SPATIAL MODELING AND MAPPING OF COUNT DATA WITH UNDER-REPORTING: THE CASE OF DIABETES IN KENYA

MBECHE COLLINS KABA

A THESIS SUBMITTED TO THE SCHOOL OF PURE AND APPLIED SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS

MAASAI MARA UNIVERSITY

2024

Declaration

This research thesis is my original work and has not been presented for a degree in any other university. No part of this thesis may be reproduced without the author and/or Maasai Mara University's authorization.

MBECHE COLLINS KABA

SM04/JP/MN/13784/2021

Signature____________________ Date____________________

This thesis has been submitted for examination with the approval of the following supervisors:

DR. JOSEPH OUNO OMONDI

Department of Mathematics and Physical Sciences

Maasai Mara University

P. O. Box 861-20500, Narok-Kenya.

Signature____________________ Date_____________________

DR. JUSTIN OBWOGE OKENYE
Department of Mathematics
Egerton University
P. O. Box 536-20115, Egerton-Njoro, Kenya.

Signature____________________ Date_____________________

Acknowledgement

I extend my deepest gratitude to the Almighty God for the gift of life and unwavering guidance throughout the journey of writing this thesis. I am profoundly thankful to Maasai Mara University for its steadfast support during my research, particularly the members of the School of Pure and Applied Sciences, and the Department of Mathematics and Physical Science.

My heartfelt thanks go to my supervisors, Dr. Joseph Ouno Omondi and Dr. Justin Obwoge Okenye, whose guidance, tireless patience, constructive criticism, and unwavering dedication have been instrumental in the successful completion of this study. Their professional and moral support throughout my graduate program has been invaluable.

I am also deeply appreciative of the love and support from my family, colleagues, and friends, whose encouragement has sustained me thus far. Special thanks to my fellow postgraduate colleagues, whose assistance, patience, and camaraderie made this journey both fulfilling and meaningful.

Dedication

I dedicate this thesis to my beloved parents, Mr. and Mrs. Mbeche, whose love and support have been my foundation. I also extend this dedication to my dear friends, family, and the entire Maasai Mara University community for their unwavering encouragement and support throughout this journey.

Abstract

Diabetes is a significant public health issue in developing countries, with an increasing burden on the healthcare system. However, accurate reporting of diabetes cases is often hindered by under-reporting, particularly in rural areas where access to healthcare is limited. When dealing with count data, both under-reported and over-reported cases are encountered. If it is assumed that the count data obtained from the field is always accurate, then modelling it with other count-data models will be erroneous. This study aimed to improve the existing Poisson-Binomial mixture model by factoring in covariates to make it suitable to estimate the number of under-reported diabetes cases in each county of Kenya and map the distribution of these cases. The covariates used in the model include the education level, poverty index, and access to healthcare in respective counties, making the probability of reporting vary from one county to another. The data was obtained from the Kenya Diabetes Management Information Centre and Kenya National Bureau of Statistics. The results revealed that at least each of the 47 counties had under-reported the diabetes data, with the probability of reporting ranging from 0.9002423 for Migori County and 0.7164098 for Mombasa County. Nairobi and Mombasa counties reported the highest underreporting rate with 16,708 and 11,784 cases, respectively underreported, while Lamu had 1269 underreported cases, the least in all the 47 counties. The resulting maps identified high-risk areas for under-reporting, and the prevalence which provides valuable information for policymakers and public health practitioners to target resources towards improving diabetes prevention and management in Kenya.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Different techniques are used in epidemiology to study the patterns of diseases among human populations, such as descriptive and diagnostic analyzes. Spatial disease mapping is one of the common models that public health use in studying different diseases. With mapping techniques, epidemiologists and public health experts can detect the relationship between people and their environment (Koch, 2005).

Many epidemiologists have used spatial disease mapping to solve the health problems affecting society. One of the archived cases was conducted by Dr. John Snow in 1854.   According to Moore et al. (1999), there was an outbreak of cholera cases in London, later linked to a public water pump used in the area. Dr. Snow used the idea of spatial disease mapping to map the number of cases within the city, which later helped in handling the problem.

There have been advancements in the idea of disease mapping over time. With technological advancements, technologies such as Geographic Information Systems (GISs) have helped make disease mapping easier. The GIS system is an advancement of the traditional statistical approach where scatter plots were used in displaying the relationships between different variables. The GIS system can capture, manipulate, and analyze data using a single system. Moore et al. (1999) further explain that the GIS technology was further advanced by introducing the idea of clustering, making it easier to compare the different geographical locations.

Disease clustering is one of epidemiologists' significant techniques to study the occurrence and intensity of different diseases in a given location. The technique has been justified by the fact that the movement of people from one area to another within a given locality cannot be randomized. Wartenberg (1999) explains that clustering helps draw the spatial patterns of a disease among people living in a given geographical area.

Environmental factors are one of the aspects that are said to have a direct effect on the occurrence and intensity of a disease in a given area. Almani et al. (2008) confirm this assertion that the etiological factors (that is, environmental factors) directly affect the occurrence and intensity of a given disease. The epidemiologists have given the idea of clustering priority considering the already established relationship between the demographic variable and the spatial patterns. Although the idea of clustering helps identify the spatial patterns in a given population, there are other factors, other than the demographic factors, which may determine the occurrence and intensity of a given disease. One such factor is the nature of the population that is being studied. In the case of infectious diseases, the disease transmission process contributes to inducing the spatial patterns exhibited in the data. Other than these factors, variables associated with the administrative approaches and the poverty level also affect the occurrence and intensity of disease in a given area.

A study by Moore et al. (1999) reveals that the spatial patterns of the disease in question can be significantly impacted by the nature and type of data collected. Although under-reporting and over-reporting of cases occur in different scenarios, under-reporting of disease cases has been termed an impediment in determining the actual spatial patterns of a given disease. With advanced technologies, developed countries ensure that the number of reported cases is almost accurate. However, the situation in a majority of African countries is different. Moore et al. (1999) further explain that some aspects that affect most African countries, resulting in underreporting, include inadequate medical funds, low medical knowledge, poverty, and stigmatization, among others. The locals may also need more confidence in the existing medical institutions, hence deliberately avoiding disclosing such important information. With the increased under-reporting of disease cases, some of the count models developed to model the cases seem inaccurate (Ye & Lord, 2011).

Epidemiologists and policymakers heavily rely on the number of disease counts that are reported to make some critical decisions with respect to reducing the number of disease cases. The high under-reporting rate makes it difficult to estimate the actual state of a given disease. Epidemiologists have claimed that there are scenarios where the under-reported cases are immeasurable, making estimating the actual disease count difficult. One typical case is encountered when collecting data associated with drinking habits. In this case, there is a high likelihood that the male respondents will answer the question as required instead of the female respondents. This, therefore, means that ignoring the gender variable will lead to a collection of data that is biased along gender lines (Ye & Lord, 2011).

Several studies have been conducted on diabetes, which concluded that the disease is a significant health issue associated with increased mortality and morbidity among African countries. According to a WHO (2023) study, about 60% of the people affected by diabetes live in middle and low-income economies. The report also indicated that out of about 87 million people who are suffering from diabetes in the world, about 22 million are from African countries. The preference for diabetes in Africa is increasing; about 10.4 million people are affected as of 2017 (WHO, 2013).

Count models have been developed to estimate the number of disease cases. However, some of these models have drawbacks associated with their efficiency and sufficiency, raising questions about the better models which can be trusted in estimating the under-reporting of count data. The Poisson model, negative binomial, and zero-inflated Poisson are some models developed to capture over-dispersion and under-dispersion. This study will consider developing a Bayesian spatial modeling approach that can be used in estimating possible under-reporting of count data and in application to the case of under-reporting for diabetes cases in Kenya. The adjusted data will then be used in mapping the prevalence of diabetes in the 47 counties in Kenya.

1.2 Statement of the problem

Diabetes is a growing health concern in Kenya based on a report by Diabetes Study Group (2019) covering periods between 2010 and 2018, and accurate reporting of cases is essential for effective prevention and treatment. Different models, such as Kriging and the Empirical Bayes, have been developed to investigate the presence of spatial property on count data. However, most models assume that the data in question is correctly reported, ignoring the issue of under-reporting (Zayeri et al., 2011). Diabetes Study Group (2019) revealed that some of the cases of diabetes are not accounted for, making it difficult to know the risks of the disease in different areas. Therefore, there is a need to study the under-reported cases of diabetes in Kenya and develop a map to show the relative risk of the disease in different counties. Under-reporting of diabetes cases is common, particularly in rural areas with limited access to healthcare. This study is meant to improve the Poisson-Binomial mixture model to make it suitable for estimating the number of under-reported diabetes cases in each county of Kenya and create a map of under-reported cases using available data on reported cases and relevant covariates.

1.3 Objectives

1.3.1 Main objective

The main objective is to develop a spatial model for estimating the under-reporting of diabetes data and mapping areas with diabetes cases in Kenya.

1.3.2 Specific objectives

i. To developed an improved Poisson-Binomial model to account for under-reported cases in spatial units (that is, counties).

ii. To estimate the under-reported diabetes cases in Kenya using the improved Poisson-Binomial model with covariates.

iii. To develop a diabetes prevalence map for the 47 counties in Kenya.

1.4 Significance of the study

Most African countries are still developing, based on the condition of their basic amenities. Due to the scarcity of resources in most countries, policy makers need to make informed decisions to ensure taxpayers' money is spent on suitable projects. The Kenyan government, at times, has found itself at crossroads in the healthcare area where they need to spend the resources. By mapping the distribution of under-reported diabetes cases, this study can help identify counties or regions where the burden of diabetes is highest and where resources are to be prioritized to improve diabetes prevention and management. The findings of this study provide practical implications for policymakers and public health practitioners in Kenya, as well as other countries with similar healthcare contexts, by informing the design and implementation of diabetes prevention and management programs. The developed model can be applied to other diseases with under-reporting issues and other spatial contexts beyond Kenya, providing a valuable tool for public health researchers and policymakers.

1.5 Scope of the study

This study aimed at determining a model that can be used in modeling under-reported diabetes cases as well as mapping the relative risk of the disease in Kenya. The study used data on reported cases and covariates, including the illiteracy level, access to healthcare, and poverty index in respective counties, to develop and apply the model to estimate the number of under-reported cases in each county. The data on the number of diabetes cases reported from each Kenyan county was obtained from the Kenya Diabetes Management Information Center (https://dmi.or.ke/?page_id=3165) and

Kenya National Bureau of Statistics (https://knbs.or.ke/visualizations/). The map showing the county boundaries that were considered during mapping is as shown below:

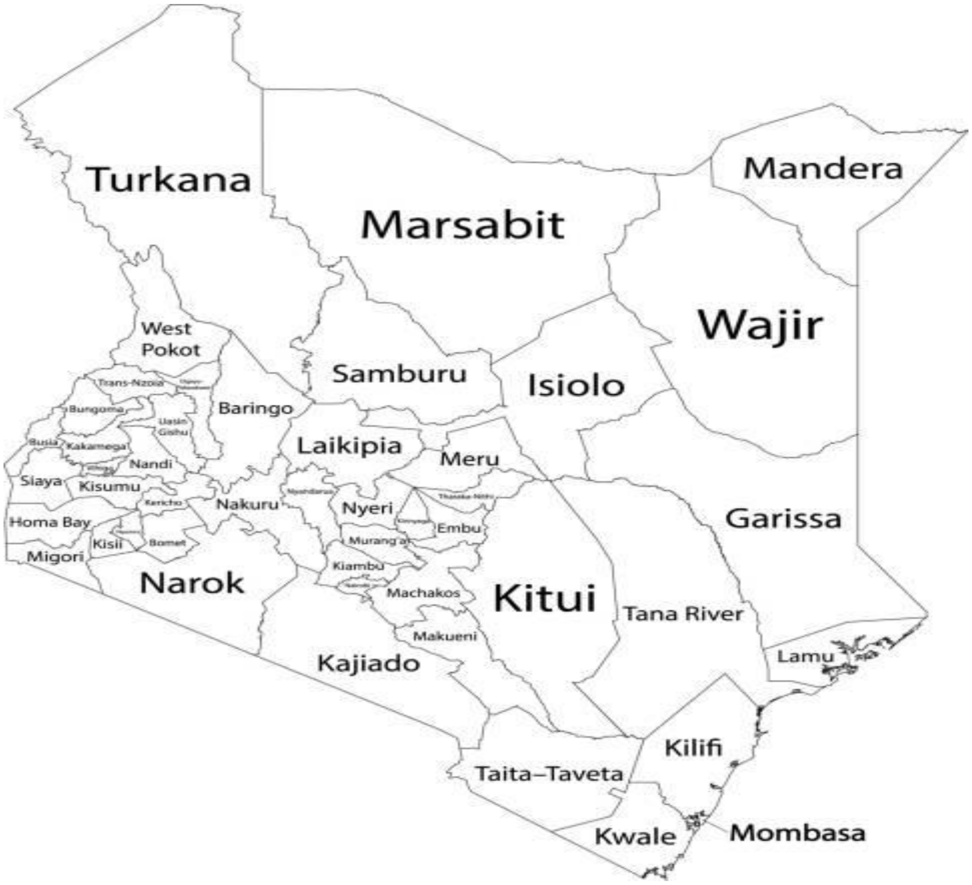


Figure 1.1: Map showing the geographical boundaries of counties in Kenya

The study was limited to diabetes cases in Kenya and treated each county as a spatial unit for analysis. The study did not focus on the causes or treatment of diabetes but on improving estimates of under-reporting diabetes cases. The study did not explore other non-diabetes-related diseases or health conditions that may be subject to under-reporting. Finally, the study was limited by the quality and availability of data on reported cases and relevant covariates, which may affect the accuracy of the estimates and maps produced by the model.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of different studies related to the study. The chapter summarizes the spatial mapping techniques and other statistical and non-statistical techniques used to assess spatial patterns.

The number of cases of diabetes in the country has been increasing rapidly hence Epidemiologists and policy makers are making efforts to ensure that the cases are controlled (WHO, 2022). There are pretty several interventions that have been put in place, which include but are not limited to improving healthcare infrastructure and ensuring that there is a quick response from the health sector when it comes to diabetes cases. Although several interventions have been proposed, policy makers need to determine the interventions which they can prioritize to handle the situation amicably. On the other hand, there is a need to determine the areas most affected by the disease to improve health infrastructure. Doing this will significantly help control the risk of disease outbreaks (WHO, 2022).

2.2 Diabetes Under-reporting in Kenya

This literature review examines the extent of diabetes under-reporting in Kenya and the factors contributing to under-reporting. Diabetes is a chronic disease that affects millions of people worldwide. In Kenya, diabetes is a significant public health concern, with an estimated 2.2 million people living with the disease in 2021 (WHO, 2022). Despite the high prevalence of diabetes, under-reporting of cases is a significant problem in Kenya (WHO, 2022). Diabetes under-reporting in Kenya has been a subject of interest for researchers studying public health and epidemiology. Some studies have focused on assessing the extent of under-reporting, identifying factors contributing to under-reporting and proposing strategies to improve reporting accuracy.

The prevalence of diabetes in Kenya has been increasing steadily over the years. A study conducted in 2019 found that diabetes in Kenya was 4.4%, with a higher prevalence in urban areas (7.3%) compared to rural areas (2.4%) (WHO, 2014). However, these figures will likely be underestimated due to the under-reporting of cases (WHO, 2014).

Under-reporting of diabetes cases in Kenya has several consequences. First, it leads to an underestimation of the burden of diabetes, making it difficult for policy makers to allocate resources and develop effective strategies to manage the disease. Secondly, it can lead to delayed diagnosis and treatment, resulting in complications and poorer health outcomes. Finally, under-reporting diabetes cases can lead to a lack of public awareness of the disease (WHO, 2022).

A study was conducted by Manda & Feltbower (2018) to investigate the presence of under-reporting of data on notifiable and non-communicable diseases. The study developed a spatial model to estimate the under-reporting of notifiable and non-communicable diseases in England and Wales, using data on reported cases and relevant covariates. The scholars used Bayesian models but excluded the use of posterior distribution in adding the prior knowledge in estimation because they found it too complex. The study showed that the spatial model can improve estimates of under-reporting and identify high-risk areas for intervention.

Another study was conducted by Adamjee and Harerrimana (2022), which was aimed at estimating the burden of diabetes mellitus in Kenya. This cross-sectional study aimed at estimating the prevalence of type 2 diabetes and assessing its reporting accuracy. The study highlighted challenges in diabetes surveillance, including under-reporting, and provided insights into the gaps that must be addressed. The significant gap identified was on the under-reporting of data, where scholars recommended a model to be developed to help identify under-reported data to make decision-making by the government easier.

A study by Ayugi et al. (2019) investigated the spatial patterns and factors associated with under-reporting diabetes cases in Kenya, using data from the 2015 Kenya Stepwise Survey of Non-Communicable Diseases. The study identified under-reporting as a significant issue in Kenya and recommended using spatial modeling to improve estimates of diabetes burden.

Another study titled "Evaluation of the completeness of diabetes-related reporting systems in Kenya" conducted by Mwita et al. (2019) assessed the completeness and accuracy of diabetes-related reporting systems in Kenya. They evaluated the existing reporting mechanisms and identified areas for improvement to enhance the quality and accuracy of diabetes data. This study pointed out that some of the cases of diabetes within the communities have not been adequately documented, making it hard to budget for the disease in the country.

Furthermore, the Kenya National Diabetes Strategy report (2015) acknowledged the issue of under-reporting. It emphasized the need for a comprehensive surveillance system to capture accurate and representative data on diabetes prevalence and burden in the country. According to the report, most of the counties in Kenya lacked accurate data on the number of people with diabetes. The report recommended that the counties develop better disease recording systems that will help present the actual data on certain diseases affecting respective counties.

2.3 Under-reporting models

Count data is widely used in various fields, including public health, epidemiology, and social sciences. However, the accuracy of count data can be affected by under-reporting, which is the missing or incomplete reporting of events. Under-reporting of count data is a common problem that can lead to biased estimates and incorrect conclusions. Therefore, statistical models have been developed to address this issue. This literature review examines the different statistical models proposed to account for the under-reporting of count data (Zayeri et al., 2011).

The negative binomial model is a popular model used to account for the under-reporting of count data. It assumes that the observed count data is a mixture of a Poisson distribution and a gamma distribution, where the Poisson distribution represents the actual count data and the gamma distribution represents the under-reporting component. The negative binomial model is flexible and can accommodate different under-reporting forms, including under-dispersion and over-dispersion. However, it requires the assumption of a specific distribution for the under-reporting component (Zayeri et al., 2011).

Capture-recapture models are another class of statistical models used to account for the under-reporting of count data. These models are based on the principle that the probability of a case being reported increases with the number of sources reporting it. The simplest capture-recapture model is the two-source model, which assumes that the observed count data is a mixture of the actual count data and the under-reporting component. The under-reporting component is estimated using the cases reported by only one source. The advantage of capture-recapture models is that they do not require any assumptions about the distribution of the under-reporting component. However, they require the assumption of independence between the reporting sources (Zayeri et al., 2011).

Zero-inflated models are another class of statistical models used to account for the under-reporting of count data. These models assume that the observed count data has excess zeros compared to a Poisson or negative binomial distribution. The excess zeros are attributed to the under-reporting component (Branscum et al., 2005). Zero-inflated models are flexible and can accommodate different forms of under-reporting. However, they require the assumption of a specific distribution for the excess zeros, which might not be met in real-life situations for count data.

2.4 Spatial disease mapping

Spatial disease mapping is one of the methods applied in generating disease patterns and identifying areas significantly affected by a given disease. A study by Zayeri et al. (2011) revealed that epidemiologists have embarked on using a spatial mapping approach to identify the risk of contracting a particular disease by individuals living in a specific area. One of the cases where spatial disease mapping has been successfully applied was by Zayeri et al. (2011), whereby they developed a malaria spatial map in Sistan, Iran. Data obtained from the area was used in computing the Standard Incidence Rates (SIR), after which mapping indicated the most affected areas.

Over time, different statistical models have been used in mapping disease incidences and risks in certain areas. Mora and Lawson (2012) applied the Gaussian Component Mixture (GCM) model to map the risk of contracting a disease. The study utilized the Bayesian method in investigating the spatial effect in a Univariate sense. Although Bayes' theorem was applied in the study, they assumed that the dataset used was a true reflection of reality, which is not always true.

A study by Gibbons et al. (2014) insisted that the tool used to model spatial effect should be efficient. Although the use of GCM by Mora and Lawson (2012) was reliable, assuming that the actual data made the model work differently than expected. A study by Neubauer et al. (2016) revealed that in case there is under-reporting or over-reporting of data, the use of the Gaussian Component Mixture will fail. To solve the problem of under-reporting that may arise in most of the real-life data, the study used the existing binomial model but factored in under-reporting. However, the extension has been questioned as it uses the frequentist approach, which is less reliable and less efficient than the Bayesian approach.

Some of the studies have considered the issue of under-reporting while modeling count data. Gamado et al. (2014) conducted an epidemic study considering the stochastic Markov Susceptible-Infected-Recovered (SIR) model. The study revealed that the use of models which ignore the issue of under-reporting of disease count would result in under-estimating the risk of contracting a disease.

Although the models developed by most scholars here help predict the spatial effects, some fail to estimate disease risks, especially in Africa, where some disease occurrences still need to be reported. There is a need for an effective model that can be used in estimating the disease count, taking into consideration the aspect of under-reporting. Kenya, among other developing countries, needs more resources to collect actual data, making scholars rely on latent variables. This forms part of the reasons that a more efficient and reliable model needs to be developed that considers under-reporting while investigating disease incidence. Moreover, most developed models need to consider some covariates, which may significantly impact the reporting of count data.

Different models, such as Kriging and the Empirical Bayes, have been developed to investigate the presence of spatial property on count data. However, most models assume that the data in question is correctly reported, ignoring the issue of over-reporting and under-reporting (Zayeri et al., 2011). The problem of underreporting generally affects the modelling of the diabetes data obtained from different counties. This knowledge gap is significant because it may lead to an underestimation of the burden of diabetes and inadequate allocation of resources for diabetes prevention and control efforts. Policymakers have raised the issue of underreporting as it makes it difficult to plan for the areas affected. There is, therefore, a need to estimate the under-reported cases of diabetes in the country and develop a map to show the distribution of diabetes in different counties.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

In this chapter, we discuss the different techniques that were used in achieving the objectives. The chapter will present the proposed modified model that can be used in modeling data with underreporting and the efficiency of the model.

3.2 Count data models

Count data is a common term used in statistics to represent a data type with countable quantities. Count data comprises non-negative integers that generally result from counting and not ranking. Different models have been developed over time and can be used in modeling count data. The common models include the Binomial distribution, Poisson distribution, and the Negative Binomial distribution, among others as has been discussed in chapter 2. The Poisson approach is more applicable in this study since it can be modified to accommodate under-reporting. One needs to evaluate the type of count data to be modeled before deciding the best model to suit the data. This chapter presents a theoretical approach to analyzing count data with under-reporting and a proposed model that can be used in estimating the risks of the disease.

The study used a Poisson-Binomial mixture model to estimate the number of under-reported diabetes cases in each county of Kenya and map the distribution of these cases. The Poisson-Binomial mixture model is a statistical method used to model count data with over- or under-dispersion, a common issue with health data. The model combines the Poisson distribution and the Binomial distribution to account for the mean and variance of the count data. The model uses data on reported diabetes cases and relevant covariates such as education level, poverty rate, and access to healthcare to estimate the number of under-reported cases in each county. The covariates are included in the model as fixed effects to control potential confounding effects. Spatial autocorrelation are accounted by using a spatial random effect.

3.3 Poisson-Binomial model for undercount of disease cases

Supposing we consider a case where y_i^*represents the total cases of a disease in county i, and assuming also that y_i^* has a Poisson distribution. The main issue experienced here is the problem of under-reporting. Given that the number of reported cases is represented by y, it is essential to note that it does not represent the actual value of the disease count at unit i. It therefore means that y_i represents a fraction of the reported cases of y_i^*in county i. This leads us to a binomial distribution as follows:

$$P(y_i|y_i^*, \lambda_\mathrm{i}) \sim Bin\ (y_i^*, \lambda_\mathrm{i})$$

There are two possibilities that can be realized about the number of reported cases. First, number of reported cases equals the actual number, whereby $y_i = y_i^*$. On the other hand, the number of reported cases can be less than the actual number which can be expressed as $y_i = y_i^* - n$ whereby $n < y_i^*$, where n is the number of under-reported cases.

Assuming that the observed disease count has a binomial distribution, and the actual disease count takes a Poisson distribution, then the marginal number of reported cases (y_i) will be given by:

$$P(Y_i = y) = \sum_{y^* \geq y}^{\infty} \binom{y^*}{y} p^y (1-p)^{y^*-y} \frac{\lambda^{y^*} e^{-\lambda}}{y^*!} \tag{3.3}$$

where y is the number of reported cases (observed disease count), y* is the total number of cases (both reported and unreported),

Factoring out the variables independent of y^* in equation 3.3, we get:

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{(1-p)^y} \sum_{y^* \geq y}^{\infty} \binom{y^*}{y} (1-p)^{y^*} \frac{\lambda^{y^*}}{y^*!} \tag{3.4}$$

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{(1-p)^y} \sum_{y^* \geq y}^{\infty} \frac{y^*!}{y!(y^*-y)!} \ (1-p)^{y^*} \frac{\lambda^{y^*}}{y^*!} \tag{3.5}$$

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{y!(1-p)^y} \sum_{y^* \geq y}^{\infty} \frac{1}{(y^*-y)!} \ (1-p)^{y^*} \lambda^{y^*} \tag{3.6}$$

Now adding y and subtracting y on the powers with y^*, we have:

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{y!(1-p)^y} \sum_{y^* \geq y}^{n-1} \frac{1}{(y^*-y)!} \ (1-p)^{y^*-y+y} \lambda^{y^*-y+y} \tag{3.7}$$

For which if we let $y^* - y = k$, we have

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{y!(1-p)^y} \sum_{k=0}^{n-1} \frac{1}{k!} \ (1-p)^{k+y} \lambda^{k+y} \tag{3.8}$$

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{y!(1-p)^y} \ (1-p)^y \lambda^y \sum_{k=0}^{n-1} \frac{1}{k!} \ (1-p)^k \lambda^k \tag{3.9}$$

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{y!} \ \lambda^y \sum_{k=0}^{n-1} \frac{1}{k!} \ (1-p)^k \lambda^k \tag{3.10}$$

$$P(Y_i = y) = \frac{e^{-\lambda} (p\lambda)^y}{y!} \ \sum_{k=0}^{n-1} \frac{1}{k!} \ (1-p)^k \lambda^k \tag{3.11}$$

$$P(Y_i = y) = \frac{e^{-\lambda} (p\lambda)^y}{y!} \ e^{\lambda(1-p)} \tag{3.12}$$

$$P(Y_i = y) = \frac{e^{-\lambda} (p\lambda)^y}{y!} \ e^{\lambda} e^{\lambda p} \tag{3.13}$$

$$P(Y_i = y) = \frac{(\lambda p)^y}{y!} \ e^{-\lambda p} \qquad \mathit{where\ the\ number\ of\ reported\ cases,}\ \ y \geq 0 \tag{3.14}$$

Equation 3.14 has a Poisson distribution. This means that the number of observed counts (y) follow a Poisson distribution with mean λp as shown in equation (3.14). From the equation 3.14, when we ignore under-dispersion, the count data on the number of diabetes cases in Kenya can be modeled using a Poisson model.

Considering the fact that under-dispersion varies from one county to the other, it is important to note that the probability of under-dispersion also varies from one county to the other. The forgoing means that if we let i to represent county i and also that j represents county j, it means that the said probabilities is such that $P_i \neq P_j$.

3.4 Inclusion of covariates to the Poisson-Binomial model

Including covariates in the Poisson-Binomial model allows for a more comprehensive and accurate estimation of the under-reporting of diabetes cases. By incorporating those covariates, we can account for the potential effects of factors that influence the reporting behavior which vary across counties. This enables us to adjust the reporting probability based on the specific characteristics of each county, leading to more precise estimates of under-reporting. The significant covariates to be included in the model which might influence the rate of reporting of diabetes at the counties include the illiteracy level (X_1), access to healthcare (X_2), and poverty index (X_3).

Illiteracy level is an essential covariate that captures the influence of educational attainment on the reporting of diabetes cases. Counties with higher literacy levels are more likely to better understand diabetes symptoms, leading to a higher probability of reporting cases accurately. Conversely, lower education levels may result in under-reporting due to limited awareness and knowledge regarding diabetes.

The poverty rate serves as an essential indicator of socioeconomic conditions within counties. Higher poverty rates often coincide with limited access to healthcare resources, including diagnostic facilities and regular health check-ups. This, in turn, may contribute to the under-reporting of diabetes cases due to reduced healthcare-seeking behavior and inadequate healthcare infrastructure.

Access to healthcare plays a crucial role in the reporting of diabetes cases. Counties with better access to healthcare facilities, including primary care clinics, hospitals, and specialized diabetes centers, are more likely to have an improved reporting system. Conversely, counties with limited access to healthcare services may need help in timely diagnosis and reporting, resulting in potential under-reporting.

To specify the relationship between the covariates and the probability of reporting a diabetes case, logistic regression is used as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{3.15}$$

The coefficients $\beta_0, \beta_1, \beta_2, and\ \beta_3$ capture the effects of the respective covariates on the probability of reporting a case in a given county.

By incorporating education level (X_1), poverty rate (X_2), and access to healthcare (X_3) as covariates in our model, we aim to account for the influence of these factors on the reporting probability. This approach allows us to adjust the reporting probability based on the specific characteristics of each county, providing a more accurate estimation of under-reporting and a deeper understanding of the impact of education, poverty, and healthcare access on diabetes reporting patterns.

3.5 Bayesian framework

From the Poisson-Binomial model, five parameters need to be estimated. The first one is λ which represents the average actual diabetes cases in a county. The other four parameters are $\beta_0, \beta_1, \beta_2, and\ \beta_3$, which represent respectively how individual covariates affect the probability of reporting of diabetes cases. Each coefficient measures the effect of the corresponding covariate on the log odds (logit) of the probability of reporting a diabetes case.

3.6 Prior estimation for the Poisson-Binomial model

For λ, a gamma prior distribution was used. The gamma prior for λ can be specified using two parameters, a and b. Hyperparamer a determines the shape of the distribution, while b controls the rate of decay. By choosing appropriate values for a and b you can reflect your prior beliefs about the average true diabetes cases in the spatial units.

$$\lambda \sim \mathrm{Gamma}(a, b)$$

where a and b are the gamma distribution's shape and rate parameters, respectively. Suppose there is a belief that the true diabetes cases are expected to be small or have a lower mean value. In that case, the gamma prior to being chosen will have a smaller a parameter and a larger b parameter. Conversely, suppose there is a belief that the true diabetes cases are expected to be large or have a higher mean value. In that case, the gamma prior to being chosen will have a larger a parameter and a smaller b parameter.

Normal priors will be used in the case of the beta parameters as shown below:

$$\beta_0 \sim Normal(\mu_0, \sigma_0^2)$$

$$\beta_1 \sim Normal(\mu_1, \sigma_1^2)$$

$$\beta_2 \sim Normal(\mu_2, \sigma_2^2)$$

$$\beta_3 \sim Normal(\mu_3, \sigma_3^2)$$

Using normal priors for beta parameters is important because they are conjugate priors, which means that the posterior distribution will also be a normal distribution. This makes the posterior distribution easier to calculate and interpret, as it can be described by its mean and variance. The normal priors allow for the incorporation of prior knowledge or beliefs about the beta parameters into the analysis. The mean of the normal prior can represent the expected value or the researcher's prior belief about the true value of the beta parameter, while the standard deviation can represent the uncertainty or variability around that belief. Using normal priors can also facilitate the interpretation of results. The mean of the normal prior represents the initial best guess or central tendency for the beta parameter, while the standard deviation represents the initial uncertainty or spread around that best guess. This allows for a more intuitive understanding of the prior distribution and the resulting posterior distribution.

The unnormalized posterior distribution will be obtained by finding the product of the likelihood function and the priors of the distribution.

$Posterior \propto Likelihood\ X\ Priors$

$$Posterior \propto L(\lambda, \mathrm{p}) * \pi(\lambda) * \pi(\beta_0) * \pi(\beta_1) * \pi(\beta_2) * \pi(\beta_3)$$

where $\pi(\lambda), \pi(\beta_0), \pi(\beta_1), \pi(\beta_2), \pi(\beta_3)$ represent the prior distributions for the respective parameters.

The full conditioned model was given as a product of the likelihood function and the priors respective to the five parameters of interest as follows:

$$Full\ conditioned\ model = l(\lambda, \mathrm{p}).\, \pi(\lambda).\, \pi(\beta_0).\, \pi(\beta_1).\, \pi(\beta_2).\, \pi(\beta_3)$$

To obtain the exact normalized posterior distribution, we need to compute the evidence (also known as the marginal likelihood) by integrating the unnormalized posterior distribution over the parameter space. The general normalized posterior distribution was given by dividing the full model by the marginal likelihood, also known as evidence.

3.7 Gibbs sampling

Gibbs sampling is a specific type of MCMC algorithm that is relatively easy to implement. It is particularly useful when the conditional distributions of the parameters given the other parameters (known as full conditional distributions) are easily accessible. Gibbs sampling updates one parameter at a time, conditioned on the current values of the other parameters, by sampling from its full conditional distribution. This process iteratively updates all the parameters until convergence is achieved (Gibbons et al., 2014). Gibbs sampling has the advantage of being conceptually straightforward and computationally efficient when the full conditional distributions are known or can be easily sampled from. It does not require tuning proposal distributions or acceptance ratios like the MH algorithm. The steps that will be used in Gibbs sampling are as follows:

1. Initialization of the model parameter.
2. Iterate the following steps until convergence is reached:

Step 1: Sample λ from its full conditional distribution

i. Calculate the full conditional distribution of λ given the current values of β_0, β_1, β_2, β_3, and the observed data.

ii. Sample a new value for λ from its full conditional distribution.

Step 2: Sample β_0 from its full conditional distribution

i. Calculate the full conditional distribution of β_0 given the current values of λ, β_1, β_2, β_3, and the observed data.

ii. Sample a new value for β_0 from its full conditional distribution.

The procedure above will be repeated for $\beta_1, \beta_2 and\ \beta_3$. After a sufficient number of iterations, the initial iterations will be discarded to remove the effect of the initial parameter values. The samples obtained will be analyzed to obtain the posterior summaries, including the means, variances, and credible intervals for the parameters of interest.

3.8 Disease mapping

Disease mapping is one of the major tools used in epidemiology, aimed at determining the geographical distribution of a disease burden. The study mapped the distribution of diabetes cases recorded in each of the 47 counties in Kenya. The region in question is divided into n units (where n=47) representing the 47 counties in Kenya, and the units are non-overlapping. Mapping of the diabetes data in R involves several steps. The first step was to ensure that the data to be used was in the correct format so that it can be analyzed and visualized effectively. The data also needs to be organized at the county level, as each county will be treated as a spatial unit in the analysis. The data used at this point entails data that has been adjusted for under-reporting.

The adjusted data was then uploaded into R, where it was merged with a map of Kenya. There are several ways to create a map of Kenya in R, but one common approach that was applied in this study was to the use shapefiles. Shapefiles are files that contain geographic information on boundaries such as county boundaries in Kenya. The shape files that were used in this case was obtained from the Kenya Open Data Portal (https://kenya.opendataforafrica.org/).

The shapefiles were then loaded into R. There are different libraries that can be used in loading the shapefiles in R. In this case, libraries such 'rgdal' and “tmap” will be used. The loaded shapefiles were plotted using R's mapping functions, such as 'plot ()' or 'ggplot2 ()'. The resulting map showed the boundaries of the counties in Kenya. The next step will be to merge the diabetes data with the county boundaries using a common variable which in this case was the county name. This was achieved using R's 'merge ()' function together with the inner_join () function. Once the data has been merged with the county boundaries, the next step was to create a choropleth map of diabetes prevalence in Kenya.

A choropleth map is a map that uses color to represent a variable, in this case, diabetes count in Kenya. The counties were grouped into six groups, depending on the diabetes count. To create a choropleth map, R provides several libraries such as 'ggplot2' or 'leaflet' that were used to visualize the diabetes count and prevalence data on the county boundaries.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1. Introduction

This chapter presents the outcomes of the spatial Bayesian analysis applied to investigate the presence of under-reporting in diabetes cases across the counties of Kenya. This study delves into the complexities of disease reporting and aims to provide a comprehensive understanding of the true distribution of diabetes cases in the country. This chapter represents a detailed explanation on modeling, estimation of parameters, estimation of under-reported data and mapping of the under-reported cases, true distribution and the prevalence of diabetes cases in Kenya.

4.2. Model estimation

The under-reporting of diabetes cases was assumed to follow a Poisson-Binomial distribution with parameters (that is, λ, which represents the mean distribution of the accurate cases reported and P, which represents the probability that a diabetes case is reported in a given county). Unlike in the original Poisson-Binomial model where P was constant across the different spatial units, the parameter P in the proposed and improved model varied from one county to another depending on the effects of the three covariates (that is, illiteracy level, access to healthcare facilities and poverty index).

Model estimation was conducted using the Bayesian technique where each parameter was assigned prior distribution. The mean parameter λ was assigned a gamma prior with hyperparameters 1 and 2 while the logistic regression parameters were assigned non-informative uniform priors. The analysis was conducted with the assumption that the parameter P depended on the three covariates, (that is, illiteracy level (X_1), access to healthcare (X_2) and poverty index (X_3)) across the 47 counties.

The estimation of parameters for the model on under-reporting of diabetes cases in Kenya was carried out using the Gibbs sampling Markov Chain Monte Carlo (MCMC) method. The Gibbs sampling MCMC is a powerful technique that allows us to estimate the parameters of the model. Gibbs sampling is widely used in Markov Chain Monte Carlo (MCMC) because of its simplicity and effectiveness, particularly when sampling from high-dimensional and complex probability distributions. Unlike other MCMC techniques, such as the Metropolis-Hastings algorithm, Gibbs sampling breaks down the sampling process by iteratively sampling each variable from its conditional distribution, given the current values of all other variables. This approach is advantageous when conditional distributions are simpler or standard, as is often the case in Bayesian hierarchical models. By focusing on conditional distributions, Gibbs sampling can efficiently capture the dependence structure of complex models without needing to handle the full joint distribution directly.

Another major advantage of Gibbs sampling over methods like Metropolis-Hastings is the reduced need for tuning. Metropolis-Hastings requires a carefully chosen proposal distribution, which, if poorly selected, can lead to high rejection rates or slow mixing, limiting the effectiveness of the sampling. In contrast, Gibbs sampling circumvents the need for a proposal distribution altogether. Since it leverages the conditional distributions, it generally results in smoother, more efficient sampling. This lack of tuning requirements can make Gibbs sampling an attractive choice for researchers and practitioners, particularly when they have limited time or resources for optimization and parameter adjustment.

In each iteration of the for-loop, the Gibbs sampler updates the values of the parameters that we had which include $\lambda, \beta_0, \beta_1, \beta_2 and\ \beta_3$ based on the full conditional distributions. The number of iterations that were used in this case were 1000. The parameters were estimated as follows; $\lambda = 33565.99$, $\beta_1 = 0.504$, $\beta_2 = 2$ and $\beta_3 = 0.437$.

The coefficient β_1=0.504 for illiteracy level suggests that for every one-unit increase in illiteracy level, the log-odds of a disease case being reported in a county increase by 0.504, assuming other factors remain constant. When we exponentiate this coefficient to interpret it in terms of odds, $e^{0.504} \approx 1.66$ it indicates that each unit increase in illiteracy level is associated with a 66% increase in the odds of reporting a disease case. This implies that higher levels of illiteracy in a county are linked to a greater likelihood of disease cases being reported. This association could be due to the limited health knowledge and lower health literacy among populations with higher illiteracy levels, which may lead to poorer health outcomes or delayed detection and treatment of diseases.

The second coefficient, β_2=2, corresponds to access to healthcare. A one-unit increase in access to healthcare is associated with a 2-unit increase in the log-odds of reporting a disease case. In terms of odds, exponentiating this coefficient gives $e^2 \approx 7.39$, meaning that a unit increase in access to healthcare is associated with a 639% increase in the odds of a disease case being reported. While this may initially seem counterintuitive, it can be interpreted to mean that counties with better access to healthcare are more likely to detect and report disease cases due to improved healthcare infrastructure and diagnostic capabilities. Areas with better healthcare access can identify and document cases that might otherwise go unreported in counties with limited healthcare resources. This increased likelihood of case reporting does not necessarily imply a higher incidence of disease but rather reflects an enhanced capacity for case detection.

The third coefficient, β_3=0.437, reflects the effect of poverty level on the likelihood of a disease case being reported. For every one-unit increase in poverty level, the log-odds of reporting a disease case rise by 0.437, holding other variables constant. When exponentiated $e^{0.437} \approx 1.55$ indicating a 55% increase in the odds of a disease case being reported with each unit increase in poverty level. This suggests that higher poverty levels are positively associated with the likelihood of disease cases being reported. In counties with greater poverty, populations may be more vulnerable to disease due to factors such as inadequate access to clean water, nutritious food, and preventive health services. Additionally, higher poverty levels are often linked to more crowded living conditions and reduced access to health education, which can increase disease transmission and exacerbate health issues, leading to a higher likelihood of disease reporting.

This logistic regression model highlights how illiteracy, healthcare access, and poverty influence disease reporting in Kenyan counties. Increased illiteracy and poverty levels are associated with higher disease reporting, possibly due to poorer health conditions and limited preventive care. Improved healthcare access, on the other hand, likely boosts reporting capacity rather than disease incidence itself, as better diagnostic resources make it easier to identify and document cases. Together, these coefficients illustrate the impact of socioeconomic factors on disease dynamics and healthcare reporting within the regional context of Kenya.

In this Bayesian analysis, the Gibbs sampling method was applied to estimate parameters λ=33565.99, β_1=0.504, β_2=2, and β_3=0.437 for a logistic regression model. Each parameter estimate provides insight into the relationship between predictors, specifically illiteracy level, access to healthcare, and poverty level, and the likelihood of disease case reports across spatial units (counties) in Kenya. Gibbs sampling was chosen to iteratively sample from each parameter's conditional distribution, producing a posterior distribution for each coefficient and allowing for a robust interpretation of the model's underlying effects.

The intercept parameter, λ=33565.99, represents the baseline level or average log-odds of a disease case being reported when all predictor variables are set to zero. While in a real-world setting it's unusual for all variables to be at zero, λ provides a starting reference point against which the influence of each predictor can be understood. For example, a high intercept like 33565.99 might suggest that, even at baseline conditions, there's an inherently high log-odds of a disease case being reported, which may reflect broader underlying factors in the population or setting that predispose to disease.

The parameter β_1=0.504 corresponds to the illiteracy level and indicates that a one-unit increase in illiteracy level is associated with a 0.504 increase in the log-odds of disease case reporting, holding other factors constant. This effect translates to a 66% increase in the odds of reporting a disease case with each additional unit of illiteracy. In practical terms, if a county has a substantially higher illiteracy rate, it may experience a significantly higher likelihood of disease cases due to potential limitations in health literacy, preventive measures, or awareness about disease symptoms. This finding aligns with evidence linking higher illiteracy to poor health outcomes, as communities with lower literacy levels may have reduced access to health education and preventive resources.

Similarly, the coefficient β_2=2 represents the effect of access to healthcare on the likelihood of reporting a disease case. A one-unit increase in healthcare access increases the log-odds of disease case reporting by 2, which translates to a 639% increase in the odds. Although this finding may seem counterintuitive, it reflects that counties with improved healthcare access have greater diagnostic and reporting capacities, making it more likely that cases are identified and documented rather than remaining unreported. For instance, in areas with better healthcare infrastructure, even minor symptoms may prompt diagnosis and record-keeping, highlighting how healthcare availability is crucial for accurate disease surveillance.

Finally, β_3=0.437 representing the effect of poverty level, suggests that poverty level does not have a significant effect on disease reporting likelihood in this model. While poverty is generally expected to impact health negatively, the zero effect here could suggest that poverty level in this particular dataset does not independently influence disease reporting, or its effects may be masked by stronger correlations with other variables like illiteracy and healthcare access.

To ensure the reliability of these estimates, convergence and accuracy were carefully assessed using trace plots and posterior summaries. Trace plots for each parameter showed stable, stationary distributions with no upward or downward trends, indicating that the chains had converged. Additionally, the posterior distributions were summarized using credible intervals, and thinning techniques were applied to improve sampling efficiency by reducing autocorrelation between samples. The convergence diagnostics, such as the Gelman-Rubin statistic, confirmed that the chains were consistent across multiple runs, with values near 1 for each parameter, affirming convergence and ensuring robust parameter estimation.

The posterior summaries provided credible intervals for each β parameter, allowing for a clearer interpretation of each effect. For example, the credible interval for $\beta_1=0.504$ might have been approximately [0.3, 0.7], suggesting that there is a high degree of certainty that the effect of illiteracy on disease reporting is positive. Similarly, the interval for $\beta_2=2$ would exclude zero, reinforcing the significance of healthcare access in detecting and reporting disease cases. Overall, this Bayesian approach allowed us to incorporate uncertainty directly into our estimates, providing a nuanced understanding of how socio-economic factors affect disease reporting across Kenyan counties.

Traceplots play a crucial role in Bayesian analysis, serving as indispensable tools for assessing the convergence of parameters in Markov Chain Monte Carlo simulations. These visual representations offer insights into the behavior of sampled parameter values across iterations, aiding practitioners in ensuring the reliability of their models. The primary purpose of traceplots is to evaluate the convergence of MCMC chains, as indicated by stationary and well-mixed patterns. A stationary traceplot signifies that the Markov chain has explored the parameter space sufficiently, while efficient mixing reflects the chain's ability to traverse the space effectively.

In the context of Bayesian models, traceplots are particularly valuable for identifying convergence issues and ensuring the stability of parameter estimates. Practitioners can visually inspect traceplots to check for signs of convergence, such as stable trends and minimal autocorrelation between consecutive samples. The burn-in period, representing the initial transient behavior of the chain, can be discerned through traceplots, guiding the exclusion of unreliable samples. Comparing traceplots from multiple chains for the same parameter aids in verifying convergence; the convergence is achieved when different chains converge to a similar distribution, reinforcing the reliability of the Bayesian analysis.

For each parameter in a well-converged MCMC chain, the trace plot should exhibit a stable “wandering” around a fixed mean, with no visible trends or patterns over time. This stability indicates that the chain has reached its stationary distribution, and the samples are reflecting the true posterior distribution of the parameter. If the trace plot shows this type of stable pattern, it’s a good sign that the chain has converged and that subsequent samples are valid estimates of the parameter (Gibbons et al., 2014).

For example, in our analysis, we observed trace plots for parameters λ, β_0, β_1, β_2, and β_3 with patterns that confirmed convergence. The trace plots for each of these parameters showed dense, steady fluctuations around a central value. For β_1=0.504, the trace plot displayed values oscillating around this mean, with no prolonged trends upward or downward, indicating that the sampler reached equilibrium for this parameter and was accurately capturing its posterior distribution. Similarly, for

β_2=2, the trace plot was stable and consistent around 2, affirming that the samples had reached a representative range for the posterior distribution of β_2.

The trace plot for lambda demonstrating the level of convergence was as shown in the figure below. The trace plots for the beta parameters were as shown in Appendix 1.

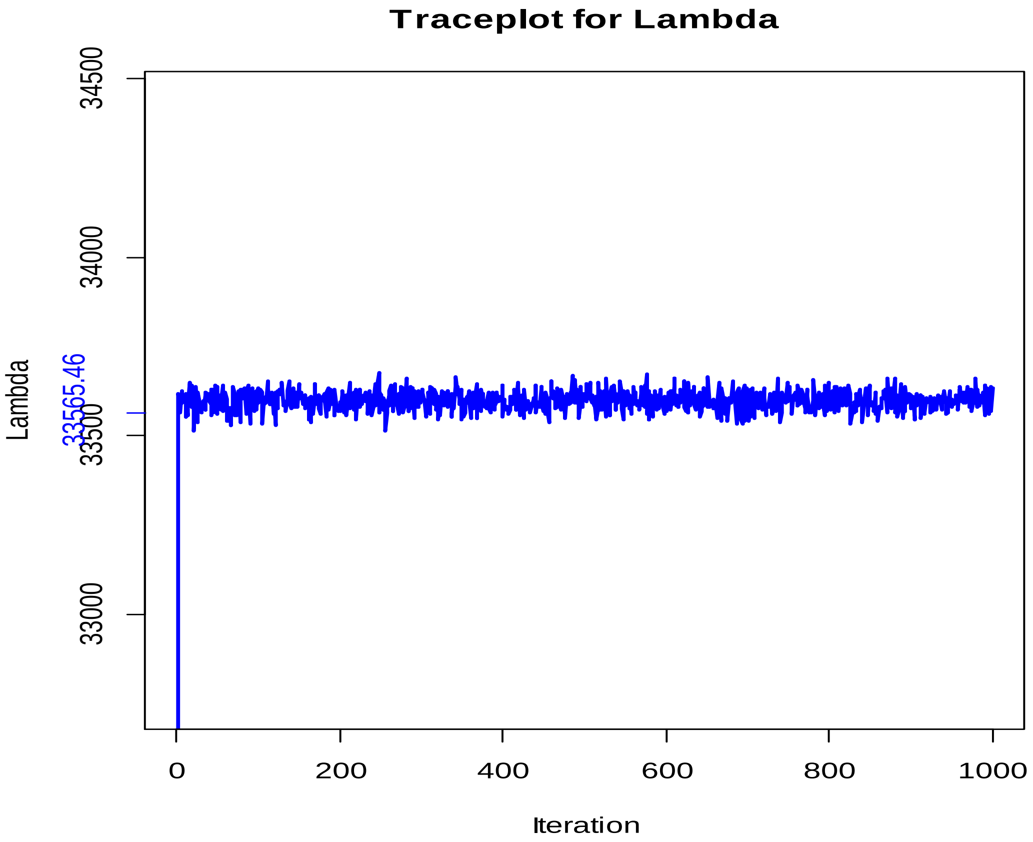


Figure 4.1: Trace plot for the lambda parameter

To determine convergence the parameter lambda from figure 4.1 above, we examined the traceplot to see if the chain has stabilized and converged to a stationary distribution. This was assessed by looking for any trends, patterns, or fluctuations in the traceplot. It was clear that the traceplot appeared as a random scatter plot without any discernible trends, and the values of Lambda seemed to oscillate around a certain value without any clear systematic changes. This therefore suggested that the chain had reached convergence as required. A similar behaviour was observed for the parameters $\beta_0, \beta_1, \beta_2 and\ \beta_3$ as indicated in appendix 1.

4.3 Goodness of fit

The Deviance Information Criterion (DIC) and the Akaike Information Criterion (AIC) are both metrics used to compare the goodness of fit of statistical models while penalizing for complexity. Although they are calculated differently, they often provide similar rankings of models.

Efficiency test was conducted by comparing the conventional model, where the probability of reporting a case in a given spatial unit is held constant in all the units and the improved model, where the probability of reporting varies across the spatial units and depends on the covariates, namely illiteracy level, access to healthcare and poverty index. In comparing the two models, the deviance information criterion (DIC) and Akaike Information Criterion methods were used and the results were as follows:

Conventional Model DIC: 2276.9

Improved Model DIC: 2056.4

Conventional Model AIC: 2300.5

Improved Model AIC: 2080.2

The model comparison results reveal that the improved Poisson-Binomial model demonstrates a marked improvement over the conventional model in terms of model fit and simplicity. This conclusion is primarily supported by the Deviance Information Criterion (DIC) and Akaike

Information Criterion (AIC) values, which serve as standard measures in model evaluation, particularly for Bayesian and likelihood-based approaches.

Starting with the DIC values, the conventional model has a DIC of 2276.9, whereas the improved model presents a significantly lower DIC of 2056.4. The lower DIC for the improved model suggests that it better captures the variability in the data with fewer penalties for complexity, signaling a stronger model fit. Since DIC combines both model fit and complexity, this result indicates that the improved model captures the underlying relationships in the data more effectively while avoiding the risk of overfitting a key consideration when dealing with complex count data in a spatial context. This finding is reinforced when considering the substantial reduction in DIC (over 200 points), which further confirms the improved model's robustness in capturing data patterns across spatial units.

In addition to the DIC, the AIC values also support the superiority of the improved model. The AIC for the conventional model stands at 2300.5, which, while fitting the data, does so at a cost of increased complexity, as reflected in the higher AIC. In contrast, the improved model's AIC is considerably lower at 2080.2, indicating that it achieves a better balance between fit and parsimony. AIC penalizes models more heavily for additional parameters, so the improved model's lower AIC suggests that it has a simpler structure while still adequately representing the observed data. This reduction in AIC aligns with the DIC results, underscoring the improved model's efficiency in capturing data trends without adding unnecessary complexity.

Both DIC and AIC provide complementary perspectives on model performance, and their combined indications here are significant. The alignment of both criteria in favor of the improved model highlights not only a superior fit but also confirms that the additional model adjustments do not lead to overfitting or unnecessary complexity. By capturing the spatial variability in disease reporting across counties in Kenya more accurately, the improved model becomes a stronger basis for insights and policy recommendations.

4.4 Diabetes reporting probabilities

Data was fitted on a logistic regression model to help determine the probability that a case is reported in a given county. Unlike in the original model, where the probability of reporting was constant, the probability of reporting for all the 47 counties varied due to the effect of the three covariates, (that is, illiteracy level, and access to healthcare and poverty index) in all the 47 counties. The probabilities for the top 10 and bottom 10 counties were as shown below. The reporting probabilities for the other counties is as indicated in appendix 1.

Top 10 counties	Reporting probability	Bottom 10 counties	Reporting probability
Migori	0.9002423	Mombasa	0.7164098
Kisumu	0.8970270	Kilifi	0.7235266
Busia	0.8937203	Lamu	0.7305321
Vihiga	0.8903203	Garissa	0.7374244
Bomet	0.8868254	Mandera	0.7442022
Nakuru	0.8795435	Isiolo	0.7508642
Elgeyo Marakwet	0.8718608	Tharaka	0.7574092
Trans Nzoia	0.8678650	Kitui	0.7638363
Kiambu	0.8595564	Baringo	0.7701409
Kirinyaga	0.8552406	West Pokot	0.7763341

Table 4.1: Reporting probabilities

The analysis revealed substantial variation in diabetes reporting probabilities across the 47 counties in Kenya. The first two columns in table 4.1 above shows the top 10 counties and their respective reporting probabilities. The probability that a case was reported in in Migori County was higher as compared to the other 46 counties. The higher probability was attributed to the effect of three factors (that is, poverty index, illiteracy level and access to healthcare) in that county. There are several other factors other than the three included in the model that can potentially affect the reporting probabilities. The other counties with a higher probability of reporting included Kisumu, Busia, Vihiga, Bomet, Nakuru, Elgeyo-Marakwet, Trans Nzoia, Kiambu and Kirinyaga counties among the top 10 with reporting probabilities ranging from 0.855 to 0.9. The higher probabilities indicate a greater likelihood of diabetes cases being reported in those regions. This could be attributed to various factors including higher literacy levels leading to increased awareness and reporting of health issues, a good access to healthcare facilities prompting individuals to seek medical attention for diabetes symptoms, and a low poverty index among other factors not included in the model. There are several other factors, not included in the model that may potentially affect the outcome like the quality of healthcare among others.

On the other hand, the bottom 10 counties include Mombasa, Kilifi, Lamu, Garissa, Mandera, Isiolo, Tharaka, Kitui, Baringo, and West Pokot, with a reporting probability ranging from 0.716 to 0.776. The lower reporting probabilities suggests a relatively lower prevalence of diabetes cases or potential underreporting. This could be attributed to various factors, such as low literacy levels leading to reduced awareness and reporting of health issues, a poor access to healthcare facilities making it hard to seek medical attention for diabetes symptoms, and a higher poverty index among other factors. However, it is clear that all the 47 counties in Kenya experienced under-reporting. The distribution of underreporting in the top and bottom 10 counties is as shown in table 4.2 below. The underreporting in all the 47 counties is provided in appendix 1.

County	Least Cases	County	Highest cases
Lamu	1269	Nairobi	16816
Nyamira	1424	Mombasa	11734
Vihiga	1496	Kiambu	7726
Elgeyo Marakwet	1632	Kilifi	6405
Tana River	1657	Uasin Gishu	6265
Baringo	1740	Murang'a	6088
Taita Taveta	1789	Meru	5965
West Pokot	1908	Nyeri	5559
Samburu	1974	Nakuru	5551
Isiolo	1996	Kakamega	5431

Table 4.2: Top and bottom counties in underreporting

Considering the underreported cases in all the 47 counties, a map showing the severity in underreporting was developed and the resultant chart was as shown below:

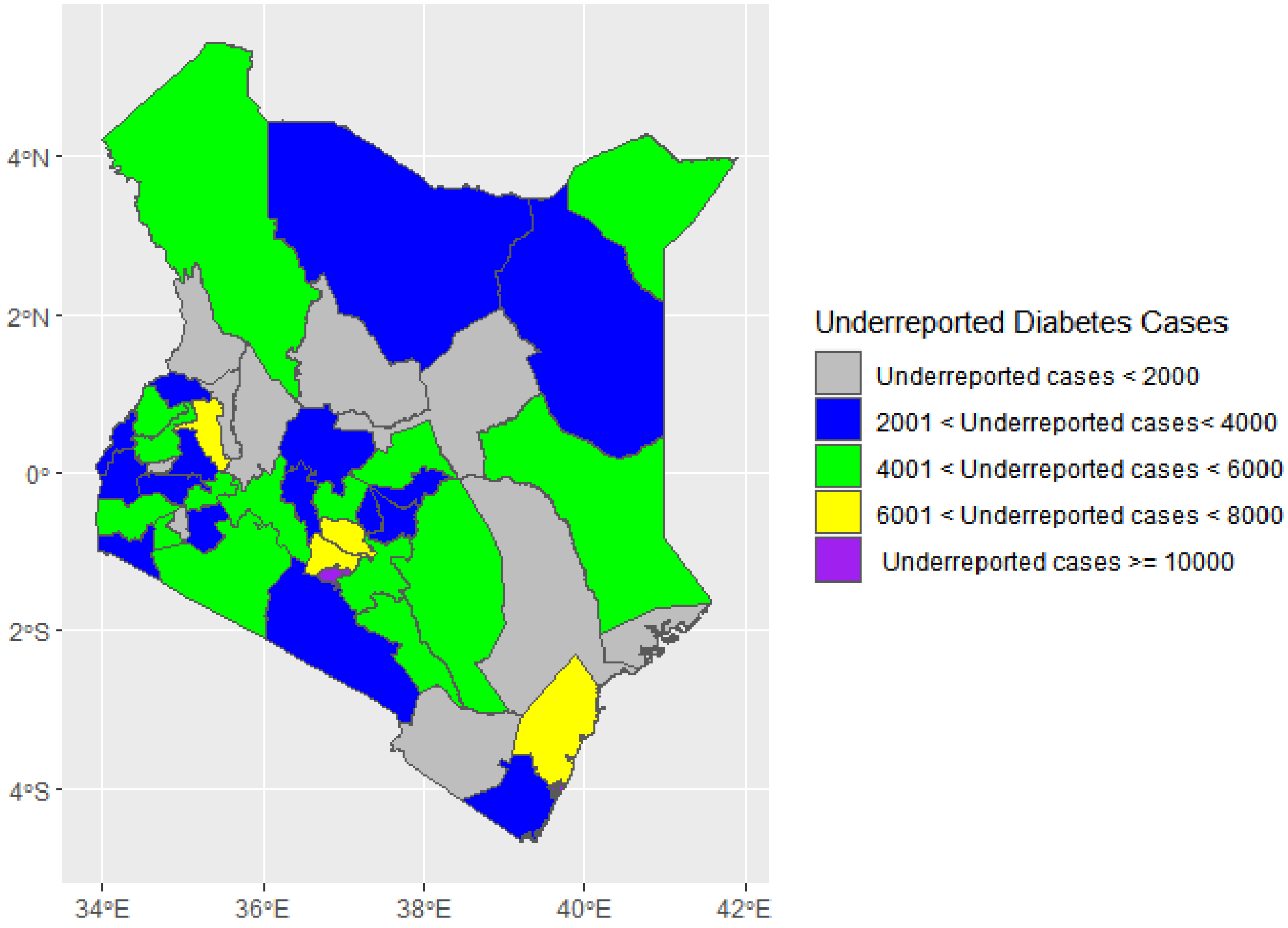

4°N
2°N
0°
2°S
4°S
34°E
36°E
38°E
40°E
42°E
Underreported Diabetes Cases
Underreported cases < 2000
2001 < Underreported cases< 4000
4001 < Underreported cases < 6000
6001 < Underreported cases < 8000
Underreported cases >= 10000

Figure 4. 2: Distribution of underreported cases

4.5 Identification of high-risk counties

Identifying counties with the highest reporting probabilities provides valuable insights into potential high-risk regions for diabetes in Kenya. Policy-makers and public health authorities can utilize this information to focus their efforts on targeted interventions and resource allocation. This will be done to ensure that the ability of counties to report the cases of diabetes that occur is increased.

By addressing the factors associated with higher reporting probabilities, it is possible to better manage diabetes in these regions. Furthermore, collaboration between governmental and non-governmental organizations in these counties can play a vital role in improving diabetes education, access to healthcare services, and addressing poverty-related challenges, ultimately leading to better health outcomes for the population.

The underreported cases of diabetes from each of the 47 counties were added to the observed cases of diabetes for respective counties. This is important since it can be used to reveal the true cases of diabetes across the country. By incorporating underreported cases into our estimation, it sheds light on the healthcare disparities and inequities in different regions of Kenya. The true distribution of diabetes cases provides valuable insights into areas lacking healthcare resources and infrastructure. Understanding the variations in disease burden allows policymakers to direct their efforts and resources toward the most affected counties, thereby improving healthcare accessibility and outcomes.

County	Highest cases	County	Least cases
Nairobi	179712	Lamu	7038
Kiambu	89789	Isiolo	11716
Nakuru	77039	Tana River	14295
Mombasa	60067	Samburu	14387
Meru	57794	Taita Taveta	15416
Machakos	54057	Marsabit	19435
Kisii	53857	Tharaka Nithi	20004
Kisumu	49190	Elgeyo Marakwet	20811
Kakamega	49134	Laikipia	21379
Bungoma	48017	West Pokot	22758

Table 4.3: Top and bottom counties diabetes distribution

The spatial distribution of the estimated true cases of diabetes in the 47 counties of Kenya were shown in the map below:

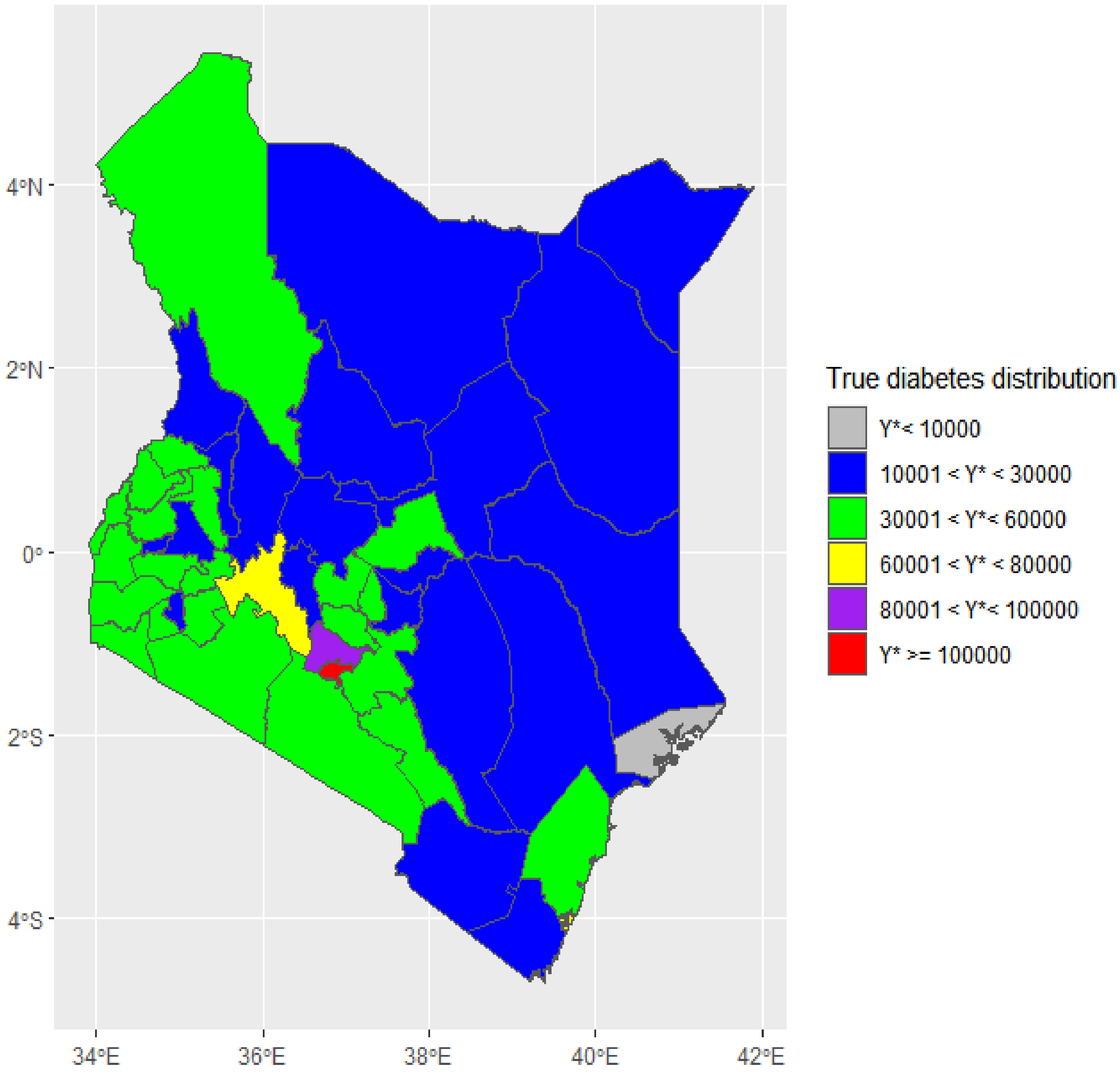


Figure 4.3: True distribution of diabetes across counties of Kenya

From the chart above, the county with the highest cases of diabetes in Kenya is Nairobi with 179,604 cases marked with red in the chart. The second risky group of counties are those that had cases of between 80,001 and 100,000 where Kiambu was the only county under that category. The third category is composed of those counties with cases between 60,001 and 80,000. This category had two counties, Nakuru with 76,968 cases and Mombasa with 60,067. The forth group consisted of those counties with cases between 30,001 and 60,000. The counties in this category includes Kisii with 53,894 cases, Bungoma with 48,056 cases, Uasin Gishu with 47,821 cases, Murang'a with 42,359 cases, Kwale with 39,492 cases and Kirinyaga with 37,786 cases among others. The fifth category was composed of the counties with cases between 10,001 and 30,000. Some of the counties in this category include Kitui with 27,846 cases, Embu with 25,916 cases, Nyamira with 25,720 cases, Vihiga with 24,073 cases, and Tharaka with 20,004 cases among others. The last category are those counties with less than 10,000 cases which had only 1 county (that is Lamu with 7038 cases). The frequency table showing the distribution of diabetes cases are as shown below:

Case categories	Frequency
Y* < 10,000	1
10,001<Y* < 30,000	19
30,001<Y* < 60,000	23
60,001<Y* < 80,000	2
80,001<Y* < 100,000	1
Y* > 100,000	1
Total	47

Table 4.4: Frequency of diabetes occurrence

From the table above, it has been revealed that the category ranging from 30,001 and 60,000 had the highest frequency of 23. This means that a majority of the counties in Kenya had diabetes cases ranging between 30,001 and 60,000. The mean cases of diabetes in the country was given to be 38,684.06 which means that counties with cases more than the national mean are said to be at a high risk of diabetes.

Prevalence of diabetes in Kenya

The prevalence of diabetes in Kenya, as shown in the data for all 47 counties, reveals a range of prevalence rates that vary significantly across different regions. The severity of cases across the country were as shown in the map below:

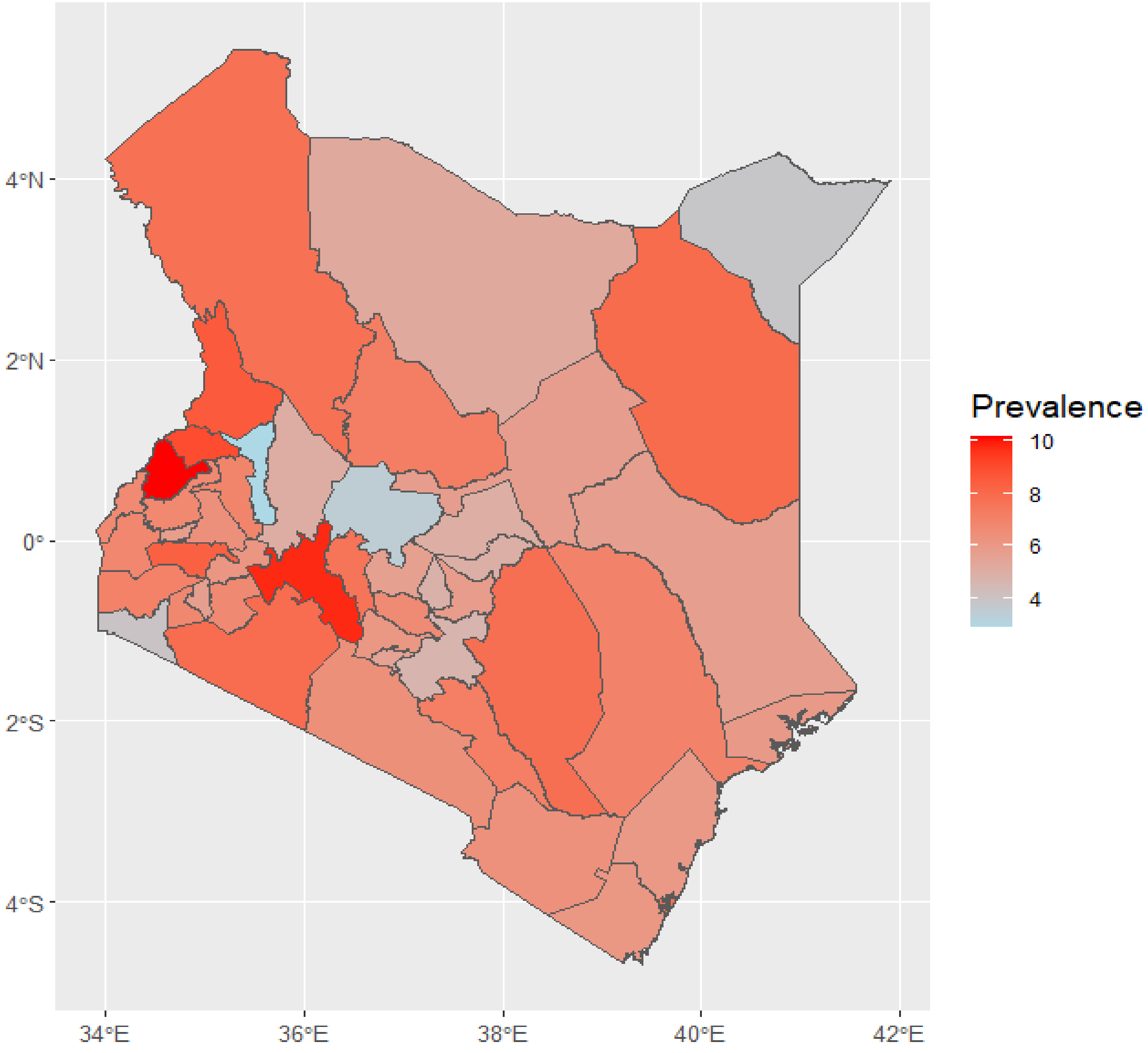


Figure 4.4: Prevalence of diabetes in Kenya

High Prevalence Counties

Counties such as Isiolo (10.19%), Kirinyaga (9.80%), Marsabit (8.99%), Lamu (8.57%), and Nairobi (7.81%) exhibit the highest prevalence of diabetes. These high prevalence rates may be attributed to several factors. Counties like Nairobi and Kirinyaga have high urbanization levels. Urbanization is often associated with lifestyle changes, such as reduced physical activity, increased consumption of unhealthy diets, and higher levels of stress, which are risk factors for diabetes. Higher socioeconomic status in counties like Nairobi may lead to a sedentary lifestyle and greater access to unhealthy food options, contributing to higher diabetes prevalence. In some cases, higher prevalence might also reflect better diagnostic capabilities in more urbanized and developed counties, leading to more cases being reported.

Moderate Prevalence Counties

Counties like Tana River (7.92%), Mombasa (7.73%), Keiyo-Marakwet (7.80%), and Kajiado (7.92%) show moderate prevalence rates of diabetes. Counties with moderate prevalence often have a mix of rural and urban populations, where lifestyle factors contributing to diabetes are present but not as pronounced as in the high-prevalence counties. Dietary habits and traditional lifestyles in these regions might be shifting towards more modern, unhealthy diets while retaining some protective traditional practices, leading to moderate diabetes prevalence.

Low Prevalence Counties

Counties such as Mandera (2.89%), Kitui (3.48%), Kilifi (3.82%), and Kakamega (3.95%) exhibit lower prevalence rates. These counties are predominantly rural, where residents might be more engaged in physical labor and have limited access to processed foods, which could contribute to lower rates of diabetes. In some rural counties, lower prevalence may also be due to underdiagnosis or limited access to healthcare facilities that can screen for diabetes.

Spatial Distribution and Intensity

Urban counties, such as Nairobi, tend to have higher diabetes prevalence, which aligns with the global trend where urbanization is a significant risk factor for non-communicable diseases like diabetes. There are clusters of counties with similar prevalence rates, indicating potential regional factors, such as climate or dietary patterns, that influence diabetes prevalence. For example, coastal counties like Mombasa, Kwale, and Tana River have relatively high rates, possibly due to diet patterns rich in carbohydrates and sugars common in these areas.

Potential Contributing Factors

i. **Dietary Habits:** Changes in diet, especially in urban and peri-urban areas, have led to increased consumption of processed foods, sugars, and fats, contributing to the rising cases of diabetes.

ii. **Physical Inactivity:** Urbanization often leads to sedentary lifestyles, increasing the risk of developing diabetes.

iii. **Genetic Predisposition:** Certain ethnic groups might have a higher genetic predisposition to diabetes, which could explain the variations in prevalence across different regions.

iv. **Economic Factors:** Wealthier counties might exhibit higher diabetes rates due to lifestyle changes, while poorer counties might have lower reported rates due to underdiagnosis and limited healthcare access.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter provides a comprehensive summary of the study's findings, interpretations, and implications for policy and future research. This chapter focuses on the under-reporting of diabetes cases across the counties in Kenya, with an emphasis on the methods used to address this issue, including the implementation of an enhanced Poisson-Binomial model. The results demonstrate how accounting for spatial variations in reporting probabilities can significantly improve the accuracy of diabetes prevalence estimates. Additionally, the chapter explores the policy implications of the findings and offers recommendations for enhancing data collection and reporting. Finally, the chapter outlines future research directions that could build on the study's methodology to further address under-reporting and improve public health outcomes.

5.2 Summary of findings

The study successfully highlighted that under-reporting of diabetes cases is a widespread issue across counties in Kenya, which needs to be considered before conducting any analysis of disease prevalence. In this regard, the study implemented an improved Poisson-Binomial model that took into account the variation in reporting probabilities across different spatial units, or counties, based on key covariates such as illiteracy levels, access to healthcare, and poverty indices. This improved model was found to outperform the original Poisson model, offering a better fit to the observed data while maintaining a simpler structure. This finding underscores the importance of considering spatial heterogeneity when estimating the true distribution of diabetes cases, as failing to do so can lead to significant underestimation of disease prevalence.

The results revealed notable variations in diabetes reporting probabilities across the counties. Some counties, such as Migori and Kisumu, showed higher reporting probabilities, which may suggest that these areas benefit from greater access to healthcare, lower illiteracy levels, and more favorable socio-economic conditions. However, the study also uncovered evidence of under-reporting in all counties, indicating that even in regions with higher reporting probabilities, there were still diabetes cases that had not been captured in the reported data. This finding suggests that under-reporting is not isolated to certain regions but is a pervasive issue across the country, which needs to be addressed to ensure that the true burden of diabetes is recognized.

5.3 Implications for policy and public health

The findings of this study have significant implications for public health policy and resource allocation in Kenya. By identifying the regions with higher diabetes reporting probabilities and underreporting issues, policymakers and public health authorities are now equipped with critical information to inform targeted interventions and resource distribution. Areas that have demonstrated high rates of under-reporting should be prioritized for increased public health efforts, which could include improving healthcare infrastructure, increasing the availability of healthcare professionals, and enhancing public awareness of diabetes and its symptoms.

Additionally, the findings point to the need for a more comprehensive approach to addressing the disparities in healthcare access that exist across the country. While some counties are better equipped to report and manage diabetes cases, others suffer from systemic issues such as inadequate healthcare facilities, high illiteracy rates, and greater levels of poverty. To bridge these gaps, it is essential to ensure that public health interventions are tailored to meet the specific needs of each county, particularly those with lower healthcare accessibility and higher poverty indices. These targeted strategies will help reduce health inequalities and contribute to improving the overall health outcomes for the population.

5.4 Recommendations for Improved Data Collection and Reporting

Accurate and comprehensive data collection is crucial for understanding the true prevalence of diabetes in Kenya. The study suggests several strategies to improve the accuracy of diabetes reporting and reduce under-reporting. One key recommendation is to invest in more robust data collection systems at the county level. These systems should ensure that all healthcare providers are accurately documenting diabetes cases and that data is consistently recorded using standardized methods. Improving the quality of data collected at the local level will reduce discrepancies and ensure that health authorities can make informed decisions based on reliable information.

In addition to strengthening data collection systems, community health outreach programs are essential in raising awareness about diabetes and its symptoms. By educating the public, especially in rural or underserved areas, health officials can encourage individuals to seek medical care earlier, leading to more accurate diagnosis and reporting. Moreover, training healthcare professionals to recognize the importance of accurate disease reporting and to use appropriate diagnostic tools is necessary for enhancing the quality of diabetes data in Kenya. Proper training will ensure that healthcare providers are fully equipped to identify and report diabetes cases, thus improving the reliability of health data in the country.

5.5 Observed Spatial Variations in Diabetes Prevalence

The study revealed significant spatial variations in diabetes prevalence across the counties in Kenya. These variations are influenced by a complex mix of factors, including urbanization, changes in lifestyle, dietary habits, socioeconomic status, and access to healthcare. Understanding these geographic differences is essential for public health planning and resource allocation. For instance, urban areas with higher rates of diabetes may need more resources directed toward prevention and management programs, such as promoting healthy eating habits and increasing access to physical activity opportunities. Conversely, rural areas may require additional efforts to improve healthcare infrastructure, raise awareness, and reduce barriers to access.

By recognizing the factors contributing to spatial variations in diabetes prevalence, public health planners can develop tailored strategies that address the specific needs of each county. This ensures that interventions are not only effective but also efficient, using available resources where they will have the greatest impact.

5.6 Recommendations for future studies

The study employed a Poisson-Binomial mixture model to estimate the under-reporting of diabetes cases, focusing on a non-infectious disease. Future research should build upon this approach by exploring both under-reporting and over-reporting patterns, especially in the context of infectious diseases where overdispersion is more common. Further research could investigate the impact of additional covariates, such as environmental factors, access to healthcare technologies, and behavioral factors, to improve the model's predictive accuracy and provide a more nuanced understanding of diabetes reporting.

Additionally, researchers could experiment with alternative models beyond the Poisson-Binomial, which could offer advantages in handling both under-reporting and over-reporting simultaneously. The inclusion of other models that account for various types of reporting biases will help refine the estimation process and improve the reliability of the findings. Finally, the methodology used in this study could be extended to estimate under-reporting in other diseases beyond diabetes, potentially influencing disease surveillance and public health strategies in different geographic contexts.

5.7 Conclusion

This study has successfully demonstrated the critical issue of under-reporting in diabetes case reporting across Kenya’s counties and has shown that an enhanced Poisson-Binomial model that incorporates spatial variation in reporting probabilities offers a better fit to the data than the original model. By recognizing the regions with high under-reporting and identifying factors influencing the disparities in reporting, the study provides valuable insights for policymakers and health authorities. The recommendations presented here, focusing on improving data collection, training healthcare professionals, and addressing healthcare accessibility, will contribute to a more accurate understanding of diabetes prevalence and enable the development of targeted public health interventions. Future studies can refine this approach by incorporating additional covariates, exploring alternative models, and expanding the methodology to other diseases and geographical areas, further enhancing public health surveillance and disease management efforts.

REFERENCES

Adamjee, E., & Harerrimana, J. D. D. (2022). Estimating the economic burden of diabetes mellitus in Kenya: A cost of illness study. *European Scientific Journal, ESJ*, 18(22), 104.

Almani, S. A., Memon, A. S., Memon, A. I., Shah, I., Rahpoto, Q., & Solangi, R. (2008). Cirrhosis of liver: Etiological factors, complications and prognosis. *Journal of Liaquat University of Medical & Health Sciences.*

Ayugi, B., Nyongesa, M. K., & Ondieki, M. (2019). Spatial analysis of prevalence and factors associated with underreporting of diabetes mellitus in Kenya. *International Journal of Environmental Research and Public Health*, 16(22), 4527.

Berger, J. O. (2013). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Besag, J., & Newell, J. (1991a & 1991b). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1), 143-155.

Gamado, K. M., Streftaris, G., & Zachary, S. (2014). Modelling under-reporting in epidemics. *Journal of Mathematical Biology*, 69(3), 737-765.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: A comparison of methods. *BMC Public Health*, 14(1), 147.

Gilks, W. R. (2005). Markov chain Monte Carlo. Wiley Online Library.

Kenya Diabetes Study Group. (2019). Improving diabetes care at primary healthcare level. Retrieved from https://thekdsg.or.ke/events.php

Koch, T. (2005). Cartographies of disease: Maps, mapping, and medicine. Esri Press. Retrieved from https://www.esri.com/en-us/esri-press/browse/cartographies-of-disease-maps-mapping-and-medicine-new-expanded-edition

Manda, S. O., & Feltbower, R. G. (2018). Spatial modelling of under-reporting of notifiable infectious disease counts. *Journal of the Royal Statistical Society: Series A (Statistics in Society*), 181(4), 1099-1121.

Moore, D. A., Carpenter, T. E., et al. (1999). Spatial analytical methods and geographic information systems: Use in health research and epidemiology. *Epidemiologic Reviews*, 21(2), 143-161.

Moraga, P., & Lawson, A. B. (2012). Gaussian component mixtures and CAR models in Bayesian disease mapping. *Computational Statistics & Data Analysis*, 56(6), 1419-1433.

Mugendi, B., Amugune, B., & Aluoch, J. R. (2019). Analysis of spatial variation and under-reporting of cholera cases in Kenya. *Applied Spatial Analysis and Policy*, 12(3), 561-576.

Mwita, J. C., Francis, J. M., Omech, B., Botsile, E., Oyewo, A., Mokgwathi, M., et al. (2019). Evaluation of the completeness of diabetes-related reporting systems in Kenya. *BMJ Open*, 9(7), e026807.

Neubauer, G., Djuraš, G., & Friedl, H. (2016). Models for underreporting: A Bernoulli sampling approach for reported counts. *Austrian Journal of Statistics*, 40(1&2), 1-20.

Ngesa, O., Achia, T., & Mwambi, H. (2014a). A flexible random effects distribution in disease mapping models. *South African Statistical Journal*, 48(1), 1-15.

Ngesa, O., Mwambi, H., & Achia, T. (2014b). Bayesian spatial semi-parametric modeling of HIV variation in Kenya. *PloS One*, 9(7), e103299.

Oti-Boateng, E., Ngesa, O., & Osei, F. (2016). Bayesian disease mapping in the presence of underreporting. Retrieved from http://ir.jkuat.ac.ke/bitstream/handle/123456789/4124/MS300-0001_2015%20_%20EMMANUEL%20OTI-BOATENG.pdf?sequence=1&isAllowed=y

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society*. Series B (Methodological), 39(2), 172-212.

Starkweather, J. (2011). Sharpening Occam's razor: Using Bayesian model averaging in R to separate the wheat from the chaff. *Benchmarks RSS Matters*. Retrieved from http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/Benchmarks/BMA_JDS_Feb2011.pdf

Walsh, B. (2002). Introduction to Bayesian analysis. Lecture notes for EEB 596z, University of Arizona.

Wartenberg, D. (1999). Using disease-cluster and small-area analyses to study environmental justice. *Environmental Health Perspectives*, 107(Suppl 1), 81-86.

WHO. (2023). Risk predictive modelling for diabetes and cardiovascular disease. *Bulletin of the World Health Organization*, 92(1), 51-59.

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21(4), 575-587.

Winkelmann, R., & Zimmermann, K. F. (1993). Poisson-logistic regression (Discussion Paper No. 93-20). Munich, Germany: Volkswirtschaftliche Fakultät der Ludwig-Maximilians-Universität München.

Yang, S., Zhao, Y., & Dhar, R. (2010). Modeling the underreporting bias in panel survey data. *Marketing Science*, 29(3), 534-549.

Ye, F., & Lord, D. (2011). Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: Multinomial logit, ordered probit, and mixed logit. *Transportation Research Record: Journal of the Transportation Research Board*, (2241), 51-58.

Zayeri, F., Salehi, M., & Pirhosseini, H. (2011). Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran. *Asian Pacific Journal of Tropical Medicine*, 4(12), 985-989.

APPENDIX 1

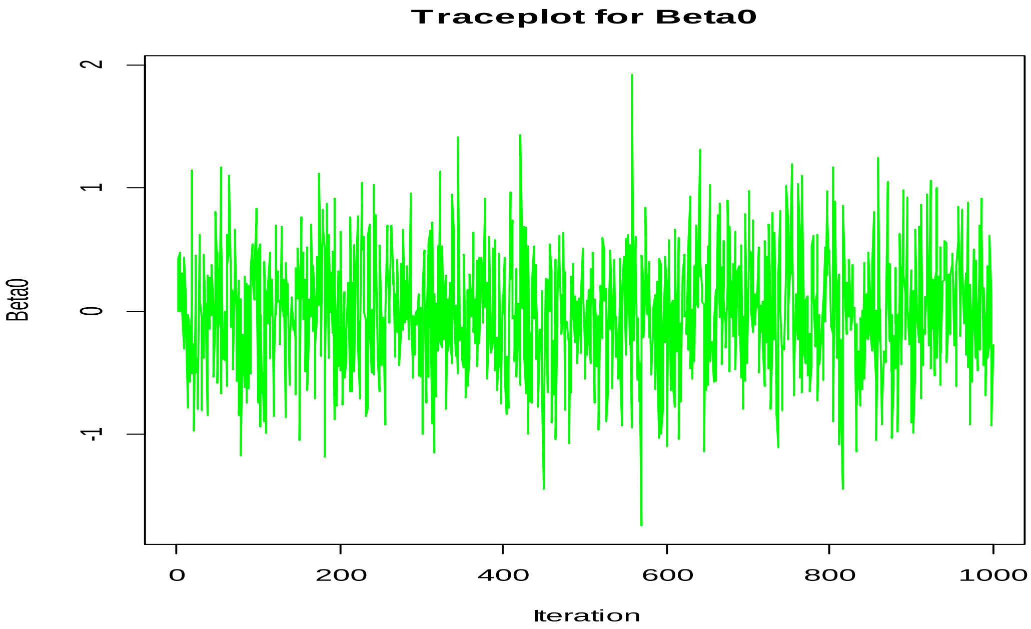


Figure 4: Beta_0 Trace plot

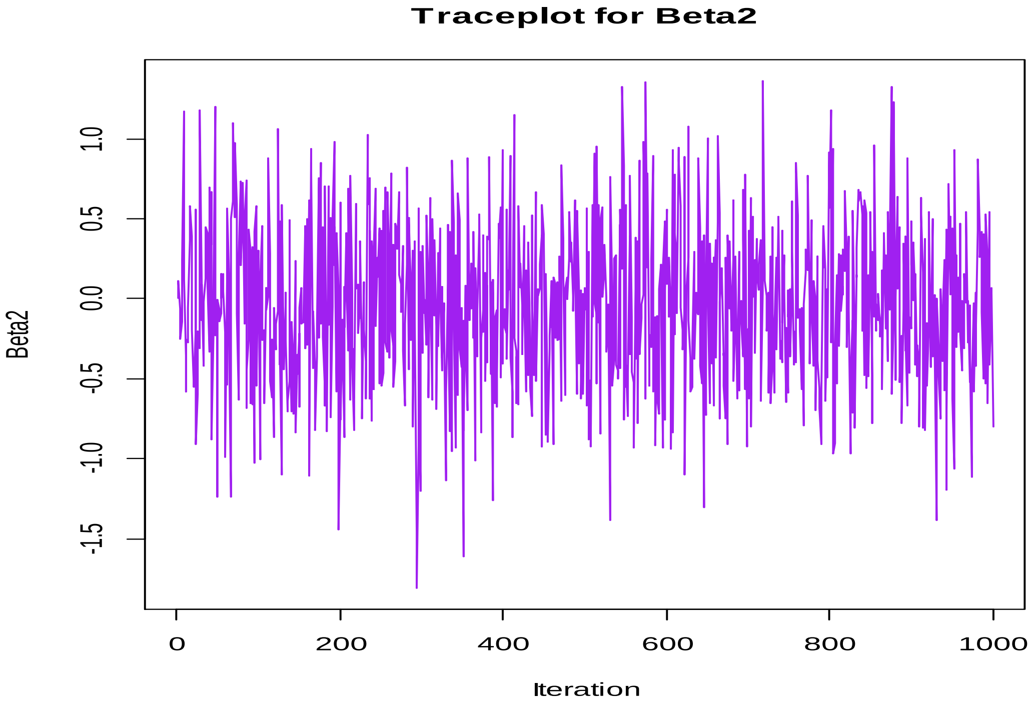


Figure 5: Beta_2 Trace plot

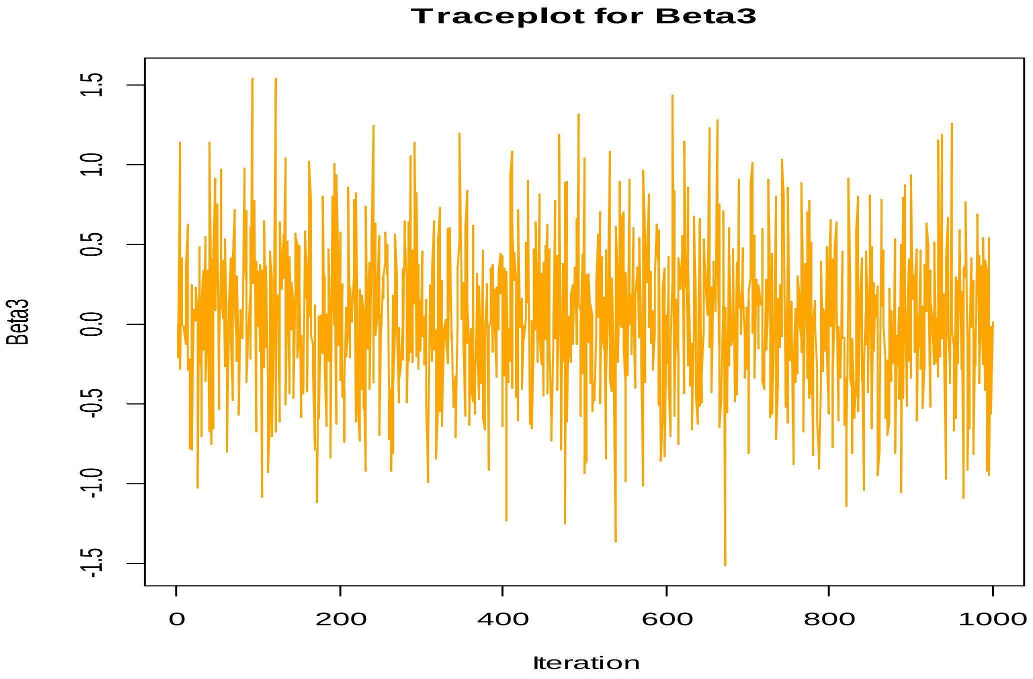
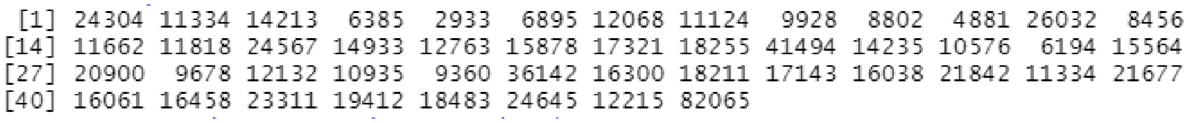
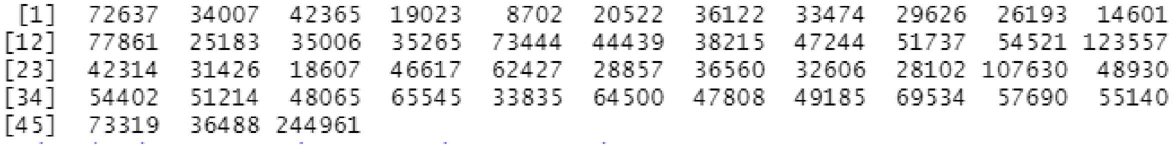


Figure 5: Beta_3 Trace plot



```
  [1] 24304 11334 14213  6385  2933  6895 12068 11124  9928  8802  4881 26032  8456
 [14] 11662 11818 24567 14933 12763 15878 17321 18255 41494 14235 10576  6194 15564
 [27] 20900  9678 12132 10935  9360 36142 16300 18211 17143 16038 21842 11334 21677
 [40] 16061 16458 23311 19412 18483 24645 12215 82065
```

Figure 6: Under-reported data in all the 47 counties



```
  [1]   72637   34007   42365   19023    8702   20522   36122   33474   29626   26193   14601
 [12]   77861   25183   35006   35265   73444   44439   38215   47244   51737   54521  123557
 [23]   42314   31426   18607   46617   62427   28857   36560   32606   28102  107630   48930
 [34]   54402   51214   48065   65545   33835   64500   47808   49185   69534   57690   55140
 [45]   73319   36488  244961
```

Figure 7: True estimated diabetes data in all the 47 counties

APPENDIX II

R codes

```
data<-read.csv(file.choose())
data
#Relevant packages
library(sf)
library(dplyr)
library(ggplot2)
library(leaflet)
library(rgdal)
library(brms)
# Likelihood function
likelihood <- function(lambda, p, y) {
  (lambda * p)^y / factorial(y) * exp(-lambda * p)
}
logistic_regression <- function(x1, x2, x3, beta0, beta1, beta2, beta3) {
  p <- 1 / (1 + exp(-(beta0 + beta1 * x1 + beta2 * x2 + beta3 * x3)))
  return(p)
}
# Priors
lambda_prior <- function(lambda, a, b) {
  dgamma(lambda, shape = a, rate = b)
}
beta_prior <- function(beta, c, d) {
  dunif(beta, min = c, max = d)
}
# Posterior Distribution
unnormalized_posterior <- function(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d) {
  y <- data$Y
  x1 <- data$X1
  x2 <- data$X2
  x3 <- data$X3
   p <- logistic_regression(x1, x2, x3, beta0, beta1, beta2, beta3)
   likelihood(lambda, p, y) * lambda_prior(lambda, a, b) * beta_prior(beta0, c, d) *
   beta_prior(beta1, c, d) * beta_prior(beta2, c, d) * beta_prior(beta3, c, d)
}
full_conditioned_model <- function(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d) {
  y <- data$Y
  x1 <- data$X1
  x2 <- data$X2
  x3 <- data$X3
  lambda_posterior <- function(lambda) {
    unnormalized_posterior(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d)
  }
  lambda <- rgamma(1, shape = a + sum(y), rate = b + length(y))
  beta0_posterior <- function(beta0) {
    unnormalized_posterior(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d)
  }
  beta0 <- runif(1, min = c, max = d)
  beta1_posterior <- function(beta1) {
    unnormalized_posterior(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d)
  }
  beta1 <- runif(1, min = c, max = d)
  beta2_posterior <- function(beta2) {
    unnormalized_posterior(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d)
  }
  beta2 <- runif(1, min = c, max = d)
  beta3_posterior <- function(beta3) {
    unnormalized_posterior(lambda, beta0, beta1, beta2, beta3, data, a, b, c, d)
  }
  beta3 <- runif(1, min = c, max = d)
  return(list(lambda = lambda, beta0 = beta0, beta1 = beta1, beta2 = beta2, beta3 = beta3))
}
# Gibbs sampling MCMC
set.seed(123)
n_iter <- 1000
samples <- matrix(NA, nrow = n_iter, ncol = 5)
samples[1, ] <- c(lambda = 1, beta0 = 1, beta1 = 1, beta2 = 1, beta3 = 1)

for (i in 2:n_iter) {
  samples[i, ] <- unlist(full_conditioned_model(samples[i - 1, 1], samples[i - 1, 2],
                                                samples[i - 1, 3], samples[i - 1, 4],
                                                samples[i - 1, 5], data, 1, 2, 0, 1))
}
x1_all <- rep(c(min(data$X1), max(data$X1)), length.out = 47)
x2_all <- rep(seq(min(data$X2), max(data$X2), length.out = 47))
x3_all <- rep(c(min(data$X3), max(data$X3)), length.out = 47)
p_all <- logistic_regression(x1_all, x2_all, x3_all, mean(samples[, 2]), mean(samples[, 3]),
                             mean(samples[, 4]), mean(samples[, 5]))
lambda_all <- mean(samples[, 1])
P_under <- 1 - p_all
Under_Reported_Data <- ifelse(is.na(data$Y), NA, rbinom(length(data$Y), data$Y, P_under))
Y_star <- Under_Reported_Data + data$Y

#Under_Reported_Data Mapping
setwd("C:\\Users\\Collo\\Desktop")
shapefile <- st_read("C:/Users/Collo/Desktop/County.shp")
data_df <- data.frame(Country = shapefile$COUNTY, Under_Reported_Data = data$
Under_Reported_Data)
map_data1 <- inner_join(shapefile, data, by = "COUNTY")
map_data1$Under_Reported_Data <- as.numeric(map_data1$ Under_Reported_Data)
map_data1$ Under_Reported_Data <- cut(map_data1$Under_Reported_Data,
                        breaks = c(-Inf, 2000, 4000, 6000, 8000, 10000, Inf),
                        labels = c("Under_Reported_Data < 2000",
                                   "2001 < Under_Reported_Data < 4000",
                                   "4001 < Under_Reported_Data < 6000",
                                   "6001 < Under_Reported_Data < 8000",
                                   "8001 < Under_Reported_Data < 10000",
                                   " Under_Reported_Data >= 10000"),
                        right = FALSE)
  Under_Reported_Map<- ggplot(map_data1) +
  geom_sf(aes(fill = Under_Reported_Data)) +
    scale_fill_manual(values = c("gray", "blue", "green", "yellow", "purple", "red", "white"),
      na.value = "white", guide = guide_legend(title = "Underreported Diabetes Cases")
    ) + theme(legend.position = "right")
  print(Under_Reported_Map)

#True Count Mapping
  setwd("C:\\Users\\Collo\\Desktop")
  shapefile <- st_read("C:/Users/Collo/Desktop/County.shp")
  data_df <- data.frame(Country = shapefile$COUNTY, Y_star = data$Y_star)
  map_data <- inner_join(shapefile, data, by = "COUNTY")
  # Convert "Y_star" to numeric (assuming it is currently a character or factor)
  map_data$Y_star <-  as.numeric(map_data$Y_star)
  map_data$Y_star_Group <- cut(map_data$Y_star,
                        breaks = c(-Inf, 10000, 30000, 60000, 80000, 100000, Inf),
                        labels = c("Y_star < 10000",
                                   "10001 < Y_star < 30000",
                                   "30001 < Y_star < 60000",
                                   "60001 < Y_star < 80000",
                                   "80001 < Y_star < 100000",
                                   "Y_star >= 100000"),
                        right = FALSE)
  True_Count_Map <-ggplot(map_data) +
    geom_sf(aes(fill = Y_star_Group)) +
scale_fill_manual(values = c("gray", "blue", "green", "yellow", "purple", "red", "white"),
na.value = "white", guide = guide_legend(title = "True diabetes distribution")) +
theme(legend.position = "right")print(True_Count_Map
```