WAREHAM, N. J. (2011). Urbanization, physical activity, and metabolic health in sub-Saharan Africa. American Diabetes Association: *Diabetes Care,* 34**,** 491-496.

ASPRAY, T.J., MAGUSI, F., RASHID, S., WHITING, D (2000). Rural and urban difference in Tanzania: the role of obesity, physical inactivity and urban living.

AZEVEDO, M. & ALLA, S. (2008). Diabetes in sub-saharan Africa: kenya, mali, mozambique, Nigeria, South Africa and zambia. *International journal of diabetes in developing countries,* 28**,** 101.

BARKER, L. E., KIRTLAND, K. A., GREGG, E. W., GEISS, L. S. & THOMPSON, T. J. (2011). Geographic distribution of diagnosed diabetes in the US: a diabetes belt. American Journal of Preventive medicine, 40, 434-439.

BARRON, D. N. (1992). The analysis of count data: Over dispersion and autocorrelation. Sociological methodology, Vol 22 (1992), pp 179-220. American Sociological Association. JSTOR, www.jstor.org/stable/270996.

BASÁÑEZ, M.-G. (2009). Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA: Spatial Analysis in Epidemiology. Oxford University Press, Springer, New York.

BATES, D. (2005). Fitting linear mixed models in R. R news, 5, 27-30.

BATES, D. M. (2010). lme4: Mixed-effects modelling with R. Oxford University Press, Springer, New York.

BATES, D., MAECHLER, M., BOLKER, B. & WALKER, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1, 1-23.

BELLAVANCE, F., DIONNE, G. & LEBEAU, M. (2009). The value of a statistical life: A meta-analysis with a mixed effects regression model. Journal of Health Economics, 28, 444-464.

BERK, R. & MACDONALD, J. M. (2008). Overdispersion and Poisson regression. *Journal of Quantitative Criminology,* 24**,** 269-284.

BERNARDO, J. M. & RUEDA, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review,* 70**,** 351-372.

BEST, N. G., ICKSTADT, K. & WOLPERT, R. L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association,* 95**,** 1076-1088.

BIESSELS, G. J., VAN DER HEIDE, L. P., KAMAL, A., BLEYS, R. L. & GISPEN, W. H. (2002). Ageing and diabetes: implications for brain function. *European journal of pharmacology,* 441**,** 1-14.

BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. & WHITE, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution, 24, 127-135.

BONOW, R. O. & GHEORGHIADE, M. (2004). The diabetes epidemic: a national and global crisis. *The American Journal of Medicine,* 116**,** 2-10.

BOOTH, G. L., KAPRAL, M. K., FUNG, K. & TU, J. V. (2006). Relation between age and cardiovascular disease in men and women with diabetes compared with non-diabetic people: a population-based retrospective cohort study. *The Lancet,* 368**,** 29-36.

BOUCHER, J.-P. & DENUIT, M. (2006). Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance. *ASTIN Bulletin: The Journal of the IAA,* 36**,** 285-301.

BRUNSDON, C, FOTHERINGHAM, A.S, AND CHARLTON, M.E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis 28(4): 281-298.

BRUNSDON, C, FOTHERINGHAM, A.S, AND CHARLTON, M.E. (1998). Geographically weighted regression: Modelling spatial non-stationarity. The statistician 47(3): 431-443.

BURNHAM, K. P. & ANDERSON, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, 33, 261-304.

CAMERON, A. C. & TRIVEDI, P. K. (2013). *Regression analysis of count data*, Cambridge university press, New York, .

CARD, N. A. (2015). Applied meta-analysis for social science research, The Guilford Press, New York, NY 10012.

CHARLTON, M AND FOTHERINGHAM, A.S. (2009). Geographically weighted regression. European Journal of Geostatistics.

CHEN, D.R., WEN, T.H. (2010). Elucidating the changing socio-spatial dynamics of neighbourhood effects on adult obesity risk in Taiwan from 2001 to 2005. Health Place; 16(6): 1248-58.

CHRISTENSEN, O. F. & WAAGEPETERSEN, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics Journal,* 58**,** 280-286.

CHRISTIANSEN, C. L. & MORRIS, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association,* 92**,** 618-632.

CLAESKENS, G. & HJORT, N. L. (2008). Model selection and model averaging, Cambridge University Press Cambridge.

CLAYTON, D. G. (1996). Generalized linear mixed models. Markov chain Monte Carlo in practice. University Press, Springer, New York.

EMERGING RISK FACTORS COLLABORATION. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. The Lancet, 375, 2215-2222.

CONDOLEO, R., MUSELLA, V., MAURELLI, M. P., BOSCO, A., CRINGOLI, G. & RINALDI, L. (2016). Mapping, cluster detection and evaluation of risk factors of ovine toxoplasmosis in Southern Italy. *Geospatial health,* 11.

CONGDON, P. (2014). Applied Bayesian modelling, John Wiley & Sons, New York.

COOPER, H., HEDGES, L. V. & VALENTINE, J. C. 2009. The handbook of research synthesis and meta-analysis, Russell Sage Foundation.

COWLES, M. K. 2013. Applied Bayesian statistics: with R and Open BUGS examples, Springer Science & Business Media.

COXE, S., WEST, S. G. & AIKEN, L. S. 2009. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. Journal of personality assessment, 91, 121-136.

DAUXOIS, J.-Y., DRUILHET, P. & POMMERET, D. 2006. A Bayesian choice between Poisson, binomial and negative binomial models. Test, 15, 423-432.

DAY, C. 2016. Reflections from IDF-WDC 2015. *British Journal of Diabetes,* 16**,** 37-38.

DE GROOT, M., AUSLANDEN, W., WILLIAMS, J. H., SHERRADEN, M. & HAIRE-JOSHU, D. 2003. Depression and poverty among African American women at risk for type 2 diabetes. Annals of Behavioural Medicine, 25, 172-181.

DE RAMIREZ, S. S., ENQUOBAHRIE, D., NYADZI, G., MJUNGU, D., MAGOMBO, F., RAMIREZ, M., SACHS, S. E. & WILLETT, W. 2010. Prevalence and correlates of hypertension: a cross-sectional study among rural populations in sub-Saharan Africa. *Journal of human hypertension,* 24**,** 786-795.

DEY, D. K., GHOSH, S. K. & MALLICK, B. K. 2000. Generalized linear models: A Bayesian perspective. CRC Press.

DOBSON, A. J. & BARNETT, A. 2008. An introduction to generalized linear models. CRC press, Florida, United State.

DRUCK, S., CARVALHO, M., CÂMARA, G. & MONTEIRO, A. 2004. Spatial analysis of geographic data. Spatial analysis of geographic data. Embrapa Cerrados, Planaltina, Brazil.

DUNCAN, G. J., DALY, M. C., MCDONOUGH, P. & WILLIAMS, D. R. 2002. Optimal indicators of socioeconomic status for health research. *American journal of public health,* 92**,** 1151-1157.

ELLIOT, P., WAKEFIELD, J. C., BEST, N. G. & BRIGGS, D. 2000. *Spatial epidemiology: methods and applications*, Oxford University Press.

ELLIOTT, P. & WARTENBERG, D. 2004. Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives***,** 998-1006.

ESTEGHAMATI, A., GOUYA, M. M., ABBASI, M., DELAVARI, A., ALIKHANI, S., ALAEDINI, F., SAFAIE, A., FOROUZANFAR, M. & GREGG, E. W. 2008. Prevalence of diabetes and impaired fasting glucose in the adult population of Iran. *Diabetes care,* 31**,** 96-98.

EVERSON, S. A., MATY, S. C., LYNCH, J. W. & KAPLAN, G. A. 2002. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *Journal of psychosomatic research,* 53**,** 891-895.

FÆRCH, K., HULMÁN, A. & PJ SOLOMON, T. 2016. Heterogeneity of pre-diabetes and type 2 diabetes: implications for prediction, prevention and treatment responsiveness. Current diabetes reviews, 12, 30-41.

FISCH, A., PICHARD, E., PRAZUCK, T., LEBLANC, H., SIDIBE, Y. & BRÜCKER, G. 1987. Prevalence and risk factors of diabetes mellitus in the rural region of Mali (West Africa): a practical approach. *Diabetologia,* 30**,** 859-862.

FISCHER, M. M. & GETIS, A. 2009. *Handbook of applied spatial analysis: software tools, methods and applications*, Oxford University Press, Springer Science & Business Media, New York.

FITZMAURICE, G. M., LAIRD, N. M. & WARE, J. H. 2012. Applied longitudinal analysis, John Wiley & Sons, New York, United State.

FONG, Y., RUE, H. & WAKEFIELD, J. 2010. Bayesian inference for generalized linear mixed models. Biostatistics, 11, 397-412.

FOTHERINGHAM, S. & ROGERSON, P. 2013. *Spatial analysis and GIS*, CRC Press , Florida , United State.

FRANK, B.O., ALFRED, A.D and ALFRED, S (2012). Evaluating Spatial and Space-Time Clustering of Cholera in Ashanti-Region-Ghana, Cholera, Dr. Sivakumar Gowder (Ed.), ISBN: 978-953-51-0415-5, InTech, Available from: http://www.intechopen.com/books/cholera/evaluating-spatial-and-space-time-clustering-ofcholera-in-ashanti-region-ghana

FROME, E. L. & CHECKOWAY, H. 1985. Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology,* 121**,** 309-323.

GARTNER, D. R., TABER, D. R., HIRSCH, J. A. & ROBINSON, W. R. 2016. The spatial distribution of gender differences in obesity prevalence differs from overall obesity prevalence among US adults. *Annals of epidemiology,* 26**,** 293-298.

GASKIN, D. J., THORPE JR, R. J., MCGINTY, E. E., BOWER, K., ROHDE, C., YOUNG, J. H., LAVEIST, T. A. & DUBAY, L. 2014. Disparities in diabetes: the nexus of race, poverty, and place. American journal of public health, 104, 2147-2155.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. 2014. Bayesian data analysis, Chapman & Hall/CRC Boca Raton, FL, USA.

GREEN C, HOPPA, R.D., YOUNG, T.K., BLANCHARD, J.F. (2003) Geographic analysis of diabetes prevalence in an urban area. Social Science Medicine 57: 551–560.

GODFREY, R., JULIEN, M. (2005) urbanisation and health. Journal of clinical medicine. Vol 5 no2- 137-141

GOOVAERTS P. (2010). Geostatistical analysis of county-level lung cancer mortality rates in the South Eastern United States. Geographic Anal 2010; 42(1): 32-52.

HALL, V., THOMSEN, R. W., HENRIKSEN, O. & LOHSE, N. 2011. Diabetes in Sub Saharan Africa 1999-2011: epidemiology and public health implications. A systematic review. *BMC public health,* 11**,** 564.

HANEWINCKEL, R., JONGMAN, H. P., WALLIS, L. A. & MULLIGAN, T. M. 2010. Emergency medicine in Paarl, South Africa: a cross-sectional descriptive study. *International Journal of Emergency Medicine,* 3**,** 143-150.

HIGGINS, J., THOMPSON, S. G. & SPIEGELHALTER, D. J. 2009. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172, 137-159.

HILBE, J. M. 2011. *Negative binomial regression*, Cambridge University Press.

HIPP, J.A. and CHALISE, N (2015). Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010. Prevention Chronic Disease 2015; 12:140404.

HOLLIS, S. 2015. Chen D-GD and Peace KE, Applied meta-analysis with R. SAGE Publications.

HOLTGRAVE, D. R. & CROSBY, R. 2006. Is social capital a protective factor against obesity and diabetes? Findings from an exploratory study. *Annals of epidemiology,* 16**,** 406-408.

HOSSAIN, P., KAWAR, B. & EL NAHAS, M. 2007. Obesity and diabetes in the developing world—a growing challenge. *New England journal of medicine,* 356**,** 213-215.

HUTCHINSON, M. K. & HOLTMAN, M. C. 2005. Analysis of count data using Poisson regression. Research in nursing & health, 28, 408-418.

IDF (2015). International Diabetes Federation 6th Atlas. [Online]:https://www.idf.org . [Accessed : May, 2016].

ISMAIL, N. & JEMAIN, A. A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. Casualty Actuarial Society Forum, 2007. Citeseer, 103-158.

JEMAL, A., KULLDORF, M., DEVESA, S.S., HAYNES, R.B., FRAUMENI, F (2002) A geographic analysis of prostate cancer mortality in the United States, 1970–89. International Journal for Cancer 101: 168–174.

KANMOGNE, G. D., KUATE, C. T., CYSIQUE, L. A., FONSAH, J. Y., ETA, S., DOH, R., NJAMNSHI, D. M., NCHINDAP, E., FRANKLIN, D. R. & ELLIS, R. J. 2010. HIV-associated neurocognitive disorders in sub-Saharan Africa: a pilot study in Cameroon. *BMC neurology,* 10**,** 60.

KAPLAN R.M, ATKINS. J, WILSON, D.K (1987). The Cost-utility of diet and exercise interaction in non-insulin dependent diabetes mellitus. Health Promotion International. Vol 2, Issue 4.

KAUFMAN, L. & ROUSSEEUW, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons.

KENGNE, A. P., SOBNGWI, E., ECHOUFFO-TCHEUGUI, J.-B. & MBANYA, J.-C. (2013). New insights on diabetes mellitus and obesity in Africa-Part 2: prevention, screening and economic burden. *Heart,* 99**,** 1072-1077.

KENGNE, A.P, AMOH, A.G.B., MBANYA, J.C (2005). Cardiovascular complication of diabetes mellitus in sub-Saharan Africa. Heart disease in Africa, Circulation 105.544312. vol112, issue 23.

KING, H., AUBERT, R. E. & HERMAN, W. H. 1998. Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections. Diabetes care, 21, 1414-1431.

KOVALCHIK, S. 2013. Tutorial on meta-analysis in R.

KULLFORFF, M., FEUER, E.J, MILLER, B.A., FREEDMAN, L.S. (1997) Breast cancer clusters in the northeast United States: a geographic analysis. American Journal of Epidemiology 146: 161–170.

KUTNER, M. H., NACHTSHEIM, C. & NETER, J. 2004. *Applied linear regression models*, McGraw-Hill/Irwin.

LACHIN, J. M. 2011. Analysis of count data. *Biostatistical Methods: The Assessment of Relative Risks, Second Edition*, 381-427.

LAWSON, A. B. 2013. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*, CRC press.

LAWSON, A. B., BROWNE, W. J. & RODEIRO, C. L. V. 2003. *Disease mapping with WinBUGS and MLwiN*, John Wiley & Sons.

LEANDRO, G. 2008. Meta-analysis in Medical Research: The handbook for the understanding and practice of meta-analysis, John Wiley & Sons.

LESAFFRE, E. & LAWSON, A. B. 2012. *Bayesian biostatistics*, John Wiley & Sons.

LEVINE, J. A. 2011. Poverty and obesity in the US. American Diabetes Association.

LEVITT, N.S. (2008) Diabetes in Africa: epidemiology, management and health care challenges. BMJ 10.1136. 14730. Volume 94, issue11.

LEVITT, A.M., BINDER, S., SACKS, J.J., HUGHES, J.M., (1999). EMERGING INFECTIOUS DISEASE: PUBLIC HEALTH ISSUES FOR THE 21ST CENTURY. Science Journal, Vol 284, Issue 5418, pp1311-1313.

LEUNG, Y., MEI, C.L., AND ZANG, W.X (2000). Statistical tests for spatial non-stationarity based on the Geographically weighted regression model. Journal of environmental and planning 32(1): 9-32.

LIM, S. S., VOS, T., FLAXMAN, A. D., DANAEI, G., SHIBUYA, K., ADAIR-ROHANI, H., ALMAZROA, M. A., AMANN, M., ANDERSON, H. R. & ANDREWS, K. G. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet,* 380**,** 2224-2260.

LINDÉN, A. & MÄNTYNIEMI, S. 2011. Using the negative binomial distribution to model over dispersion in ecological count data. Ecology, 92, 1414-1421.

LITTELL, R. C. 1996. SAS, Wiley Online Library.

LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D. & SCHABENBERGER, O. 2007. SAS for mixed models, SAS institute.

LIU, Y., JIANG, S., LIU, Y, WANG, R., LI, Y., YUAN, Z., WANG, L., AND XUE, F (2011). Spatial epidemiology and spatial ecology study of worldwide drug-resistance tuberculosis. International journal of Health Geographic 10:50.

LUNN, D. J., THOMAS, A., BEST, N. & SPIEGELHALTER, D. 2000. Win BUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing, 10, 325-337.

LYKOU, A. & NTZOUFRAS, I. 2011. Win BUGS: a tutorial. Wiley Interdisciplinary Reviews: Computational Statistics, 3, 385-396.

MBANYA, J. C. N., MOTALA, A. A., SOBNGWI, E., ASSAH, F. K. & ENORU, S. T. 2010. Diabetes in sub-saharan africa. *The lancet,* 375**,** 2254-2266.

MCCULLAGH, P. & NELDER, J. 1989. Generalised linear modelling. *Chapman and Hall: New York*.

MICHIMI, A., WIMBERLY, M.C. (2010). Spatial patterns of obesity and associated risk factors in the conterminous U.S. American Journal of Preventive Medicine; 39: 1-12.

MILLAR, R. B. 2009. Comparison of hierarchical Bayesian models for over dispersed count data using DIC and Bayes' factors. Biometrics, 65, 962-969.

MONTGOMERY, D. C., PECK, E. A. & VINING, G. G. 2015. *Introduction to linear regression analysis*, John Wiley & Sons, New York, United State.

MOTALA, A. A. 2002. Diabetes trends in Africa. *Diabetes/metabolism research and reviews,* 18**,** S14-S20.

MOTALA, A. A., OMAR, M. A. & PIRIE, F. J. 2003. Epidemiology of type 1 and type 2 diabetes in Africa. *European Journal of Cardiovascular Risk,* 10**,** 77-83.

NAMOSHA, E., SARTORIUS, B., TANSER, F. (2013). Spatial Clustering of All-Cause and HIV-Related Mortality in a Rural South African Population (2000–2006). PLoS ONE 8(7): e69279.

NELDER, J. A. & BAKER, R. J. 1972. *Generalized linear models*, Wiley Online Library.

NORMAND, S.-L. T. 1999. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. Statistics in medicine, 18, 321-359.

NTZOUFRAS, I. 2011. *Bayesian modeling using WinBUGS*, John Wiley & Sons, New York, United States.

OGURTSOVA, K., DA ROCHA FERNANDES, J., HUANG, Y., LINNENKAMP, U., GUARIGUATA, L., CHO, N., CAVAN, D., SHAW, J. & MAKAROFF, L. 2017. IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice,* 128**,** 40-50.

OLSEN, J., SARACCI, R. & TRICHOPOULOS, D. 2010. *Teaching epidemiology: A guide for teachers in epidemiology, public health and clinical medicine*, OUP Oxford.

OMAR, M., SEEDAT, M., MOTALA, A., DYER, R. & BECKER, P. 1993. The prevalence of diabetes mellitus and impaired glucose tolerance in a group of urban South African blacks. *South African medical Journal* 83**,** 641-643.

ORGANIZATION, 2000. *The world health report 2000: health systems: improving performance*, World Health Organization.

OSEI, F.B., DUKER, A., AND STEIN, A. (2012). Evaluating Spatial and Space-Time Clustering of Cholera in Ashanti-Region-Ghana. Available from: http://www.intechopen.com/books/cholera/evaluating-spatial-and-space-time-clustering-ofcholera-in-ashanti-region-ghana.

PAN, W. 2001. Akaike's information criterion in generalized estimating equations. Biometrics, 57, 120-125.

PARNELL, S. & PIETERSE, E. A. 2014. Africa's urban revolution.

PEER, N., KENGNE, A.-P., MOTALA, A. A. & MBANYA, J. C. 2014. Diabetes in the Africa Region: an update. *Diabetes research and clinical practice,* 103**,** 197-205.

PFEIFFER, D., ROBINSON, T. P., STEVENSON, M., STEVENS, K. B., ROGERS, D. J. & CLEMENTS, A. C. 2008. Spatial analysis in epidemiology. Oxford University Press, New York.

POCH, M. & MANNERING, F. 1996. Negative binomial analysis of intersection-accident frequencies. Journal of transportation engineering, 122, 105-113.

PRESTON, D. L. 2005. Poisson regression in epidemiology. *Encyclopedia of biostatistics*.

RAMACHANDRAN, A., MARY, S., YAMUNA, A., MURUGESAN, N. & SNEHALATHA, C. 2008. High prevalence of diabetes and cardiovascular risk factors associated with urbanization in India. American Diabetes Association: *Diabetes care,* 31**,** 893-898.

RAMACHANDRAN, A., SNEHALATHA, C., LATHA, E., MANOHARAN, M. & VIJAY, V. 1999. Impacts of urbanisation on the lifestyle and on the prevalence of diabetes in native Asian Indian population. *Diabetes research and clinical practice,* 44**,** 207-213.

RAMACHANDRAN, A., SNEHALATHA, C., VIJAY, V. & KING, H. 2002. Impact of poverty on the prevalence of diabetes and its complications in urban southern India. Diabetic Medicine, 19, 130-135.

RENARD, D. 2011. Roger S. Bivand, Edzer J. Pebesma, Virgilio Gomez-Rubio: Applied Spatial Data Analysis with R. *Mathematical Geosciences,* 43**,** 607-609.

RISTE, L., KHAN, F. & CRUICKSHANK, K. 2001. High prevalence of type 2 diabetes in all ethnic groups, including Europeans, in a British inner city. American Diabetes Association: Diabetes Care, 24, 1377-1383.

ROGLIC, G. & UNWIN, N. 2010. Mortality attributable to diabetes: estimates for the year 2010. *Diabetes research and clinical practice,* 87**,** 15-19.

ROGLIC, G., UNWIN, N., BENNETT, P. H., MATHERS, C., TUOMILEHTO, J., NAG, S., CONNOLLY, V. & KING, H. 2005. The Burden of Mortality Attributable to Diabetes Realistic estimates for the year 2000. American Diabetes Assocation: *Diabetes care,* 28**,** 2130-2135.

ROMESBURG, C. 2004. *Cluster analysis for researchers*, Lulu. com.

ROSENBERG, M.S., SOKAL R.R., ODEN, N.L., DIGIOVANN, D. (1999). Spatial autocorrelation of cancer in Western Europe. European Journal of Epidemiology 15: 15–22.

SAKAMOTO, Y., ISHIGURO, M. & KITAGAWA, G. 1986. Akaike information criterion statistics. Journal of the American Statistical Association, Vol 83, Issue 403.

SANKOH OA, YAZOUME Y, SAUEBORN R, MULLER O, BECHER H (2001) Clustering of childhood mortality in rural Burkina Faso. International Journal of Epidemiology 30: 485–492.

SCHWARZ, G. 1978. Estimating the dimension of a model. The annals of statistics, 6, 461-464.

SCHWARZER, G. 2007. Meta: An R package for meta-analysis. R news, 7, 40-45.

SCHWITZGEBEL, V. M. 2014. Many faces of monogenic diabetes. *Journal of diabetes investigation,* 5**,** 121-133.

SEBER, G. A. & LEE, A. J. 2012. *Linear regression analysis*, John Wiley & Sons, New York, United State.

SHAW, J. E., SICREE, R. A. & ZIMMET, P. Z. 2010. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice,* 87**,** 4-14.

SHRESTHA, S.S., KIRTLAND, K.A., THOMPSON, T.J., BARKER, L., GREGG, E.W., AND GEISS, L. (2012). Spatial Clusters of County-level Diagnosed Diabetes and Associated Risk Factors in the United States. Open Diabetes Journal. 5, 29-37.

SCHMIEDEL, S., GEOFFREY, M.J., BLETTNER, M., SCHUZ, J (2011). Spatial clustering of leukaemia and type 1 diabetes in children in Denmark. Cancer Causes Control (2011) 22:849–857.

SCHUURMAN, N., PETERS, P.A., OLIVER, L.N (2009). Are obesity and physical activity clustered? A spatial analysis linked to residential density. Obesity (Silver Spring); 17(12): 2202-09.

SNIJDERS, T. A. & BOSKER, R. J. 2011. Multilevel analysis: An introduction to basic and advanced multilevel modelling. International Encyclopaedia of Statistical Science. Springer International Publishing, Berlin, Heidelberg.

SOBNGWI, E., MAUVAIS-JARVIS, F., VEXIAU, P., MBANYA, J. & GAUTIER, J. 2001. Diabetes in Africans. Part 1: epidemiology and clinical specificities. *Diabetes & metabolism,* 27**,** 628-634.

SOBNGWI, E., NDOUR-MBAYE, M., BOATENG, K. A., RAMAIYA, K. L., NJENGA, E. W., DIOP, S. N., MBANYA, J.-C. & OHWOVORIOLE, A. E. 2012. Type 2 diabetes control and complications in specialised diabetes care centres of six sub-Saharan African countries: the Diabcare Africa study. *Diabetes research and clinical practice,* 95**,** 30-36.

STATSILK- clearing house of links to shapefile maintaining pages. (Online) http://www.statsilk.com/maps/download-free-shapefile-maps. [Accessed: May 13, 2016].

TANGO, T. 2010. Statistical methods for disease clustering. Springer Science & Business Media. Berlin, Heidelberg.

TANSER, F., BARNIGHAUSEN, T., COOKE, G.S, NEWELL, M.L (2009). Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International Journal of Epidemiology 38: 1008–1016, doi:10.1093/ ije/dyp148

TIAO, G. C. & BOX, G. E. 1973. Some comments on "Bayes" estimators. *The American Statistician,* 27**,** 12-14.

TISHKOFF, S. A. & WILLIAMS, S. M. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nature - Reviews Genetics,* 3**,** 611-621.

TOURAY, K., ADETIFA, M.I., JALLOW, A., RIGBY, J., JEFFRIES, D., CHEUNG, Y. B., DONKOR, S., ADEGBOLA, R. A., and HILL, P. C. (2010). Spatial analysis of tuberculosis in an Urban West African setting: is there evidence of clustering? Journal of Tropical Medicine and International Health doi:10.1111/j.1365-3156.2010. 02533.x volume 15 no 6 pp 664–672.

TUOMI, T., SANTORO, N., CAPRIO, S., CAI, M., WENG, J. & GROOP, L. 2014. The many faces of diabetes: a disease with increasing heterogeneity. *The Lancet,* 383**,** 1084-1094.

UNITED NATION (UN). (Online from: www.https//unitednation.org) [Accessed : July 20,2016]

VAN HOUWELINGEN, H. C., ARENDS, L. R. & STIJNEN, T. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in medicine, 21, 589-624.

VANDENHEEDE, H., DEBOOSERE, P., STIRBU, I., AGYEMANG, C. O., HARDING, S., JUEL, K., RAFNSSON, S. B., REGIDOR, E., REY, G. & ROSATO, M. 2012. Migrant mortality from diabetes mellitus across Europe: the importance of socio-economic change. *European journal of epidemiology,* 27**,** 109-117.

VER HOEF, J. M. & BOVENG, P. L. 2007. Quasi-Poisson versus Negative Binomial regression: How should we model over dispersed Count Data? Journal of Ecology, 88, 2766-2772.

WALLER, L. A. & GOTWAY, C. A. 2004. *Applied spatial statistics for public health data*, John Wiley & Sons, New York, United State.

WANG, H., LIDDELL, C. A., COATES, M. M., MOONEY, M. D., LEVITZ, C. E., SCHUMACHER, A. E., APFEL, H., IANNARONE, M., PHILLIPS, B. & LOFGREN, K. T. 2014. Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet,* 384**,** 957-979.

WHEELER, D AND TIEFELSDORF, M (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical systems 7(2): 161-187.

WHITE, G. C. & BENNETTS, R. E. 1996. Analysis of frequency count data using the negative binomial distribution. Journal of Ecology, 77, 2549-2557.

WHITING, D. R., GUARIGUATA, L., WEIL, C. & SHAW, J. 2011. IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. Journal of the Intenational Diabetes Federation. *Diabetes research and clinical practice,* 94**,** 311-321.

WHITING, D.R., GAURIGUATA, L (2011). Global estimates of diabetes prevalence for 2013 and projection for 2035. Diabetes research and clinical practice. Journal of the international Diabetes Federation.

WILD, S., ROGLIC, G., GREEN, A., SICREE, R. & KING, H. 2004. Global prevalence of diabetes estimates for the year 2000 and projections for 2030. *Diabetes care,* 27**,** 1047-1053.

WILLIAMS, E. J. 1959. *Regression Analysis*, John Wiley & Sons, New York, United.

WORLD BANK DATA (2015). World Bank Data. (Online : www.databank.worldbank,org) . [Accessed , March 22, 2016]

XIANG, K. & SONG, D. 2016. Spatial Analysis of China Province-Level Perinatal Mortality. *Iranian Journal of Public Health,* 45**,** 614.

YAZDY, M. M., WERLER, M. M., ANDERKA, M., LANGLOIS, P. H. & VIEIRA, V. M. 2015. Spatial analysis of gastroschisis in Massachusetts and Texas. *Annals of epidemiology,* 25**,** 7-14.

ZACKS, S. 1983. Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. Recent advances in statistics, 25, 245-269.

ZEILEIS, A., KLEIBER, C. & JACKMAN, S. 2008. Regression models for count data in R. Journal of statistical software, 27, 1-25.

ZOU, G. 2004. A modified Poisson regression approach to prospective studies with binary data. *American journal of epidemiology,* 159**,** 702-706.

Appendix A: Data Indicators, Sources, and Definitions

	INDICATORS	DEFINITION	SOURCES
1	DIABETES PREVALENCE (% OF POPULATION AGES 20 TO 69)	Diabetes prevalence refers to the percentage of people ages 20-79 that have type 1 or type 2 diabetes.	World Bank national accounts data, and OECD National Accounts data files.
2	GDP per capita (current US$)	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.	World Bank national accounts data, and OECD National Accounts data files.
3	HEALTH EXPENDITURE	Total health expenditure is the sum of public and private health expenditure. It covers the provision of health services (preventive and curative), family planning activities, nutrition activities, and emergency	World Health Organization Global Health Expenditure database (see http://apps.who.int/nha/datab

		aid designated for health but does not include provision of water and sanitation.	ase for the most recent updates).
4	Urban population (% of total)	Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.	United Nations, World Urbanization Prospects.
5	Literacy rate, adult total (% of people ages 15 and above)	Adult literacy rate is the percentage of people ages 15 and above whom can both read and write with understanding a short simple statement about their everyday life. .	United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics.
6	Daily smoking of any tobacco product (age-standardized rate)	Prevalence estimates for daily smoking of any tobacco product are age-standardized rates for adults aged 15 years and over, estimated using the method described in the Method of Estimation below. "Tobacco smoking" includes cigarettes, cigars, pipes or any other smoked tobacco products. The definition of "daily tobacco	WHO http://apps.who.int/gho/data/view.main.TOB30011

		smoking" varies between surveys, but often means current smoking of any tobacco product at least once a day.	
7	Smoking prevalence, females (% of adults)	Prevalence of smoking, female is the percentage of women ages 15 and over who smoke any form of tobacco, including cigarettes, cigars, pipes or any other smoked tobacco products. Data include daily and non-daily or occasional smoking.	World Health Organization, Global Health Observatory Data Repository (http://apps.who.int/ghodata/).
8	Smoking prevalence, males (% of adults)	Prevalence of smoking, male is the percentage of men ages 15 and over who smoke any form of tobacco, including cigarettes, cigars, pipes or any other smoked tobacco products. Data include daily and non-daily or occasional smoking.	. World Health Organization, Global Health Observatory Data Repository (http://apps.who.int/ghodata/).
9	Population ages 15-64 (% of total)	Total population between the ages 15 to 64 is the number of people who could potentially be economically active. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship--except for refugees	World Bank staff estimates based on age distributions of United Nations Population Division's World Population Prospects

		not permanently settled in the country of asylum, which are generally considered part of the population of the country of origin.	
10	Prevalence of insufficient physical activity among adults aged 18+ years	Percent of defined population attaining less than 150 minutes of moderate-intensity physical activity per week, or less than 75 minutes of vigorous-intensity physical activity per week, or equivalent.	WHO: http://apps.who.int/gho/data/view.main.2463
11	Recorded alcohol per capita (15+ years) consumption of pure alcohol Recorded APC	Recorded APC is defined as the recorded amount of alcohol consumed per capita (15+ years) over a calendar year in a country, in liters of pure alcohol. The indicator only considers the consumption which is recorded from production, import, export, and sales data often via taxation. Numerator: The amount of recorded alcohol consumed per capita (15+ years) during a calendar year, in liters of pure alcohol. Denominator: Midyear resident population (15+ years) for the same calendar year, UN World Population Prospects, medium variant.	WHO http://apps.who.int/gho/data/view.main.52160 Global Information System on Alcohol and Health (GISAH

12	OBESITY: Prevalence of obesity, BMI ≥ 30	Percentage of defined population with a body mass index (BMI) of 30 kg/m2 or higher.	WHO: http://apps.who.int/gho/data/view.main.2450A
13-15	Development index		
	• Human Development Index (HDI)	• A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. See Technical note 1 at http://hdr.undp.org/en for details on how the HDI is calculated.	• HDRO calculations based on data from UNDESA (2015), UNESCO Institute for Statistics (2015), United Nations Statistics Division (2015), World Bank (2015a), Barro and Lee (2014) and IMF (2015).
		• Average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level.	

• Mean year of Schooling	Aggregate income of an economy generated by its production and its ownership of factors of production, less the incomes paid for the use of factors of production owned by the rest of the world, converted to international dollars using PPP rates, divided by midyear population.	• UNESCO Institute for Statistics (2015), Barro and Lee (2014), UNICEF Multiple Indicator Cluster Surveys and ICF Macro Demographic and Health Surveys.
Gross national income (GNI) per capita:		World Bank (2015a), IMF (2015) and United Nations Statistics Division (2015)

Density of physicians (per 1 000 population)	Number of medical doctors (physicians), including generalist and specialist medical practitioners, per 1 000 population.	WHO http://apps.who.int/gho/data/view.main.92100
Physician density per 1000 population	Physicians include generalist and specialist medical practitioners.	World Health Organization's Global Health Workforce Statistics, OECD, supplemented by country data. WHO compiles data on health workforce from four major sources: population censuses, labour force and employment surveys, health facility assessments and routine administrative information systems (including reports on public expenditure, staffing, and payroll as well as professional training, registration and licensure)? Most of the data from administrative sources are derived from published national health sector reviews and/or official country reports to WHO offices. In general, the denominator data for physicians' density (i.e. national

			population estimates) are obtained from the United Nations Population Division's World Population Prospects database.
16	Per capita government expenditure on health at average exchange rate (US$) i.e. Health expenditure per capita, all countries, selected years Estimates by country	Per capita, general government expenditure on health (GGHE) expressed at average exchange rate for that year in US dollar. Current prices.	WHO http://apps.who.int/gho/data/view.main.HEALTHEXPCAPLATESTv
17	Age dependency ratio, old (% of working-age population)	Age dependency ratio, old, is the ratio of older dependents--people older than 64--to the working-age population--those ages 15-64. Data are shown as the proportion of dependents per 100 working-age population.	World Bank staff estimates using the World Bank's population and age distributions of the United Nations Population Division's World Population Prospects. The World Bank's population estimates are from various sources including the United

Nations Population Division's World Population Prospects; census reports and statistical publications from national statistical offices; Eurostat's Demographic Statistics; United Nations Statistical Division, Population and Vital Statistics Report (various years); U.S. Census Bureau: International Database; and Secretariat of the Pacific Community, Statistics and Demography Programme.

Appendix B: Data Sources

I. Prevalence of Diabetes, this information was obtained at:

http://data.worldbank.org/indicator/SH.STA.DIAB.ZS/countries?display=defaulthttp3A2F2Fdata.worldbank.org.cn2Findicator2FNY.GDP.PCAP.CD2Fcountries2F1W3Fdisplay3Ddefault . For all the countries.

2. The data on percentage of country GDP for all the countries was found at http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

3. The data on percentage of unemployment for all the countries was found at http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

And

http://unstats.un.org/unsd/demographic/products/socind/

3. The data on percentage of population ages 15 above was found at http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

4. The data on percentage pf physical activity for the countries was found at: http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

5. The data on percentage of population living in poverty for all the countries was found at: http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#
6. The data on percentage of country health expenditure was found at http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

7. The data on percentage of nurses and mid wives for the countries except ⋯⋯.was found at: http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#
8. The data on percentage of physicians per head for the countries except ⋯⋯was found at:

The data on percentage of population living in poverty was found at http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SH.STA.DIAB.ZS#

9. The data on Prevalence of insufficient physical activity among adult by country was found at: http://apps.who.int/gho/data/node.main.A893?lang=en and

http://gamapserver.who.int/gho/interactive_charts/ncd/risk_factors/physical_inactivity/atlas.html?indicator=i1&date=Male

10. The data on Percentage of population living under 1.90 and 3.10 a day 2011 dollars (Purchasing power parity) was found at https://en.wikipedia.org/wiki/List_of_countries_by_percentage_of_population_living_in_poverty
11. The data on Tertiary school gross enrolment was found at: http://unstats.un.org/unsd/demographic/products/socind/

12. Resources

Source of Information (Data Source):

http://data.worldbank.org/indicator/SH.STA.DIAB.ZS/countries?display=defaulthttp3A2F2Fdata.worldbank.org.cn2Findicator2FNY.GDP.PCAP.CD2Fcountries2F1W3Fdisplay3Ddefault

http://www.un.org/en/databases/

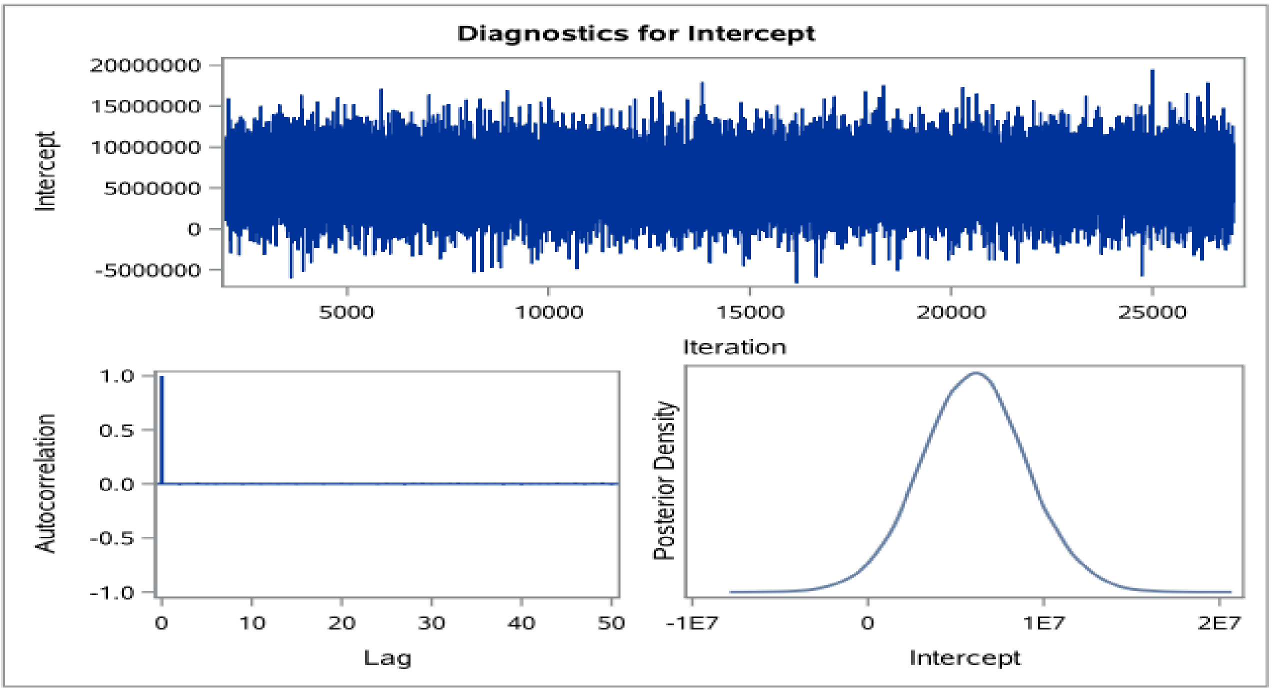
13. Global and regional trends by UN Regions, 1990-2025

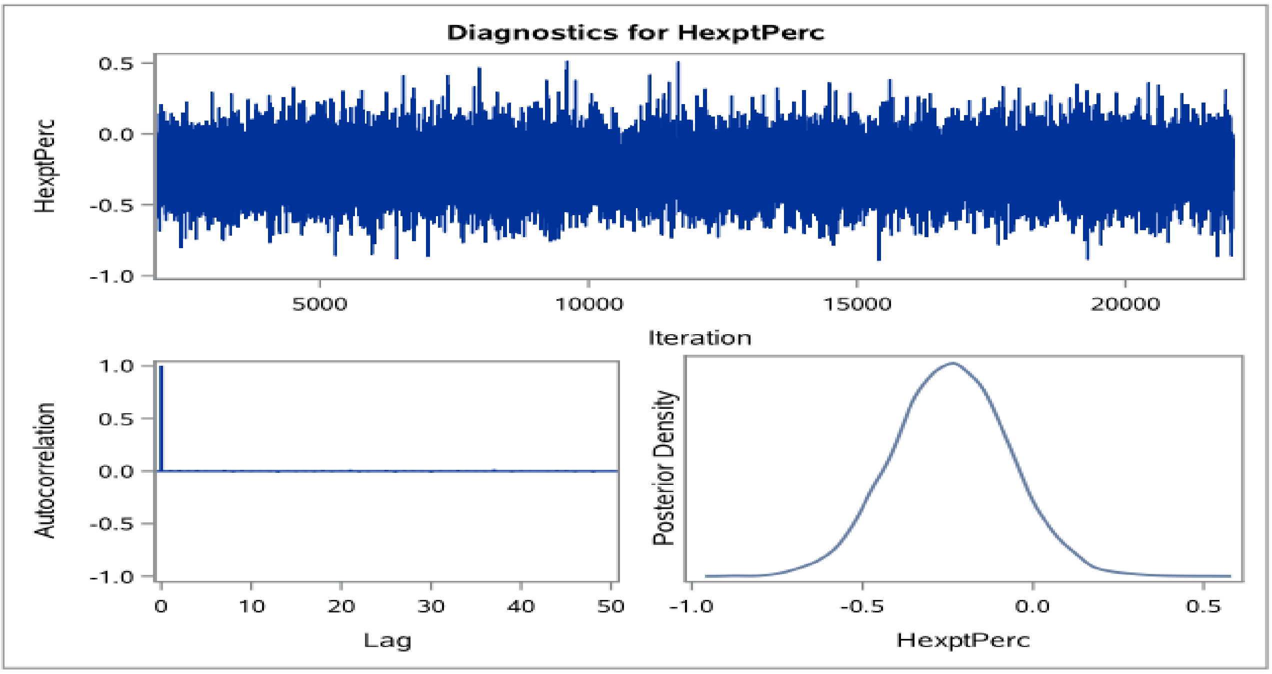
Overweight: 1990-2015: Overweight prevalence http://apps.who.int/gho/data/view.main.NUTUNOVERWEIGHTv

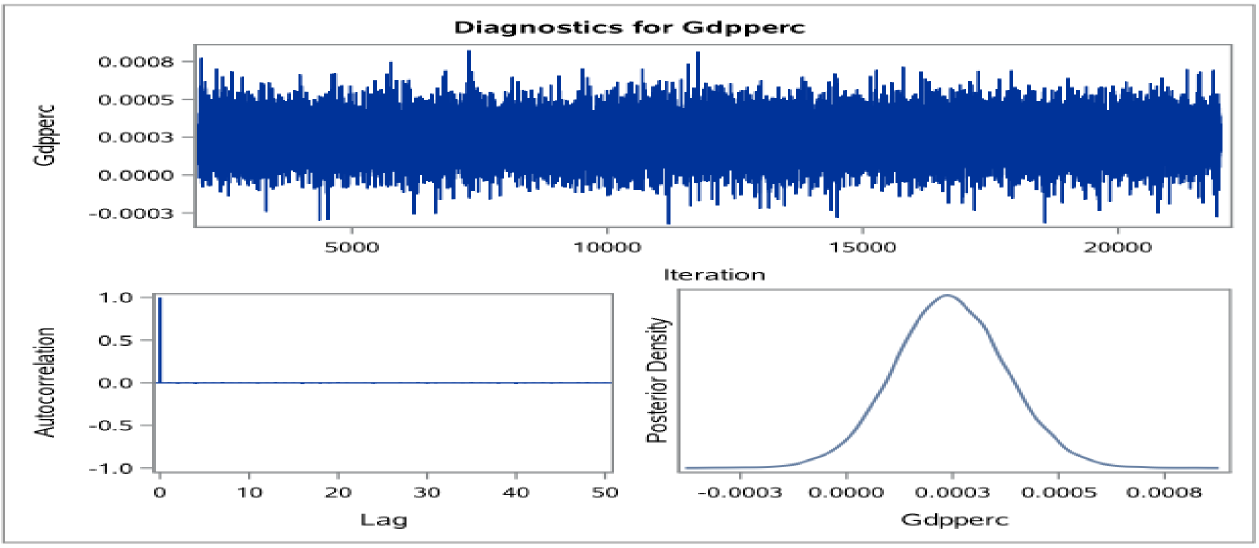
Urban population: http://apps.who.int/gho/data/view.main.100015

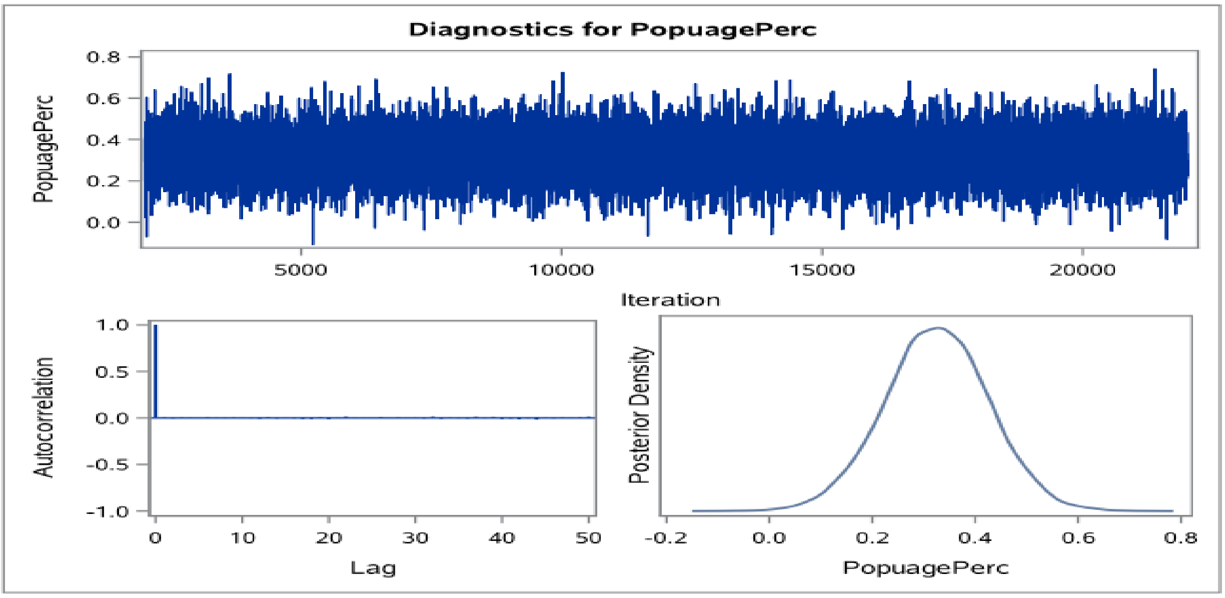
http://gamapserver.who.int/gho/interactive_charts/ncd/risk_factors/physical_inactivity/atlas.html?indicator=i1&date=Male

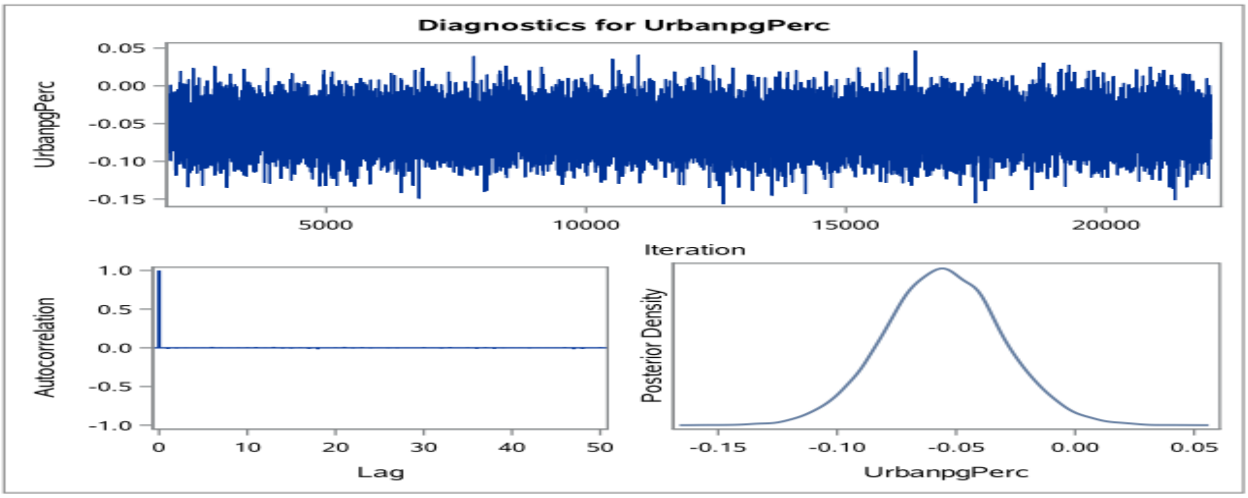
Appendix C: MCMC Plots of Bayesian Linear Regression

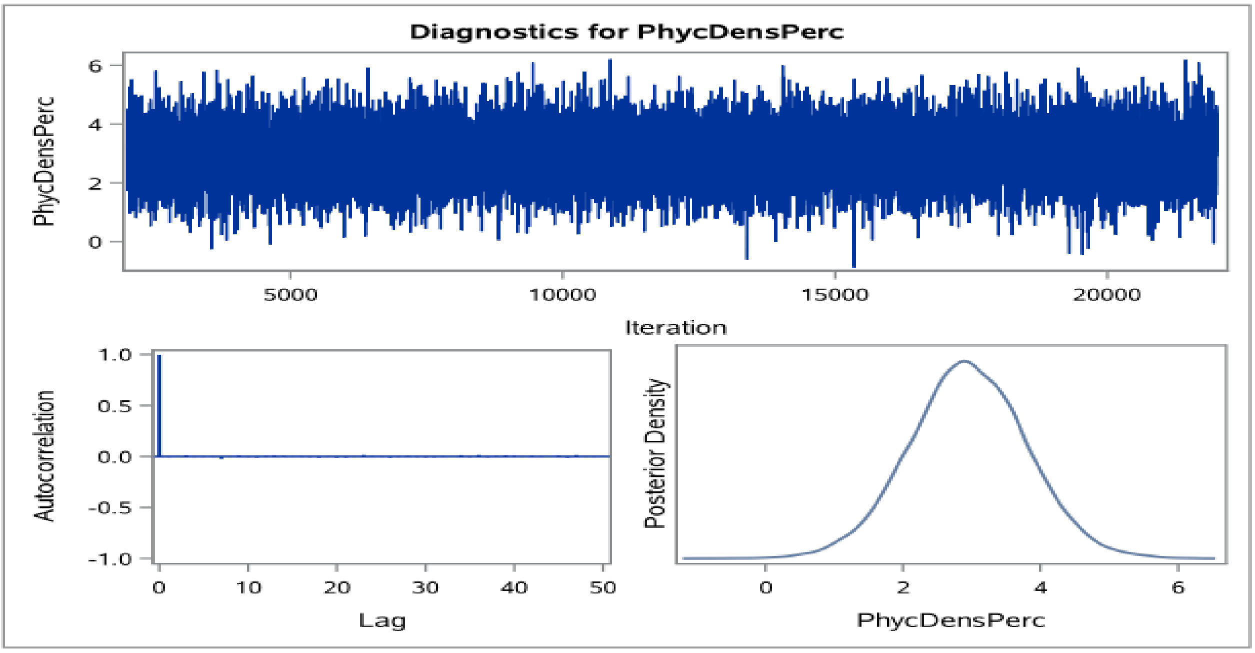


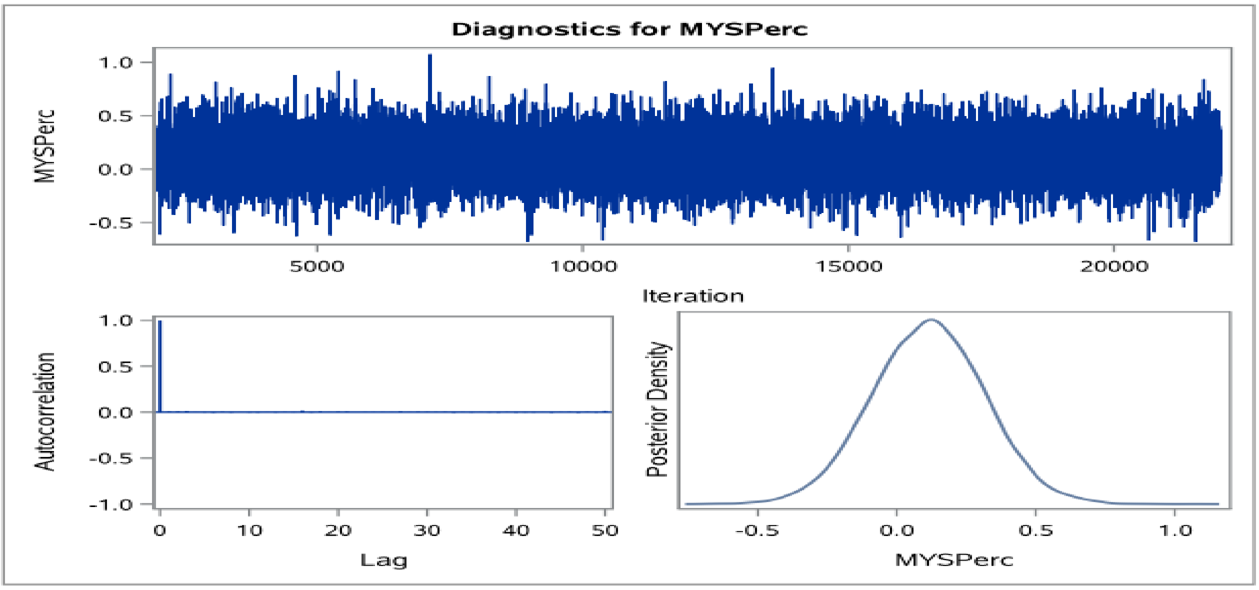


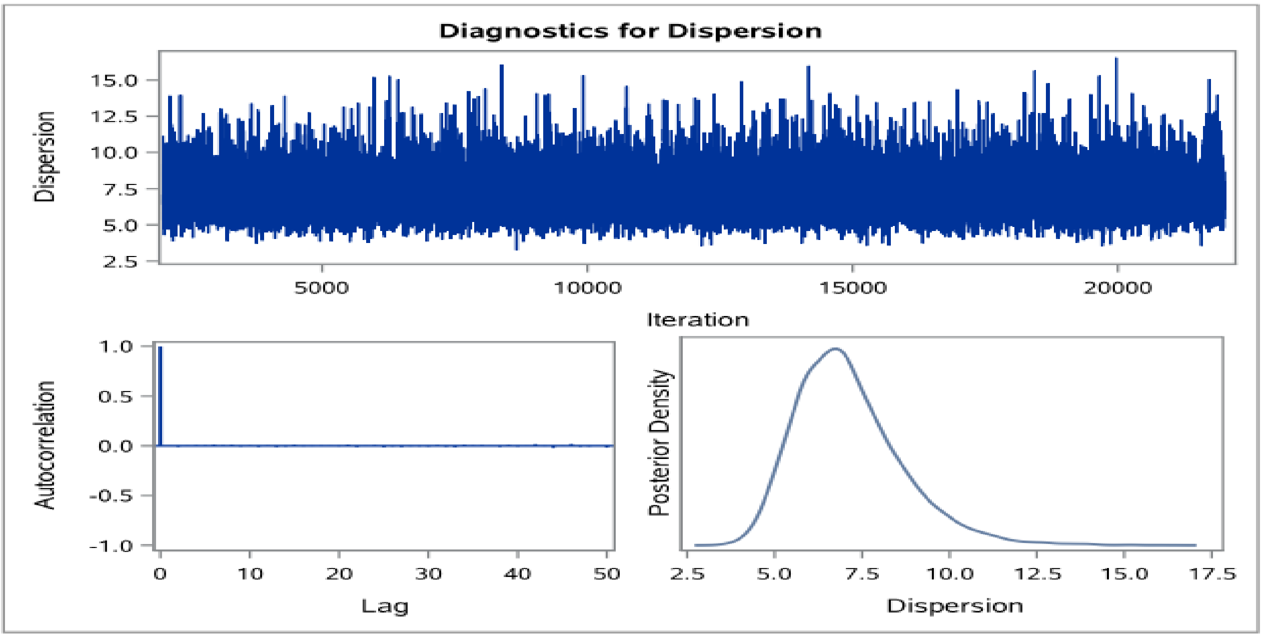

Diagnostics for Gdpperc
Gdpperc
0.0008
0.0005
0.0003
0.0000
-0.0003
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.0003
0.0000
0.0003
0.0005
0.0008
Gdpperc


Diagnostics for PopuagePerc
PopuagePerc
0.8
0.6
0.4
0.2
0.0
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.2
0.0
0.2
0.4
0.6
0.8
PopuagePerc

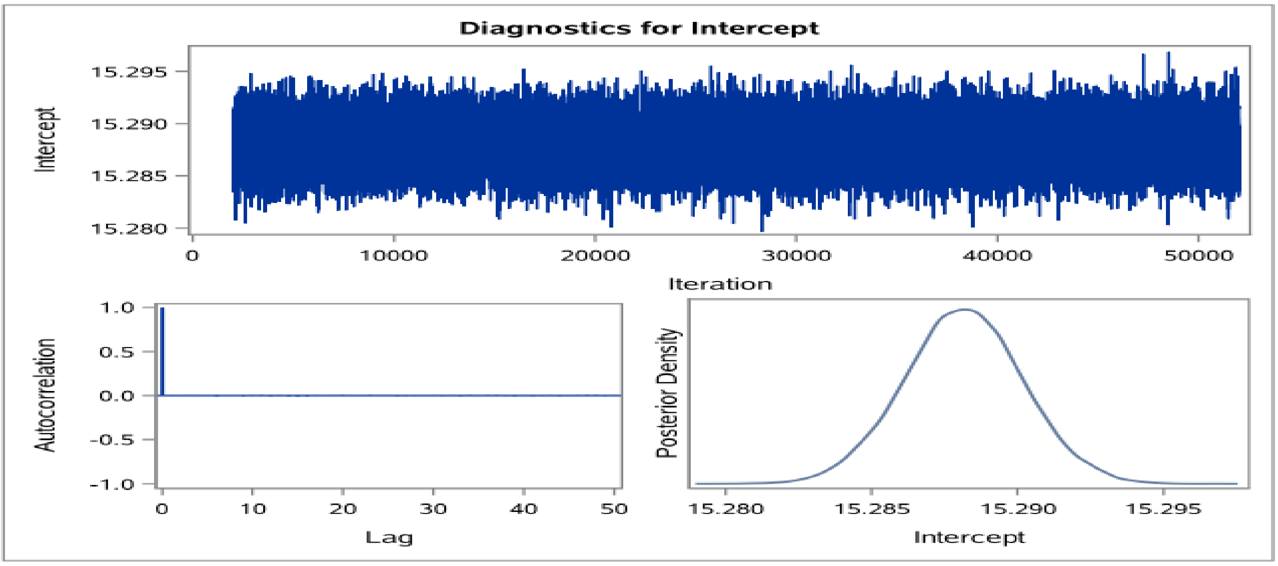

Diagnostics for UrbanpgPerc
UrbanpgPerc
0.05
0.00
-0.05
-0.10
-0.15
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.15
-0.10
-0.05
0.00
0.05
UrbanpgPerc

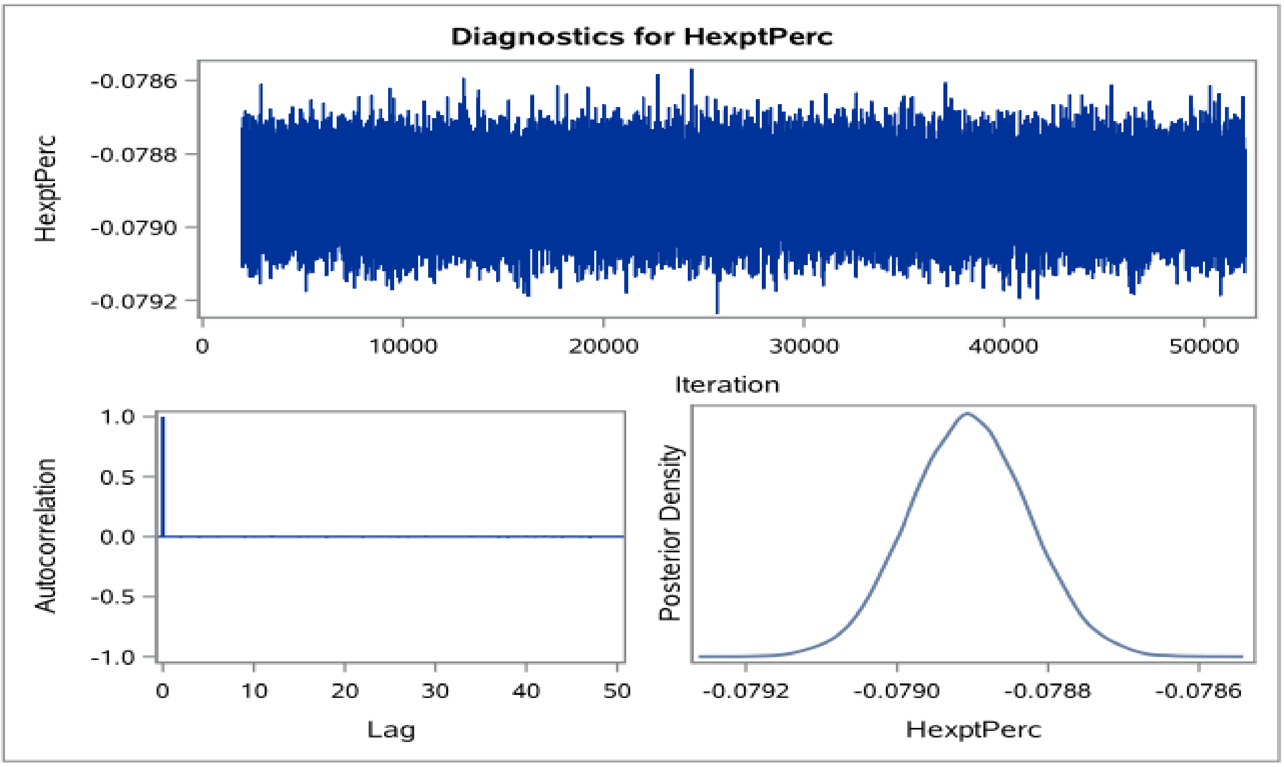

Diagnostics for PhycDensPerc
PhycDensPerc
6
4
2
0
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0
2
4
6
PhycDensPerc

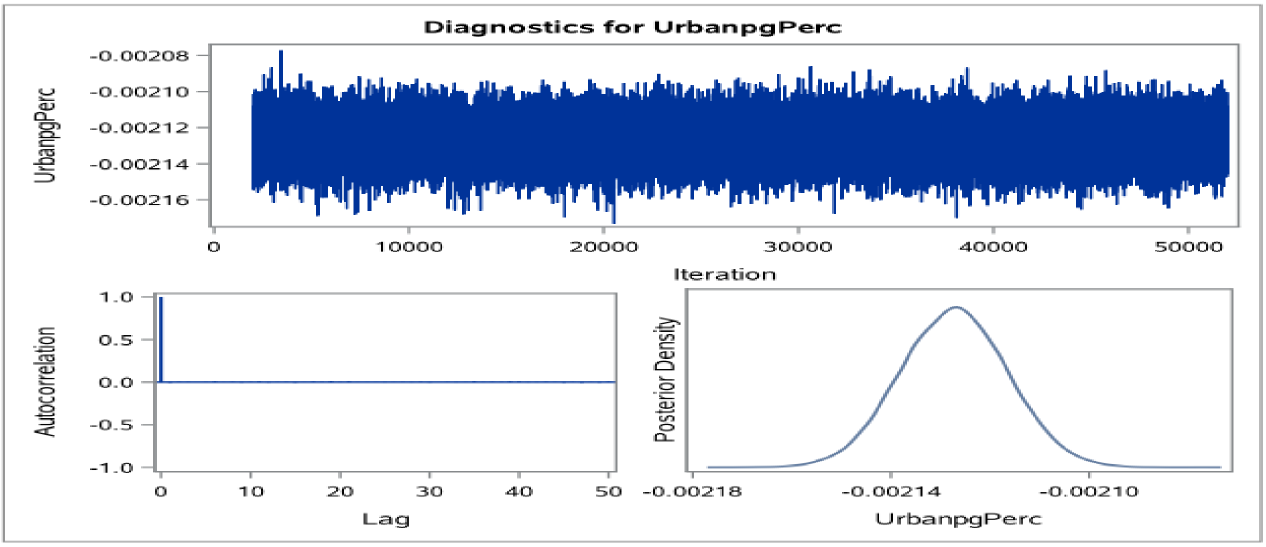

Diagnostics for MYSPerc
MYSPerc
1.0
0.5
0.0
-0.5
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.5
0.0
0.5
1.0
MYSPerc

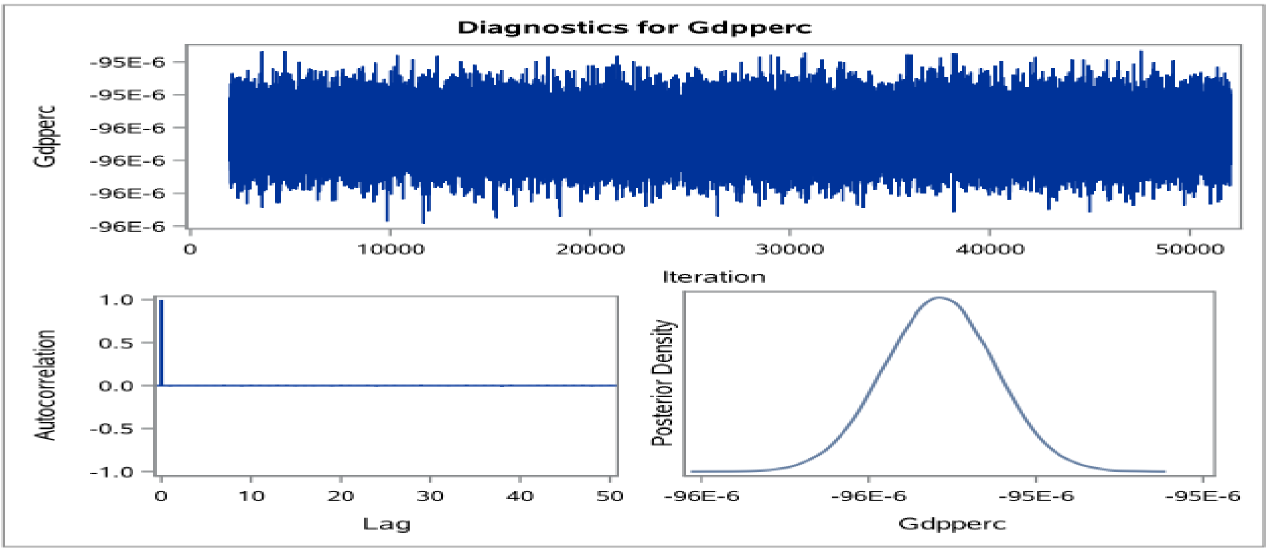

Diagnostics for Dispersion
Dispersion
15.0
12.5
10.0
7.5
5.0
2.5
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Dispersion

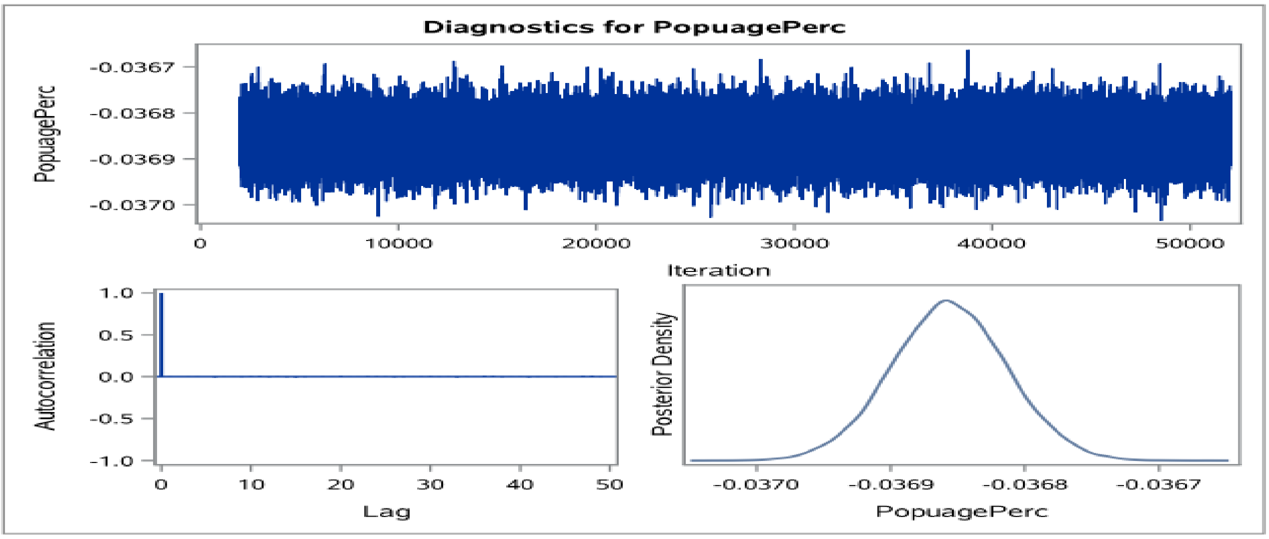
Appendix D: MCMC Plots for Bayesian Poisson Regression

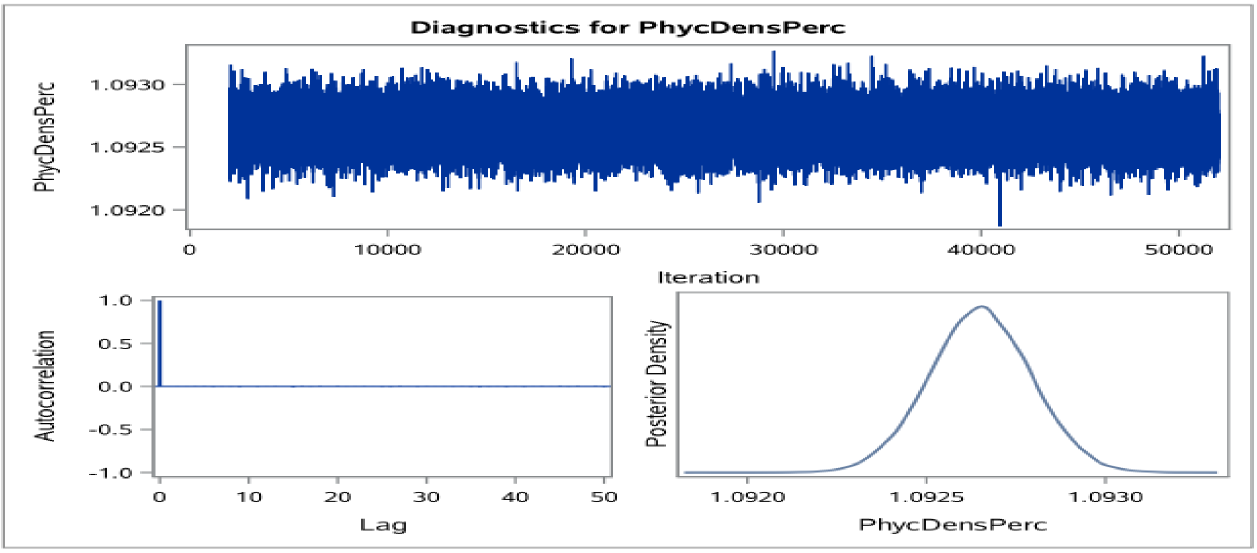


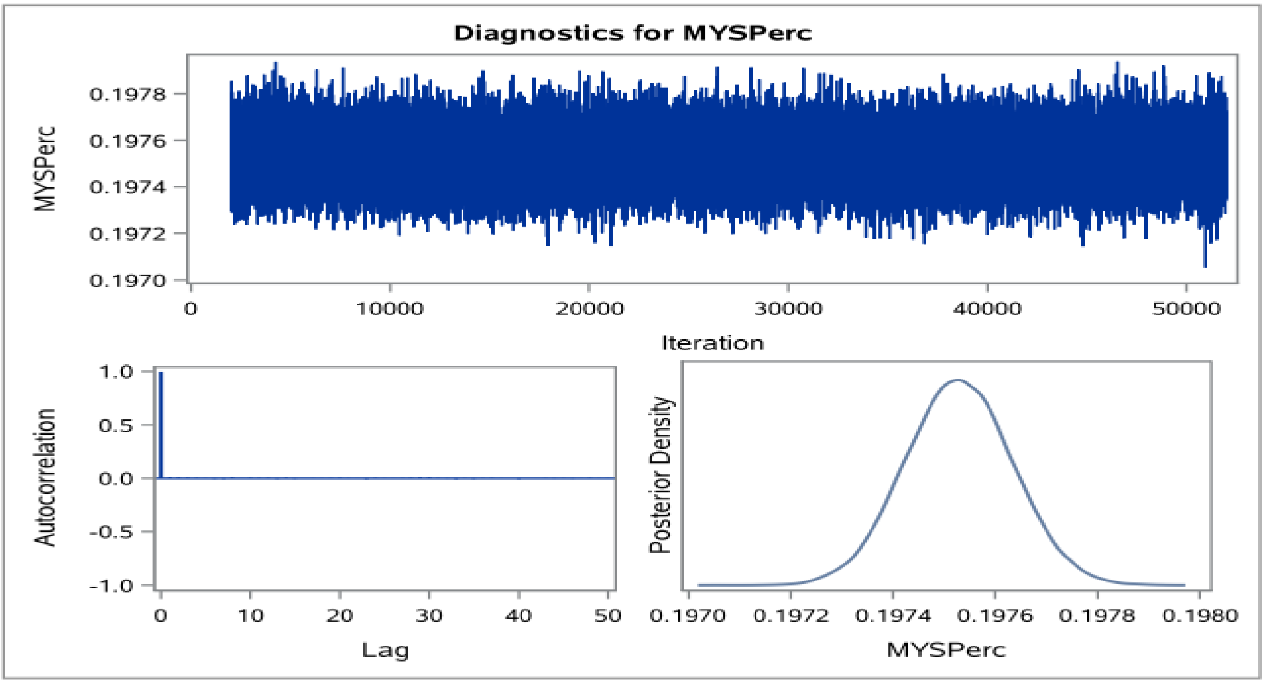

Diagnostics for HexptPerc
HexptPerc
-0.0786
-0.0788
-0.0790
-0.0792
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.0792
-0.0790
-0.0788
-0.0786
HexptPerc


Diagnostics for UrbanpgPerc
UrbanpgPerc
-0.00208
-0.00210
-0.00212
-0.00214
-0.00216
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.00218
-0.00214
-0.00210
UrbanpgPerc

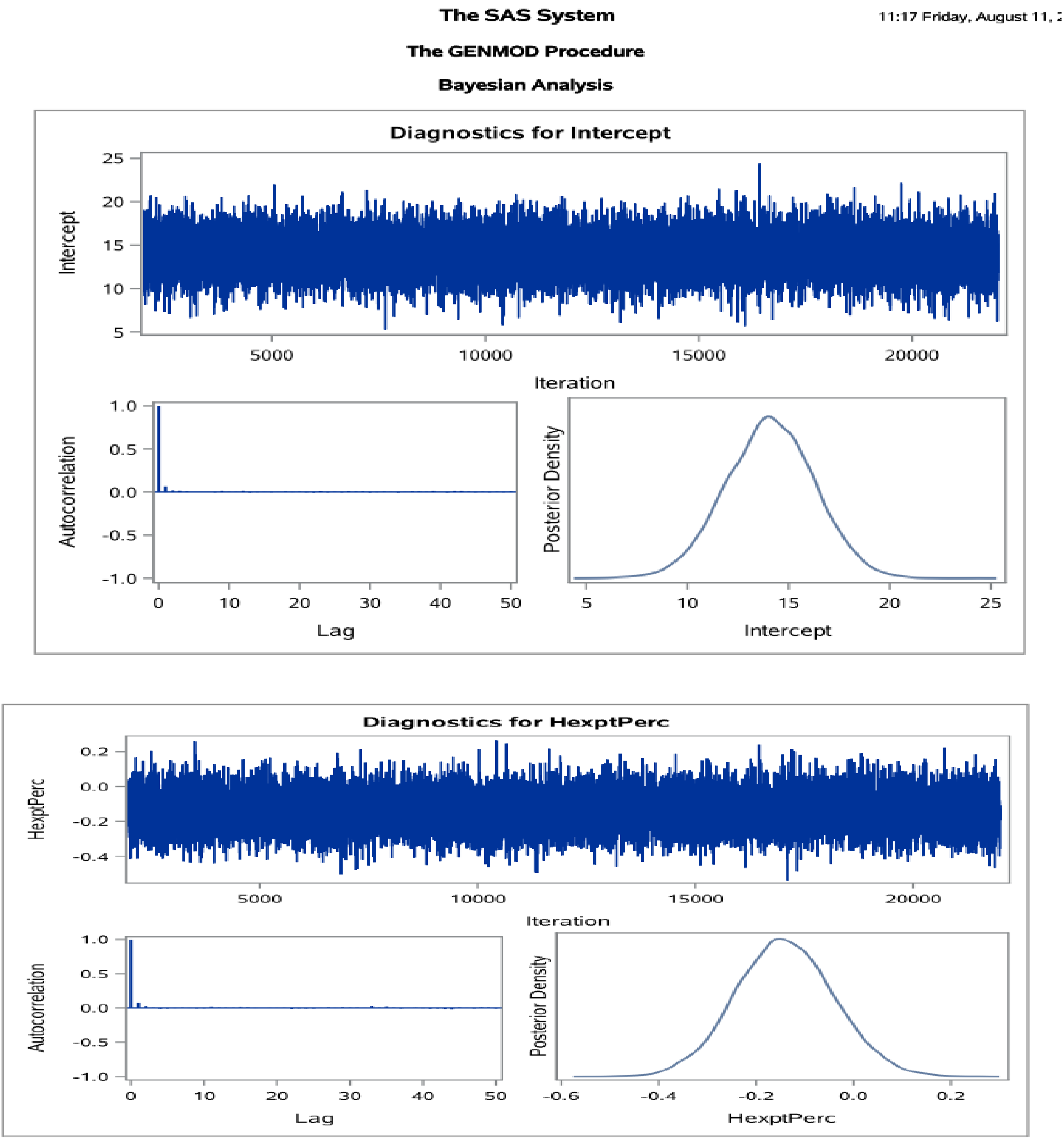

Diagnostics for Gdpperc
Gdpperc
-95E-6
-95E-6
-96E-6
-96E-6
-96E-6
-96E-6
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-96E-6
-96E-6
-95E-6
-95E-6
Gdpperc


Diagnostics for PopuagePerc
PopuagePerc
-0.0367
-0.0368
-0.0369
-0.0370
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.0370
-0.0369
-0.0368
-0.0367
PopuagePerc


Diagnostics for PhycDensPerc
PhycDensPerc
1.0930
1.0925
1.0920
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
1.0920
1.0925
1.0930
PhycDensPerc


Diagnostics for MYSPerc
MYSPerc
0.1978
0.1976
0.1974
0.1972
0.1970
0
10000
20000
30000
40000
50000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0.1970
0.1972
0.1974
0.1976
0.1978
0.1980
MYSPerc

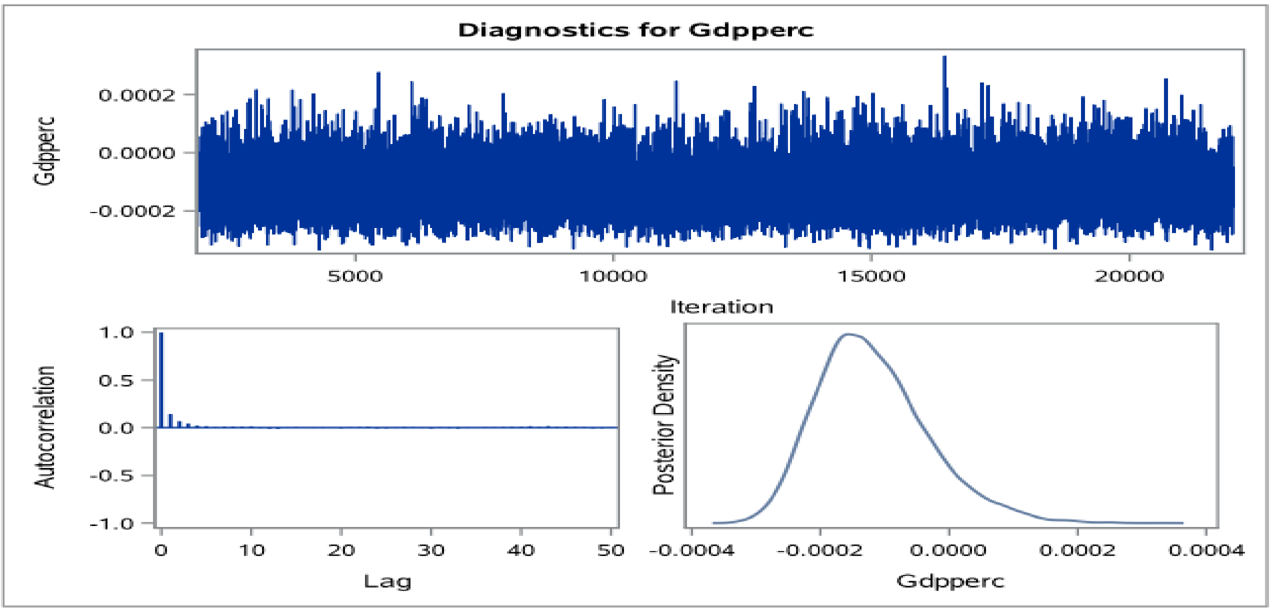
Appendix E: MCMC Plots for Negative Binomial

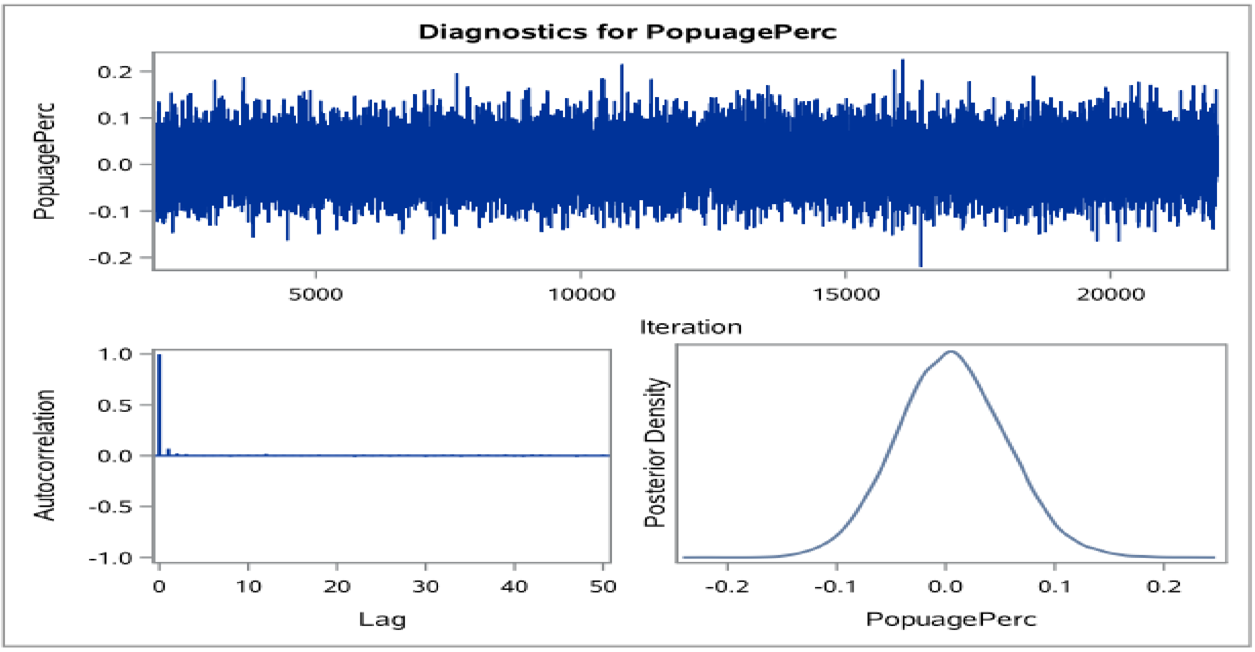


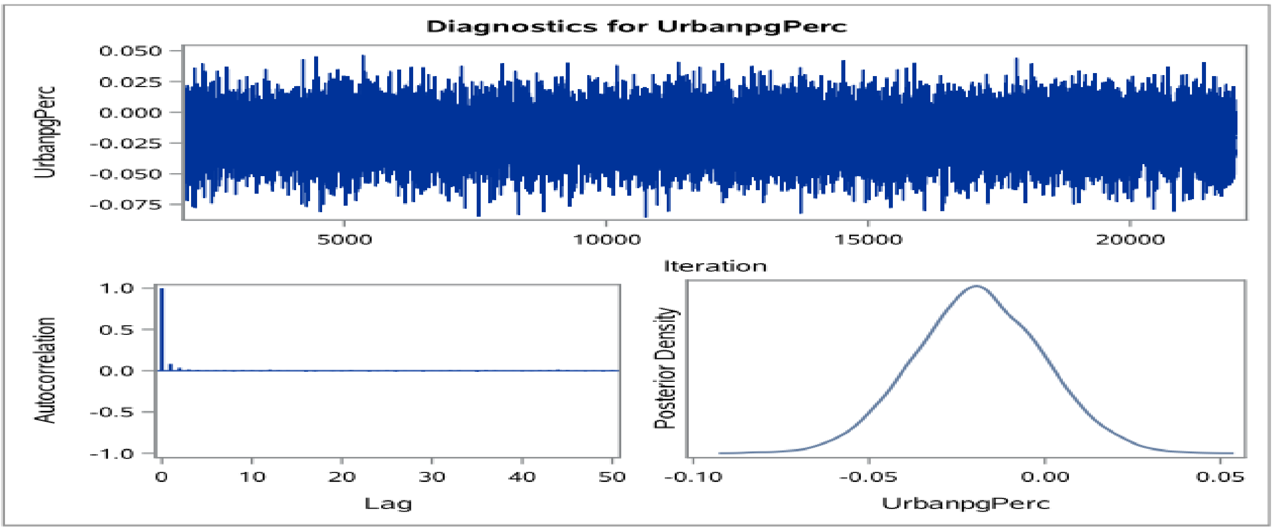
The SAS System 11:17 Friday, August 11, 2

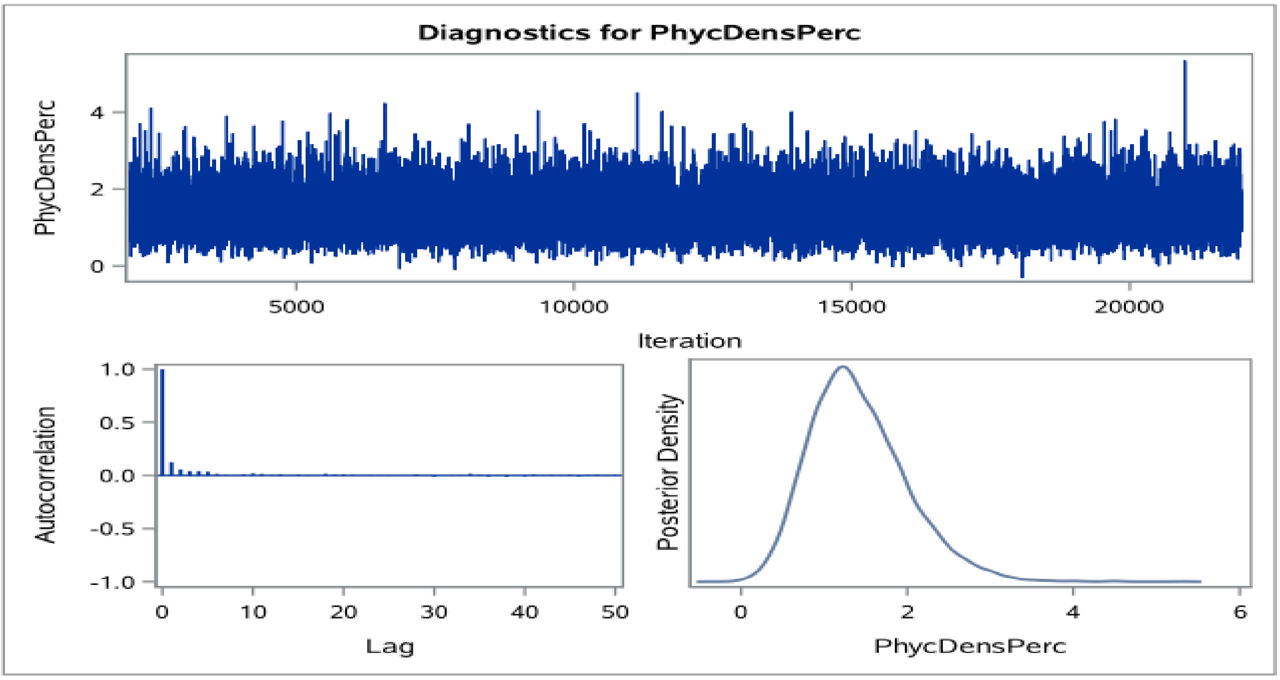
The GENMOD Procedure

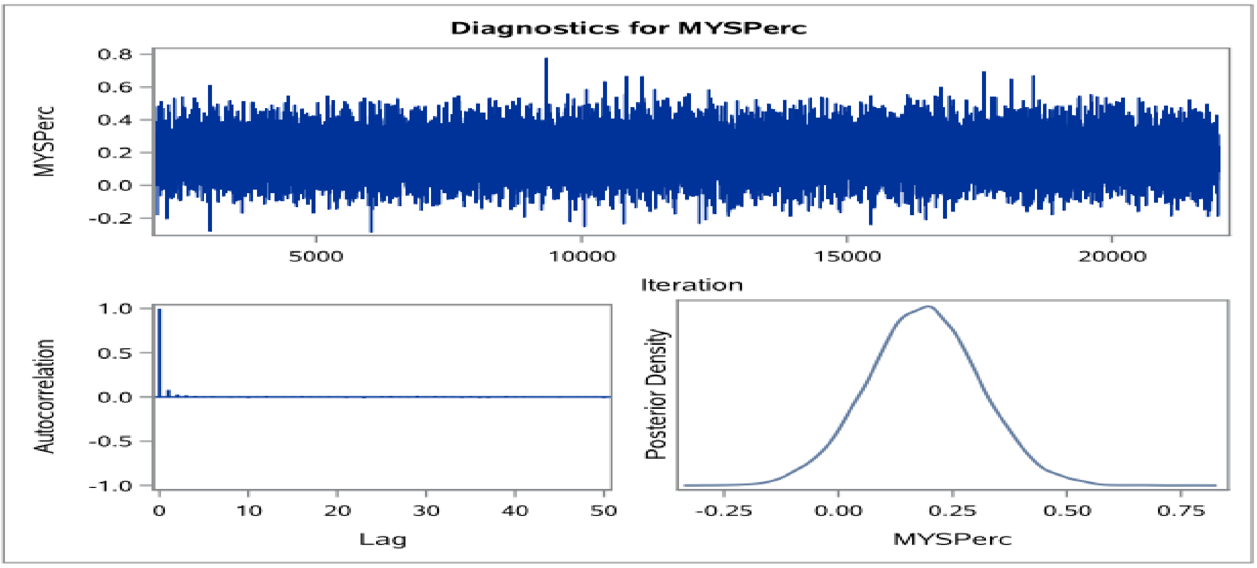
Bayesian Analysis

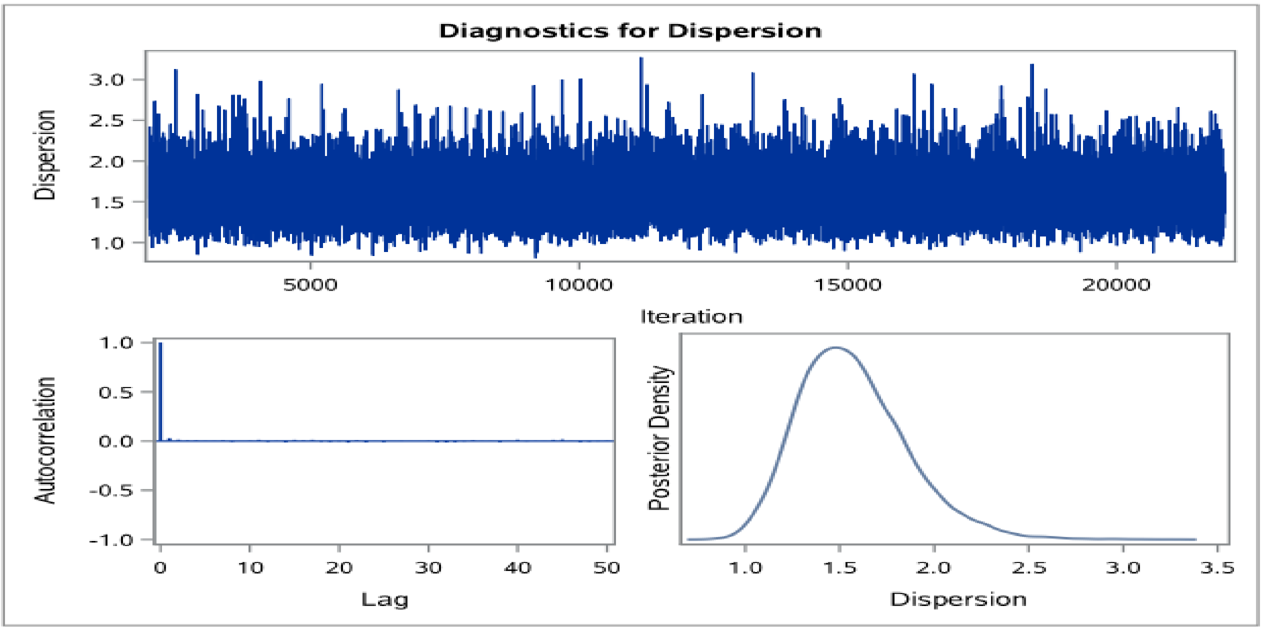

Diagnostics for Gdpperc
Gdpperc
0.0002
0.0000
-0.0002
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.0004
-0.0002
0.0000
0.0002
0.0004
Gdpperc


Diagnostics for PopuagePerc
PopuagePerc
0.2
0.1
0.0
-0.1
-0.2
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.2
-0.1
0.0
0.1
0.2
PopuagePerc


Diagnostics for UrbanpgPerc
UrbanpgPerc
0.050
0.025
0.000
-0.025
-0.050
-0.075
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.10
-0.05
0.00
0.05
UrbanpgPerc


Diagnostics for PhycDensPerc
PhycDensPerc
4
2
0
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0
2
4
6
PhycDensPerc


Diagnostics for MYSPerc
MYSPerc
0.8
0.6
0.4
0.2
0.0
-0.2
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.25
0.00
0.25
0.50
0.75
MYSPerc


Diagnostics for Dispersion
Dispersion
3.0
2.5
2.0
1.5
1.0
5000
10000
15000
20000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
1.0
1.5
2.0
2.5
3.0
3.5
Dispersion