VECTOR AUTOREGRESSION MODELING OF MALARIA INCIDENCE AND MORTALITY RATES IN MIGORI COUNTY, KENYA

CHACHA PAUL JACKSON

A THESIS SUBMITTED TO THE SCHOOL OF SCIENCE IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN BIOSTATISTICS OF THE DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, UNIVERSITY OF ELDORET, KENYA

2025

DECLARATION

Declaration by the Student

This 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 prior written permission of the author and/or University of Eldoret.

Jackson Chacha

______________________________ **Date**________________________

SSCI/MAT/M/004/21

Approval by the Supervisors

This thesis has been submitted with our approval as the university supervisors.

______________________________ **Date**________________________

Prof. Argwings Otieno
Department of Mathematics and Computer Science,
University of Eldoret,
ELDORET, KENYA.

______________________________ **Date**________________________

Dr. Julius Koech, PhD.
Department of Mathematics and Computer Science,
University of Eldoret,
ELDORET, KENYA.

DEDICATION

To my son and my beautiful daughter.

ACKNOWLEDGEMENTS

I am thankful to Almighty God, the giver of life, for granting me life, knowledge, and wisdom, which enable me to pursue this research. I also appreciate my supervisors, whose guidance has been instrumental in helping me develop this proposal.

ABSTRACT

Malaria prevalence in poorer countries has been a persistent public health concern, disproportionately affecting vulnerable populations such as children and pregnant women. Despite notable progress in scaling up malaria control interventions in Kenya, malaria incidence rates continue to vary widely across counties, with endemic regions like Migori County experiencing persistent challenges. This study aimed to identify key factors associated with malaria incidence and mortality in Migori County using secondary data from the Kenya National Health Management System. Multiple statistical models, including regression, Vector Autoregression (VAR), and Vector Autoregression with Exogenous Variables (VARX), were applied to examine the temporal dynamics of malaria. While malaria incidence rates declined over time, mortality rates remained relatively stable. Regression results indicated that insecticide-treated net usage and effective treatment significantly influenced both incidence and mortality rates. However, model residuals showed substantial variability and signs of poor fit, highlighting the need for improved model specifications. The VAR model revealed issues of residual autocorrelation, while the VARX model, which incorporated exogenous variables, showed improved but still imperfect performance. Bayesian VAR (BVAR) models provided consistent findings across methodologies but also underscored ongoing challenges in modeling temporal malaria data accurately. Therefore, this study concludes that while current models offer valuable insights, they remain limited in capturing the full complexity of malaria dynamics. It recommends methodological enhancements, such as using advanced techniques like Generalized Method of Moments (GMM) or machine learning, conducting rigorous residual diagnostics, and incorporating environmental, socio-economic, and behavioral variables. Expanding the dataset across regions and timeframes could also improve the robustness and generalizability of future research aimed at informing more effective malaria control strategies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

ACF: Autocorrelation Function

AMET: Anti-Malaria Effective Treatment

ANOVA: Analysis of Variance

ITN: Insecticide Treated Net

VAR: Vector Autoregression

VARX: Vector Auto Regression with Exogeneous Variables.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Malaria remains a critical global public health challenge, with its burden most heavily felt in tropical and subtropical regions. According to the World Health Organization (WHO), there were approximately 241 million cases of malaria and 627,000 related deaths globally in 2020, with Sub-Saharan Africa accounting for 95% of all cases and 96% of deaths (WHO, 2021). This disproportionate burden underscores the ongoing vulnerability of African populations, particularly children under the age of five and pregnant women, who represent the majority of malaria-related morbidity and mortality. East African countries—including Kenya, Uganda, and Tanzania—face particularly high transmission rates, driven by favorable climatic conditions such as warm temperatures and high humidity, which support the breeding and survival of the *Anopheles* mosquito, the primary vector for malaria (Chaves *et al*., 2012; Djihinto *et al*., 2022).

In Kenya, malaria continues to be one of the leading causes of outpatient attendance and hospital admissions, with approximately 70% of the population at risk of infection (Ministry of Health, 2021). National malaria prevalence among children aged 6–59 months was 6% in 2020, but regional disparities are significant. For instance, Migori County—located in the Lake Victoria Basin where environmental conditions favor year-round mosquito breeding—reported a prevalence rate of 27.0% among children under five in the same year (Guissou *et al*., 2023). This is nearly five times higher than the national average, highlighting the urgent need for localized analysis and intervention.

Despite extensive investment in control measures—such as insecticide-treated nets (ITNs), indoor residual spraying (IRS), and prompt access to effective antimalarial treatment—the effectiveness of these interventions varies across counties. According to the Kenya Malaria Indicator Survey (2020), ITN ownership in Migori County stood at 83%, while usage among children under five was only 67%, indicating a significant gap between ownership and actual use. Nationally, ITN usage among children in the same age group averaged 56%, suggesting that while Migori fares better than average in ITN access, proper utilization remains a concern. Similarly, access to effective malaria treatment within 24 hours of symptom onset is suboptimal, especially in rural and hard-to-reach areas (Ototo *et al.*, 2022).

This study is therefore justified by the pressing need to better understand the local dynamics of malaria incidence and mortality in high-burden areas such as Migori County. Traditional epidemiological methods have contributed to important progress in malaria control, but they often fall short in capturing the complex temporal interactions among factors influencing malaria outcomes. To bridge this gap, the present study employs Vector Autoregression (VAR) models with Exogenous Variables (VARX) to analyze malaria incidence and mortality over time. These models are well-suited for understanding interdependence and feedback effects between multiple time series variables—such as past incidence and mortality rates—and how they are influenced by exogenous factors like ITN usage, ITN access, prevalence rates, and effective treatment coverage.

The specific objectives of this study include modeling the temporal spread of malaria incidence and mortality in Migori County using VAR and VARX approaches, assessing the significance of exogenous intervention variables, estimating the effects of these interventions on malaria outcomes, evaluating the model fit, and forecasting future trends. By incorporating exogenous variables into the models, the study offers a data-driven understanding of what drives malaria persistence in the region. The insights gained will support more targeted public health interventions, improve resource allocation, and contribute to Kenya’s broader malaria control and elimination agenda. Moreover, the findings will enhance the global understanding of malaria dynamics in high-risk tropical settings and inform strategies beyond the Kenyan context.

1.2 Statement of the problem

Malaria remains a significant public health concern in Kenya, particularly in Migori County, which is situated in the malaria-endemic Lake Victoria region. Despite national and global efforts to combat the disease, Migori County continues to experience high malaria incidence and mortality rates. According to the Kenya Malaria Indicator Survey 2020, the Lake Victoria region, which includes Migori County, reported a malaria prevalence of 19% among children aged 6-59 months, significantly higher than the national average of 5.6%. This persistent burden underscores the need for more effective strategies to predict and control malaria transmission in the region.

Current malaria control efforts in Migori County face challenges due to the complex interplay of factors influencing disease transmission, including climatic conditions, socioeconomic factors, and healthcare access. Traditional time series analysis methods often fail to capture the dynamic relationships between these variables and malaria incidence and mortality rates. Vector Autoregression (VAR) modeling offers a promising approach to address this gap by allowing for the simultaneous analysis of multiple time series variables and their interdependencies. However, the application of VAR modeling to malaria epidemiology in Migori County remains unexplored, presenting a critical knowledge gap in local malaria control efforts.

The VAR model has mostly been used by researchers to numerically describe the behavior of a variable over time (Al-hajj *et al.*, 2017; Sharma *et al.*, 2018 and Zhou *et al.*, 2024). Therefore, in a VAR model of order k, VAR (k), each component of vector Xt is linearly related to its own lagged values up to k periods as well as on the lagged values of all other variables up to lag k. Furthermore, the Vector Autoregression Model with Exogenous Variables (VARX) has been extensively used for predicting malaria incidence and mortality using time series data.

The lack of a comprehensive, data-driven forecasting model for malaria incidence and mortality in Migori County hinders effective resource allocation and intervention planning. According to the Migori County Health Department's 2021 annual report, malaria accounts for approximately 30% of all outpatient visits and 20% of hospital admissions in the county. These statistics highlight the substantial burden on the local healthcare system and emphasize the urgent need for improved predictive tools. By conducting a study using VAR modeling, researchers can provide valuable insights into the temporal dynamics of malaria transmission in Migori County, potentially leading to more targeted and efficient malaria control strategies.

1.3 Significance of the Study

This study is of paramount importance due to its potential to revolutionize how malaria epidemiology is understood and addressed in high-burden regions. By employing vector autoregression (VAR) modeling, this research provides a sophisticated analytical framework that captures the intricate temporal relationships between malaria incidence and mortality rates. This approach offers a more comprehensive view of how fluctuations in one variable can influence the other over time, leading to a deeper understanding of the dynamics driving malaria transmission and outcomes. Such insights are crucial for tailoring interventions that address the specific needs of Migori County and similar areas.

The findings from this study have direct implications for public health strategies in Migori County. By elucidating the temporal patterns and causal relationships between malaria incidence and mortality, the research can inform more effective and responsive control measures. This is particularly important in a region where seasonal and environmental factors heavily influence malaria transmission. The ability to predict trends and anticipate the impact of various interventions will enable health authorities to allocate resources more efficiently, implement timely preventive measures, and ultimately reduce the disease burden more effectively.

Moreover, this study contributes significantly to the global body of knowledge on malaria epidemiology. By demonstrating the application of VAR modeling in a real-world context, it provides a valuable methodological contribution that could inspire similar studies in other endemic regions. The insights gained can enhance the understanding of malaria dynamics on a broader scale, supporting the development of innovative strategies to combat malaria worldwide. As such, this research not only addresses local public health challenges but also contributes to the global effort to reduce malaria's impact and move towards eradication.

1.4 Objectives

This study shall be guided by the following research objectives.

1.5 General Objective

The general objective of the study is to analyze malaria incidence and mortality rates in Migori County, Kenya, using Vector Autoregression Models with Exogenous Variables.

1.5.1 Specific Objectives

The specific objectives shall be;

i. To analyze the temporal trends of malaria incidence and mortality using Vector Autoregression Models with Exogenous Variables in Migori County, Kenya.
ii. To estimate the effects of ITN use, ITN access, effective anti-malaria treatment, and malaria infection prevalence on malaria incidence and mortality rates in Migori County, Kenya.
iii. To assess the goodness of fit of the model for malaria incidence and mortality rates in Migori County, Kenya.
iv. To predict malaria incidence and mortality in Migori County, Kenya using ITN use, ITN access, effective anti-malaria treatment, and malaria infection prevalence.

1.6 Scope and Outline of Study.

This research thesis is outlined as follows;

1. Chapter One: This chapter provides the background of the study topic, the statement of the problem, the significance of the research, and the research objectives.

2. Chapter Two: This chapter provides a detailed literature review aimed at substantiating the need for this research while acknowledging the various efforts made to address similar research problems.
3. Chapter Three: This chapter outlines the research area, data sources, and the methodology used to address each of the objectives.
4. Chapter four: This chapter discusses the results of each specific objective
5. Chapter Five: This chapter outlines the discussions based on the results obtained from each specific objective.
6. Chapter Six: This chapter provides a summary of the conclusions and recommendations derived from the results of the specific objectives.
7. Appendices: This is the last section of the research thesis and it gives the Work-plan and the estimated budget for the research work.

CHAPTER TWO

LITERATURE REVIEW

Malaria, caused by *Plasmodium* parasites primarily *Plasmodium falciparum* in Sub-Saharan Africa is transmitted through bites from infected *Anopheles* mosquitoes (Nyawanda *et al*., 2023). The public health burden of malaria is intensified by ecological conditions favorable for mosquito breeding, including high temperatures, humidity, and stagnant water sources (Ankamah *et al*., 2018). This chapter examines the literature related to the spread of malaria epidemics in both developed and developing countries. It focuses particularly on the analysis of malaria spread and its dynamics, as well as the application of autoregression models in epidemiology.

2.1 Empirical Literature Review

2.1.1 Regional Perspectives on Malaria Modeling

In Kenya, Achieng *et al*. (2020) used weekly cross-sectional data from Kisumu County (2014–2017) to examine malaria trends via ARIMA (2,0,1) and negative binomial regression. Their study revealed seasonal fluctuations in malaria incidence and identified ITN usage and treatment access as critical predictors. Similarly, Nyawanda *et al*. (2023), using Bayesian negative binomial models in western Kenya, found climate variability significantly influenced malaria resurgence post-2016 despite widespread ITN distribution. The incidence was inversely associated with land surface temperature and positively associated with rainfall. In Uganda, Eunice *et al*. (2017) used Poisson and negative binomial models to explore monthly malaria cases in Apac District (2007–2016), reporting strong positive correlations between rainfall and malaria incidence, especially in late-year months. Aregawi *et al*. (2014), focusing on Ethiopia, conducted a time series analysis linking declining malaria mortality (2006–2011) to scaled-up antimalarial interventions, though causality could not be isolated due to ecological variability.

Ankamah *et al*. (2018) employed Vector Autoregressive (VAR) models in Ghana to assess how rainfall, temperature, and humidity influenced malaria incidence. They confirmed strong Granger causality from climatic variables, with impulse response analysis showing maximum malaria impact in months aligned with peak rainfall and humidity. In Nigeria, Amadi and Erandi (2024) enhanced incidence prediction through cluster-based regression models integrating meteorological data, revealing that geographic heterogeneity in incidence could not be fully explained by weather patterns alone.

Diao *et al*. (2023) used Poisson and negative binomial regressions in Senegal to develop a forecast model adjusting for rainfall and ITN distribution, with the Poisson model outperforming. Koné and Atchadé (2023), using spatial regression models in Burkina Faso, highlighted the influence of population literacy and rurality on malaria distribution, with the Spatial Durbin Model (SDM) providing the best fit.

Darkoh *et al*. (2017) explored climatic influences on malaria in Botswana using time series methods, identifying November as the month with peak incidence. Minimum and maximum temperatures (lagged by three months) were significant predictors, while rainfall was not. This suggests temperature as a more reliable driver of malaria dynamics in this region.

2.1.2 Modeling Approaches in Malaria Research

Time series models such as ARIMA, VAR, and Vector Error Correction Models (VECM) have been extensively used. For example, Badaoui *et al*. (2023) in Morocco found the VAR model superior for predicting cutaneous leishmaniasis (CL), demonstrating the model's robustness in handling disease dynamics when stationarity assumptions are met. In Kenya, Warsono *et al.* (2019) applied a $VARX(3,0)$ model in energy modeling—showing promise for applications in disease modeling, especially when exogenous variables are present.

Impulse response functions (IRFs) and variance decompositions are commonly used for understanding dynamic interactions. Bayarbat and Li (2020) used these tools to study the link between health expenditure and economic growth in Mongolia, while Ankamah *et al.* (2018) used them to quantify climatic effects on malaria in Ghana.

Generalized linear models, including Poisson and negative binomial regressions, are widely applied in malaria incidence modeling. Eunice *et al.* (2017) and Diao *et al.* (2023) demonstrated the superiority of negative binomial models in accounting for overdispersion in malaria data. Muriithi *et al.* (2024) applied supervised machine learning models—such as random forest and support vector machines—in Kenya, achieving high predictive accuracy (97.3%) for malaria test outcomes, with *Plasmodium falciparum* presence being the most influential predictor.

Nkiruka *et al.* (2021) extended machine learning analysis across six Sub-Saharan countries using k-means clustering and XGBoost algorithms, revealing non-seasonal climate variability as a key outbreak driver. Similarly, Ogbuagada *et al.* (2022) used multivariate VAR models to analyze malaria incidence among vulnerable groups in Nigeria, finding that malaria cases among adults were consistently higher.

Recent advancements in VAR modeling have addressed some of the classical model limitations. For instance, Cordoni *et al.* (2024) introduced a recursive structural VAR model with nonlinear contemporaneous dependencies, enabling the estimation of structural impulse responses. Karlsson *et al.* (2023) proposed using a generalized hyperbolic skew

Student’s t-distribution for innovations in VAR models, improving robustness in the presence of skewed and heavy-tailed health data.

Virolainen (2024) presented a structural VAR with an endogenously switching conditional covariance matrix, while Wiesen and Beaumont (2024) introduced the joint impulse response function (jIRF) to improve the accuracy of volatility shock assessments. Such techniques, while originally developed in econometrics, are increasingly applicable in health and epidemiology due to their capacity to handle nonlinearity, contemporaneous effects, and cross-sectional dependence.

2.1.3 Identified Research Gaps

While the reviewed studies demonstrate a wide array of modeling techniques—ranging from ARIMA and GLM to advanced structural VARs and machine learning—most remain limited in their scope of incorporating both biological control strategies (e.g., ITN use, ACT access) and time-dependent effects simultaneously. Few studies explicitly assess the interaction between vector-control interventions and temporal malaria trends using multivariate frameworks such as VARX or Bayesian VARs. Additionally, many studies either focus on climate variability or static socio-demographic characteristics, without integrating the operational and usage aspects of interventions like ITNs and effective treatment into temporal models.

This study addresses these gaps by employing a VAR model with exogenous variables (VARX) to analyze malaria incidence and mortality in Migori County, Kenya, where malaria remains highly endemic. By incorporating intervention-related exogenous variables—such as ITN use, access to effective treatment, and prevalence among vulnerable populations—this research advances the understanding of both temporal dependencies and public health responses in malaria dynamics. It builds upon regional and methodological insights to offer an integrated, data-driven framework for malaria modeling, capable of informing targeted interventions in high-burden contexts.

2.2 Conceptual Framework

Figure 2.1 presents the conceptual framework for the study variables for this study.

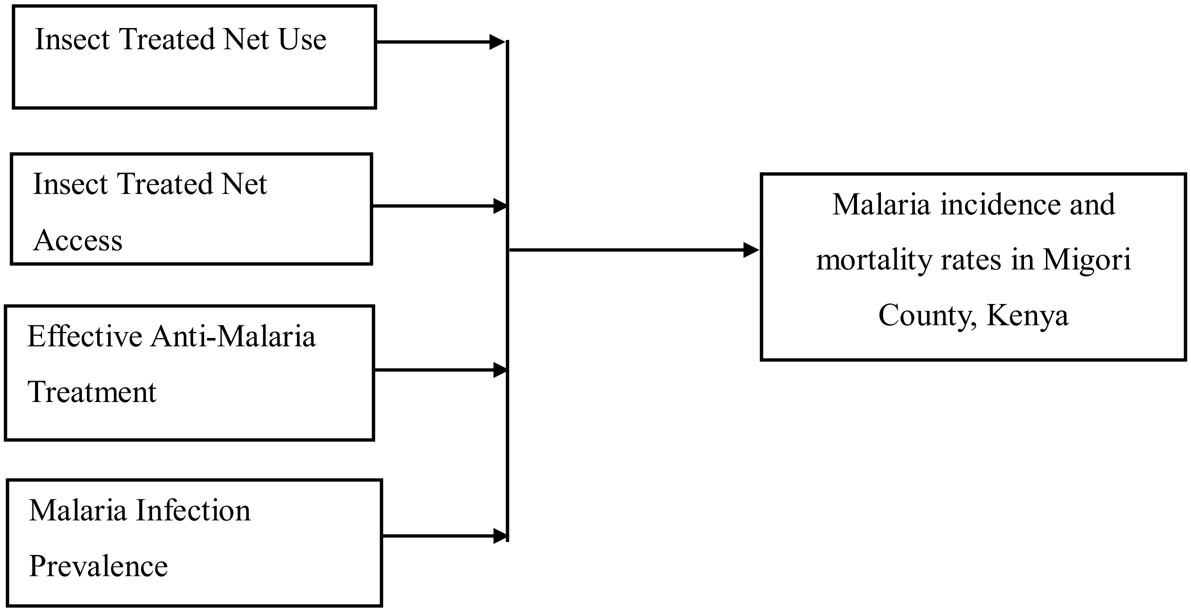


Figure 2.1: Conceptual Framework

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the study methodology to be used in solving each of the study objectives. A map of the study area has also been indicated in this section.

3.2 Study Area and Data

3.2.1 Study Area

This study was conducted in Migori County, Kenya. Figure 3.1 shows the map of Migori County where the study took place.

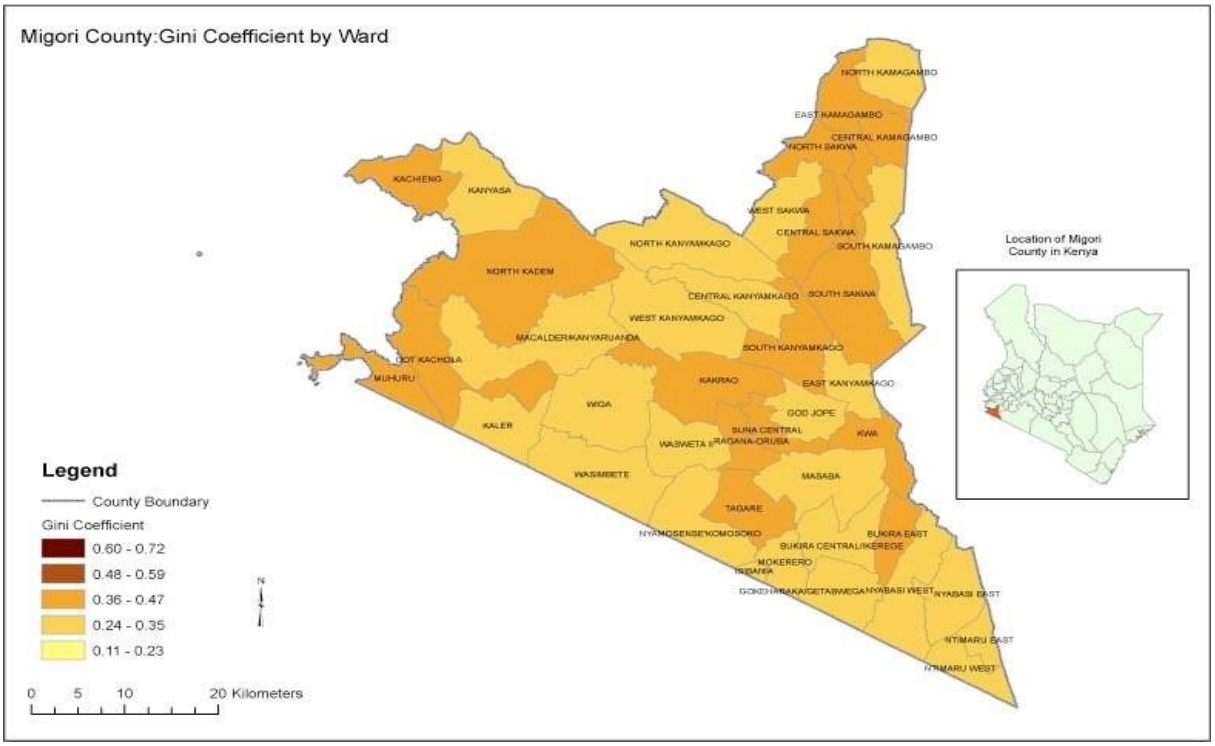


Figure 3.1: Gini Index Migori County Map by Ward

3.2.2 Study Data

This research utilized aggregated monthly secondary data from the Kenya National Health Management Information System (KHIS) spanning the period January 2015 to December 2021, to examine the temporal spread and dynamics of malaria incidence and mortality in Migori County. Data abstraction followed a defined eligibility criterion, focusing on indicators relevant to malaria surveillance, including malaria incidence, malaria-specific mortality, insecticide-treated net (ITN) use, ITN access, antimalarial treatment coverage, and malaria infection prevalence. These variables were operationalized as monthly proportions or rates, aggregated at the county level. The data were cleaned and structured using Microsoft Excel and then analyzed using R statistical software, applying Vector Autoregressive (VAR) and Vector Autoregressive with Exogenous Variables (VARX) models to assess temporal interdependencies.

3.2.3 Study Design

The study employed both descriptive and retrospective time series research designs to describe the nature of malaria incidence and mortality rates in Migori County, Kenya.

3.3 Regression Modeling of Malaria Incidence and Mortality Rates

In order to determine the influence of treatment interventions (ITN use, ITN access, effective anti-malaria treatment, and malaria infection prevalence) on malaria incidence and mortality rates in Migori County, Kenya, the study used a regression equation as given in the below subsections.

3.3.1 Regression Model Definition

The Study regression model is defined as;

$$Y_{it} = \alpha_{it} + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \beta_4 X_{4it} + \mathrm{e}_{it}$$

3.3.2 Parameter Estimation

The study regression model parameters shall be estimated respectively as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_{it} - \bar{Y}_t)(X_{it} - \bar{X}_t)}{\sum_{i=1}^{n}(X_{it} - \bar{X}_t)^2}$$

$$\hat{\alpha} = \bar{Y}_t - \frac{\sum_{i=1}^{n}(Y_{it} - \bar{Y}_t)(X_{it} - \bar{X}_t)}{\sum_{i=1}^{n}(X_{it} - \bar{X}_t)^2}\bar{X}_t$$

for which $\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}$ and $\bar{X}_t = \frac{1}{n}\sum_{i=1}^{n} X_{it}$.

3.4 The vector autoregressive model

3.4.1 Model Definition

Define $Y_t = (y_{1,t}, y_{2,t})'$ a 2×1 vector of time series variables observed at time t for $t = 1, \ldots, T$. A p^{th} order vector autoregressive model, denoted by VAR(p), is written as;

$$Y_t = \mu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t, \quad t = 1,2, \ldots, T \qquad (1)$$

where μ is a 2 dimensional intercept vector, A_i for $i = 1, \ldots, p$ are 2×2 time-invariant coefficient matrices that evaluate the long-run co-movement between time series variables at time t and those at time $t - i$, and ε_t is a 2×1 zero-mean white noise vector process (serially independent) with a timeinvariant variance-covariance matrix Σ (Chang & Shi, 2024).

3.4.2 VAR Model Analysis

For a covariance stationary VAR system, a range of popular metrics are employed to perform further analyses. One sufficient condition to ensure covariance stationarity is the stability, such that the roots of

$$\det\left(I_2 - A_{1,}z - \cdots - A_p z^p\right) = 0$$

lie outside the complex unit circle, where I_2 is the 2×2 identity matrix. If this condition is met, $(I_2 - A_1 - \cdots A_p)^{-1}$ then measures the unconditional or long-term mean of Y_t (Guefano *et al.*, 2021).

Another popular metric of the VAR analysis is the impulse response function (IRF). For a covariance stationary system, (1) can be written as a moving average (MA) model, such that

$$Y_t = c + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \cdots \tag{2}$$

It can be shown that

$$\Phi_i = \sum_{j=1}^{i} \Phi_{i-j} A_j$$

for $i = 1,2,\ldots$, where Φ_0 is a 2×2 identity matrix, and $A_j = 0$ for $j > p$. The elements of Φ_i are intuitively interpreted as the impulse response. However, due to the non-zero cross-sectional correlations among the error sequence, elements of Φ_s for $s > 0$ cannot be easily interpreted. To resolve this issue, one works on the orthogonal errors $\eta_t = L^{-1}\varepsilon_t$, where L is an invertible lower triangular matrix derived using the Cholesky decomposition as follows:-

$$\Sigma = \mathrm{LDL}'. \tag{3}$$

The MA representation (2) can then be rewritten with respect to η_t:

$$Y_t = c + \Theta_1 \eta_{t-1} + \Theta_2 \eta_{t-2} + \cdots$$

IRFs are therefore defined as impulse responses of the i^{th} series to the j^{th}orthogonal shocks $\eta_{j,t}$ or

$$\frac{\partial\, y_{i,t+s}}{\partial\, \eta_{j,t}} = \frac{\partial\, y_{i,t}}{\partial\, \eta_{j,t-s}} = \theta_{i,j}^{s} \;\; i,j = 1,2; s > 0$$

where $\theta_{i,j}^{s}$ is the ij^{th} element of Θ_s.

Another metric discussed in this study is the forecast error variance decomposition (FEVD), which is closely related to IRF. In short, the ij^{th} h-step-ahead FEVD describes the portion of variance of forecast errors in predicting $y_i, T + h$ explained by the j^{th}orthogonal shocks $\eta_{j,t}$. The mathematical definition is provided below.

$$FEVD_{i,j}(h) = \frac{\partial_{\eta_j}^2 \sum_{s=0}^{h-1} \left(\theta_{i,j}^{s}\right)^2}{\sum_{j=1}^{2} \partial_{\eta_j}^2 \sum_{s=0}^{h-1} \left(\theta_{i,j}^{s}\right)^2} \qquad (4)$$

Where $\partial_{\eta_j}^2$ is the variance of $\eta_{j,t}$ and equal to the j^{th} diagonal element of D as defined in (3). In terms of the estimation, assume there are no restrictions put on the parameters A_i for $i = 1, \ldots, p$.

3.4.3 Parameter Estimation

The VAR(*p*) as in (1) can be rewritten as

$$Y_t = B' z_t + \varepsilon_t, \quad t = 1,2, \ldots, T$$

Where z_t is an $(Np+1)\times 1$ vector and $B = (\mu, A_1, \ldots, A_p)'$

is an $(2p+1) \times 2$ matrix that contains all unknown coefficients. Alternatively, if we define $\boldsymbol{Y} = (y_{1,}y_2, \dots, y_T)', \boldsymbol{Z} = (z_{1,}z_2, \dots, z_T)'$, and $\boldsymbol{E} = (\varepsilon_{1,}\varepsilon_2, \dots, \varepsilon_T)'$, the VAR($p$) can be represented as a concise matrix form:

$$\boldsymbol{Y} = \boldsymbol{Z}B + \boldsymbol{E}.$$

Therefore, the VAR(p) can be estimated using ordinary least squares (OLS) under a multivariate regression framework, which is given by the well-known formula:

$$\hat{B}_{OLS} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \qquad (\mathbf{5})$$

Correspondingly, the covariance matrix of the error can be estimated by the sample covariance of the multivariate least squares residuals such that:

$$\hat{\Sigma}_{OLS} = \frac{1}{T-p}\left(Y - Z\hat{B}_{OLS}\right)'\left(Y - Z\hat{B}_{OLS}\right).$$

3.5 The vector autoregressive model with Exogenous Variables

3.5.1 Model Definition

The general form of the vector autoregressive model with exogenous variables (VARX) with the order of the endogenous variables p and the order of the exogenous variables q, VARX(p, q), is expressed as follows:

$$\mathbf{z}_t = \mu + \sum_{i=1}^{p} \phi_i \, \mathbf{z}_{t-i} + \sum_{j=0}^{q} \theta_j \, \mathbf{x}_{t-j} + \mathbf{e}_t,$$

where $\mathbf{z}_t$ represents the n-dimensional vector containing the endogenous variables at time t, Ψ_i, $i = 1,2, \dots, p$ express the $(n \times n)$ coefficient matrices of $\mathbf{z}_{t-i}$, $\mathbf{x}_t$ represents the r-dimensional vector containing the exogenous variables at time t, θ_j, $j = 0,1, \dots, q$ express the $(n \times r)$ coefficient matrices of $\mathbf{x}_{t-j}$, and $\mathbf{e}_t$ represents the n-dimensional random error vector. Equation (1) can also be reformulated as follows:

$$\phi(B_z)\mathbf{z}_t = \mu + \theta(B_x)\mathbf{x}_t + \mathbf{e}_t,$$

where $\phi(B_z) = I_n - \sum_{i=1}^{p} \phi_i B_z^i$, $\theta(B_x) = \sum_{j=0}^{q} \theta_j B_x^j$, $B_z^i \mathbf{z}_t = \mathbf{z}_{t-i}$, and $B_x^j \mathbf{x}_t = \mathbf{x}_{t-j}$.

There are three assumptions of the VARX model presented as follows:

i. The VARX(p, q) model is stationary if the roots of the determinants of $\phi(v)$ and $\theta(w)$ are outside the unit circle.

ii. $\mathbf{e}_t$ is independent and identically multivariate normally distributed with zero-vector mean and a constant covariance matrix Σ. Σ is a positive semi-definite matrix expressed as; $\Sigma = \left[\mathrm{Cov}\left(e_{k1,t}, e_{k2,t}\right)\right]$,where$k_1, k_2 = 1,2, \dots, n$.

iii. $\mathbf{e}_t$ and $\mathbf{x}_t$ are independent.

Many criteria can be used to determine the order of the VARX model. Lui *et al.* stated that the order of the VARX model is determined through the Akaike information criterion (AIC). The model with the smallest AIC is the best. The AIC value of a model can be determined by the following equation:

$$\mathrm{AIC} = \ln(|\Sigma|) + \frac{2L}{T},$$

where T is the size of observation data, and L is the number of model parameters.

3.5.2 Augmented Dickey-Fuller Test

The unit root test equation for the Augmented Dickey-Fuller (ADF) test from data that have constant, and trend elements is as follows:

$$\Delta z_{k,t} = \alpha_k + \beta_k t + \gamma_k z_{k,t-1} + \sum_{i=1}^{p-1} \delta_{k,i}\, \Delta z_{k,t-i} + e_{k,t},$$

where $z_{k,t}$ represents observation data k at time t, $\Delta z_{k,t}$ represents the difference between $z_{k,t}$ and $z_{k,t-1}$, α_k is a constant, β_k represents the trend parameter, $\delta_{k,i}$ is the parameter of $\Delta z_{k,t-i}$, p represents the optimal lag, and $e_{k,t}$ is a random error that is independent and identically normally distributed with zero mean and constant variance. The following are the test hypotheses used: $H_0: \gamma_k = 0$, and $H_1: \gamma_k < 0$. The test statistics used is the t-ratio of the parameter estimator $\hat{\gamma}_k$ estimated using the least squares method, stated as follows:

$$t = \frac{\hat{\gamma}_k}{\mathrm{SE}(\hat{\gamma}_k)},$$

where $\hat{\gamma}_k$ is an estimator of γ_k, and $\mathrm{SE}(\hat{\gamma}_k)$ is the standard error of $\hat{\gamma}_k$. If t is less than the critical value, then H_0 is rejected, which means the data are stationary; otherwise, if H_0 is not rejected, the data must be differenced until they are stationary.

3.5.3 Granger Causality Test

Suppose that there are two stationary time series $z_{1,t}$ and $z_{2,t}$. In certain cases, $z_{1,t}$ can be affected by $z_{1,t-i}$ and $z_{2,t-i}$, and $z_{2,t}$ can also be affected by $z_{1,t-i}$ and $z_{2,t-i}$. This relationship is called a bidirectional causality relationship. The Granger causality test is a popular statistical test used to check this relationship.

The relationship between $z_{1,t}$ with $z_{1,t-i}$, which is also called a restricted relationship, is stated as follows:

$$z_{1,t} = \alpha_1 + \sum_{i=1}^{p} \beta_{1,i}\, z_{1,t-i} + e_{1,t}.$$

Then, the relationship between $z_{1,t}$ with $z_{1,t-i}$ and $z_{2,t-i}$, which is also called the unrestricted relationship, is stated as follows:

$$z_{1,t} = \alpha_1 + \sum_{i=1}^{p} \beta_{1,i}\, z_{1,t-i} + \sum_{i=1}^{p} \beta_{2,i}\, z_{2,t-i} + e_{1,t}.$$

The test hypotheses used are as follows: $H_0\colon \beta_{2,i} = 0$, and $H_1\colon \beta_{2,i} \neq 0$. The F test statistics are determined by the following equation:

$$F = \frac{(R_{\mathrm{ur}}^2 - R_{\mathrm{r}}^2)/p}{(1 - R_{\mathrm{ur}}^2)/(T - p - 1)},$$

where R_{ur}^2 and R_{r}^2 represent the coefficients of determination in equations (6) and (7), respectively. If the F test statistic is greater than $F_{p,T-p-1}$, then H_0 is rejected, which means $z_{1,t}$ is influenced by $z_{1,t-i}$ and $z_{2,t-i}$. $z_{1,t}$ and $z_{2,t}$ have a bidirectional causality if $z_{1,t}$ is influenced by $z_{1,t-i}$ and $z_{2,t-i}$, and $z_{2,t}$ is influenced by $z_{1,t-i}$ and $z_{2,t-i}$.

3.5.4 VARX Parameter Estimation Using Maximum Likelihood (ML) Method

The VARX models with $n = 2$, $r = 1$, $p = 1$, and $q = 0$ can be stated as follows:

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta} + \mathbf{e}_t,$$

where

$$\mathbf{y}_t = \begin{pmatrix} z_{1,t} \\ z_{2,t} \end{pmatrix}, \quad \mathbf{X}_t = \begin{pmatrix} 1 & 0 & z_{1,t-1} & z_{2,t-1} & 0 & 0 & x_{1,t} & 0 \\ 0 & 1 & 0 & 0 & z_{1,t-1} & z_{2,t-1} & 0 & x_{1,t} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \phi_{11}^{(1)} \\ \phi_{12}^{(1)} \\ \phi_{21}^{(1)} \\ \phi_{22}^{(1)} \\ \theta_{11}^{(0)} \\ \theta_{21}^{(0)} \end{pmatrix}, \quad \mathbf{e}_t = \begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix}.$$

Based on equation (9), $\mathbf{e}_t$ can be stated as follows:

$$\mathbf{e}_t = \mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}.$$

Because $\mathbf{e}_t \sim N(\mathbf{0}, \Sigma)$, then $\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma)$. Therefore, the distribution function of $\mathbf{e}_t$ can be expressed as follows:

$$f(\mathbf{e}_t|\boldsymbol{\beta}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta})^\top \Sigma^{-1}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta})\right).$$

The vector estimator of $\boldsymbol{\beta}$ is obtained by maximizing the likelihood function of $\mathbf{e}_t$, which is stated as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \Sigma|\mathbf{e}_2, \dots, \mathbf{e}_T) = \frac{1}{(2\pi)^{n(T-1)/2}|\Sigma|^{(T-1)/2}} \exp\left(-\frac{1}{2}\sum_{t=2}^{T}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta})^\top \Sigma^{-1}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta})\right).$$

Generally, equation (14) is transformed into a natural logarithm to facilitate maximization. This function is then referred to as the log-likelihood function, which is expressed as follows:

$$\ln\mathcal{L}(\boldsymbol{\beta}, \Sigma \left| \mathbf{e}_2, \dots, \mathbf{e}_T) = -\frac{(T-1)n}{2}\ln(2\pi) - \frac{(T-1)}{2}\ln\right| \Sigma|$$
$$-\frac{1}{2}\sum_{t=2}^{T}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta})^\top \Sigma^{-1}(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}).$$

The vector estimator of $\boldsymbol{\beta}$ is the vector that zeroes the derivative of equation (15) with respect to $\boldsymbol{\beta}$. The following equation is used to estimate the vector $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=2}^{T} \mathbf{X}_t^{\top} \Sigma^{-1} \mathbf{X}_t \right)^{-1} \sum_{t=2}^{T} \mathbf{X}_t^{\top} \Sigma^{-1} \mathbf{y}_t.$$

3.6. Bayesian Vector Autoregression with Exogenous Variables

3.6.1 Model Definition

This section presents Bayesian VAR models with three different priors – independent Normal-Wishart prior, the Minnesota prior, and the SSVS prior. The VAR model can be written in matrix form as follows:

$$Y = X\Phi + \epsilon$$

where the $T \times n$ matrix Y is defined as $Y = (y_1, \dots, y_T)'$; the $T \times (1 + np)$ matrix X is defined as $X = (x_1, \dots, x_T)'$; the $(1 + np) \times 1$ vector is defined as $x_t = \left(1, y_{t-1}', \dots, y_{t-p}'\right)'$; the $(1 + np) \times n$ matrix Φ is defined as $\Phi = \left(\mu', \Theta_1', \dots, \Theta_p'\right)'$; and ϵ is a $T \times n$ matrix with $\epsilon = (\epsilon_1, \dots, \epsilon_T)'$ (Hou *et al.*, 2023).

Based on the VAR model, the study describes the independent Normal-Wishart prior used in the study:

$$\begin{aligned} \text{vec}(\Phi) \quad &\sim MN(\text{vec}(\Phi_0), V_0) \\ \Sigma \quad &\sim IW(\Sigma_0, \nu_0) \end{aligned}$$

where MN refers to a multivariate normal with mean $\text{vec}(\Phi_0)$ and covariance matrix V_0, and IW refers to an inverted Wishart distribution with parameters Σ_0 and degrees of freedom ν_0. Unlike the natural conjugate priors, the prior for Φ in eq. (7) and Σ in eq. (8) are independently specified (Sugita, 2022).

With the joint prior and the likelihood, the conditional posterior densities of $\mathrm{vec}(\Phi)$ and Σ are derived as follows:

$$\begin{aligned} \mathrm{vec}(\Phi)|\Sigma, Y &\sim MN(\mathrm{vec}(\Phi^*), V^*) \\ \Sigma|\Phi, Y &\sim IW(\Sigma^*, \nu^*) \end{aligned}$$

where $V^* = (V_0^{-1} + \Sigma \otimes X'X)^{-1}$ and

$$\mathrm{vec}(\Phi^*) = V^*[V_0^{-1}\mathrm{vec}(\Phi_0) + (\Sigma \otimes I_\kappa)^{-1}\mathrm{vec}(X'Y)],$$

$\Sigma^* = (Y - X\Phi)'(Y - X\Phi) + \Sigma_0$, and $\nu^* = T + \nu_0$. Given these conditional posterior specifications, the Gibbs sampler generates sample draws.

Note that, with zero prior mean $\Phi_0 = 0$ and large prior variance V_0, the posterior mean for Φ is almost identical to the maximum likelihood estimator. In this study, the hyperparameters are set at $\mathrm{vec}(\Phi_0) = 0$ and $V_0 = 100$, $\Sigma_0 = 0.1I$, and $\nu_0 = 5$ as with Sukono *et al.*, 2023.

3.6.2 Monte Carlo Simulations for Bayesian VAR Model

To estimate the parameters of the Bayesian VAR model using Monte Carlo simulations, the study applies the Gibbs sampler as described earlier. The steps involved in the Monte Carlo estimation procedure are as follows:

i. Initialize the parameters $\Phi^{(0)}$ and $\Sigma^{(0)}$ based on prior distributions as

$$\begin{aligned} \mathrm{vec}(\Phi) &\sim MN(\mathrm{vec}(\Phi_0), V_0), \\ \Sigma &\sim IW(\Sigma_0, \nu_0). \end{aligned}$$

ii. For each iteration $i = 1,2, \dots, N$ (where N is the number of Monte Carlo draws):

a. Sample $\Phi^{(i)}$ conditional on $\Sigma^{(i-1)}$ and Y:

$$\mathrm{vec}\left(\Phi^{(i)}\right)|\Sigma^{(i-1)}, Y \sim MN(\mathrm{vec}(\Phi^*), V^*),$$

where $V^* = \left(V_0^{-1} + \Sigma^{(i-1)} \otimes X'X\right)^{-1}$ and

$$\mathrm{vec}(\Phi^*) = V^* \left[V_0^{-1}\mathrm{vec}(\Phi_0) + \left(\Sigma^{(i-1)} \otimes I_\kappa\right)^{-1}\mathrm{vec}(X'Y)\right].$$

b. Sample $\Sigma^{(i)}$ conditional on $\Phi^{(i)}$ and Y:

$$\Sigma^{(i)}|\Phi^{(i)}, Y \sim IW(\Sigma^*, \nu^*),$$

where $\Sigma^* = \left(Y - X\Phi^{(i)}\right)'\left(Y - X\Phi^{(i)}\right) + \Sigma_0$ and $\nu^* = T + \nu_0$.

iii. Repeat the above steps until convergence is achieved.

The posterior mean and variance of the parameters Φ and Σ can be estimated as the sample averages of the Monte Carlo draws:

$$\hat{\Phi} = \frac{1}{N}\sum_{i=1}^{N} \Phi^{(i)}, \quad \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \Sigma^{(i)}.$$

The Gibbs sampling procedure ensures that the posterior distributions of the parameters Φ and Σ are obtained, allowing us to make inferences about the VAR model's coefficients.

3.6.3 Convergence Diagnostics

To check the convergence of the Gibbs sampler, the study uses standard diagnostics such as trace plots, the Gelman-Rubin diagnostic, and effective sample size (ESS). These ensure that the chains have mixed well and that the posterior draws accurately reflect the true parameter distribution.

i. **Trace plots**: A trace plot of the Gibbs sampler iterations should show stable fluctuations around the posterior mean.

ii. **Gelman-Rubin diagnostic**: The potential scale reduction factor (PSRF) should approach 1 as the chains converge.

iii. **Effective sample size (ESS)**: A higher ESS indicates that the samples are uncorrelated and provide accurate parameter estimates.

3.7 Diagnostic Tests for the Fitted Models

3.7.1 MAPE and Ljung-Box test

The diagnostic test is a feasibility test of the model for forecasting. This feasibility test includes checking the assumption of error and the model's mean absolute percentage error (MAPE). The model errors are tested to determine whether they are independent and identically normally distributed with zero mean and constant variance. This assumption is also known as the white noise assumption. One of the statistical tests commonly used to test whether the errors are independent or not is the Ljung-Box test. The null hypothesis (H_0) in this test is that the errors are independent, while the alternative hypothesis (H_1) is the opposite. The test statistic with the lag length M denoted by Q_M is determined using the following equation:

$$Q_M = T(T+2)\sum_{m=1}^{M}\frac{r_m^2}{T-m},$$

where

$$r_m = \frac{\sum_{t=m+1}^{T} e_{k,t}\, e_{k,t-m}}{\sum_{t=1}^{T} e_{k,t}^2}.$$

Reject H_0 if $Q_M > \chi^2_{1-\alpha,\text{df}}$.

Meanwhile, the assumption that the error is normally distributed with zero mean and constant variance can be visually checked using the quantile-quantile plot (Q-Q Plot). Suppose $F_N\left(e_{k,t}\right)$ is the value of the normal cumulative distribution function with zero mean and constant variance of $e_{k,t}$. If the scatter of point pairs $\left(e_{k,t}, F_N^{-1}\left(e_{k,t}\right)\right)$ is around a line with a gradient of one, then $e_{k,t}$ is visually normally distributed with zero mean and constant variance. If the assumption of white noise in the model error is met and the MAPE of the model is less than 20%, then the model obtained is suitable for use in forecasting.

3.7.2 Economic Test for Granger Causality

To test whether a particular series Y Granger-causes X, consider the following model (Hamilton, 1994) with autoregressive lag length p:

$$X_t = c_t + \sum_{i=1}^{p} \alpha_i \, X_{t-i} + \sum_{i=1}^{p} \beta_i \, Y_{t-i} + u_t$$

Under the null hypothesis:

$$H_0 \colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

the sum of squared residuals from model (11) is:

$$\text{RSS}_1 = \sum_{t=1}^{T} \hat{u}_t^2$$

Under the null hypothesis, the model is:

$$X_t = c_0 + \sum_{i=1}^{p} \gamma_i \, X_{t-i} + \epsilon_t$$

The sum of squared residuals from model (12) is:

$$\mathrm{RSS}_0 = \sum_{t=1}^{T} \hat{\epsilon}_t^2$$

The test statistic is:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/p}{\mathrm{RSS}_1/(T - 2p - 1)}$$

where H_0 is rejected if $F > F_{0.05;(p,T-2p-1)}$.

3.7.3 Impulse Response Function (IRF)

The VAR model can be expressed in vector MA (∞) form as:

$$X_t = \mu_0 + \epsilon_t + \Psi_1 \epsilon_{t-1} + \Psi_2 \epsilon_{t-2} + \cdots$$

where the matrix Ψ_s represents the effects of innovations. Specifically, the row i, column j element of Ψ_s identifies the effect of a one-unit increase in the j-th variable's innovation at time t (ϵ_{jt}) on the i-th variable at time $t + s$ ($X_{i,t+s}$), while keeping all other innovations constant.

If the first element of ϵ_t is changed by δ_1, the second element by δ_2, and the n-th element by δ_n, then the combined effect of these changes on the value of the vector X_{t+s} would be:

$$\begin{aligned} X_{t+s} \quad &= \Psi_s \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix} + X_t \\ &= \begin{bmatrix} X_{1,t+s} \\ X_{2,t+s} \\ \vdots \\ X_{n,t+s} \end{bmatrix}. \end{aligned}$$

A plot of the row i, column j element of Ψ_s is known as the Impulse Response Function (IRF) and is denoted as:

$$\text{IRF}_{ij}(s) = \frac{\partial X_{i,t+s}}{\partial \epsilon_{jt}}.$$

3.8 Forecasting

Forecasting is a key goal in the analysis of multivariate time series. When it comes to forecasting in a VAR (p) model, the process is similar to forecasting in a univariate AR (p) model. The initial step involves identifying the most optimal VAR model using selection criteria. Once the model has been selected, it can be utilized for forecasting purposes.

For a VARX (p,q) model with known parameters ϕ_i for $i = 1,2,\dots,p$ and ψ_j for $j = 1,2,\dots,q$, the best predictor for Γ_{T+1} (one-step forecast) based on data available at time T is:

$$\begin{aligned}\hat{\Gamma}_{T+1|T} &= c + \Phi_1\hat{\Gamma}_T + \Phi_2\hat{\Gamma}_{T-1} + \cdots + \Phi_p\hat{\Gamma}_{T-p+1} \\ &\quad + \Psi_1\hat{X}_T + \Psi_2\hat{X}_{T-1} + \cdots + \Psi_q\hat{X}_{T-q+1}.\end{aligned}$$

For longer forecasts, such as h-step ahead forecasts, the chain rule of forecasting is used:

$$\begin{aligned}\hat{\Gamma}_{T+h|T} &= c + \Phi_1\hat{\Gamma}_{T+h-1} + \Phi_2\hat{\Gamma}_{T+h-2} + \cdots + \Phi_p\hat{\Gamma}_{T+h-p} \\ &\quad + \Psi_1\hat{X}_{T+h-1} + \Psi_2\hat{X}_{T+h-2} + \cdots + \Psi_q\hat{X}_{T+h-q}.\end{aligned}$$

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter presents the results and findings of the analysis in relation to the study objectives.

4.2 The analysis of the spread of malaria incidence and mortality using Vector Autoregression Models with Exogenous Variables in Migori County, Kenya.

4.2.1 Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted in Table 4.1 to examine the existing trends in mortality and incidence rates. The descriptive statistics of the study variables are displayed in Table 4.1, accompanied by visual representations in Figures 4.1, 4.2, 4.3, and 4.4.

Table 4.1: Variations in the malaria mortality and incidence rates together with intervention mechanisms in migori county, Kenya.

	Malaria Mortality Rate	Malaria Incidence Rate	Insect Treated Net Use	Insect Treated Net Access	Anti-Malaria Effective Treatment	Malaria Infection Prevalence
Min.	6.85	87.14	41.11	42.48	47.35	4.778
1st Qu.	11.23	106.15	53.71	55.97	51.71	5.979
Median	12.82	126.64	61.96	62.25	54.55	7.723
Mean.	13.07	142.57	59.64	61.68	54.24	8.941
3rd Qu.	14.79	187.31	68.53	71.96	57.17	12.347
Max.	19.37	199.58	75.81	79.86	58.56	13.449

In Migori County, there are various malaria statistics and factors to consider. The malaria mortality rates ranged from a minimum of 6.85 to a maximum of 19.37, with a mean of

13.07 and a median of 12.82. Incidence rates varied between 87.14 and 199.58, with a mean of 142.57 and a median of 126.64. The percentage of insecticide-treated net (ITN) use ranged from 41.11% to 75.81%, with a mean of 59.64% and a median of 61.96%. ITN access was slightly higher, ranging from 42.48% to 79.86%, with a mean of 61.68% and a median of 62.25%. The Anti-Malaria Effective Treatment (AMET) values, which may be influenced by environmental or intervention conditions, ranged from 47.35 to 58.56, with a mean of 54.24 and a median of 54.55. Malaria Infection Prevalence (MIP) rates varied from 4.778 to 13.449, with a mean of 8.941 and a median of 7.723.

Figure 4.1 shows histograms and density plots of malaria incidence and mortality rates, respectively. These plots depict a relatively normal distribution for both malaria incidence and mortality rates. However, in Figure 4.2, ITN use, ITN access, and AMET exhibit an approximately normal distribution, whereas malaria infection prevalence is skewed to the left.

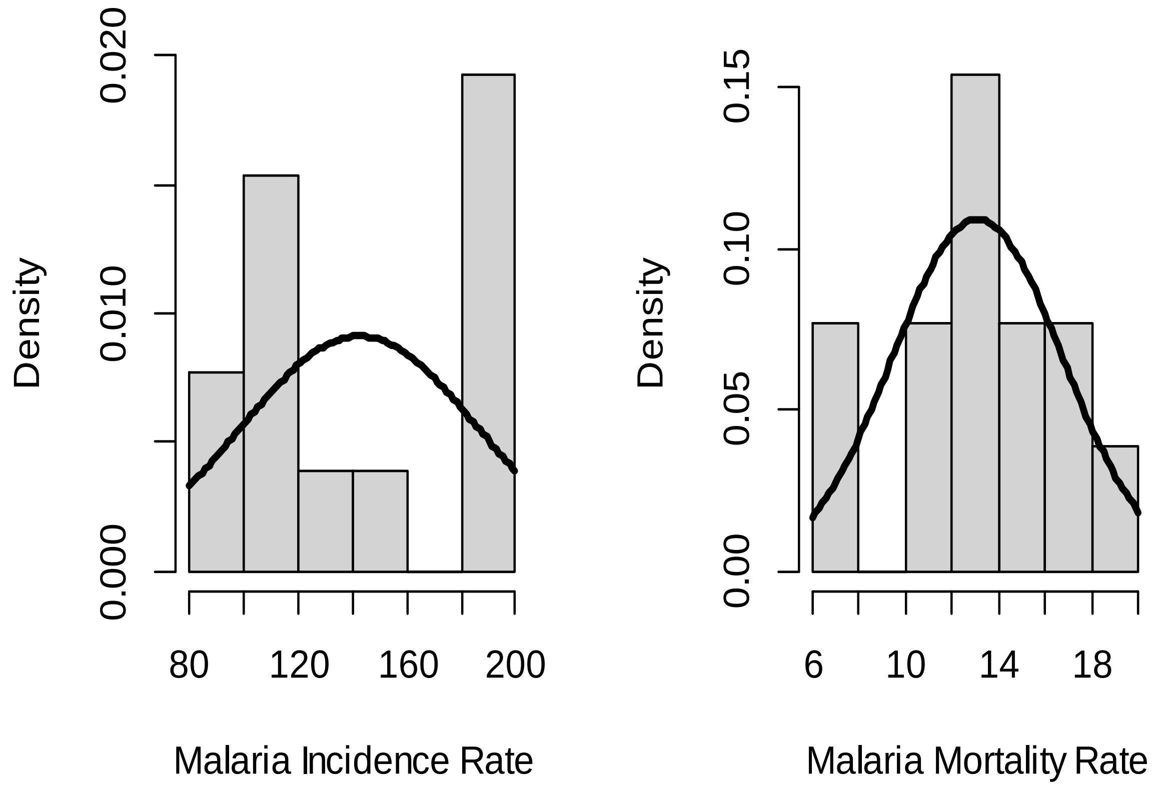


Figure 4.1: Histogram of Malaria Incidence Rate and Mortality Rate

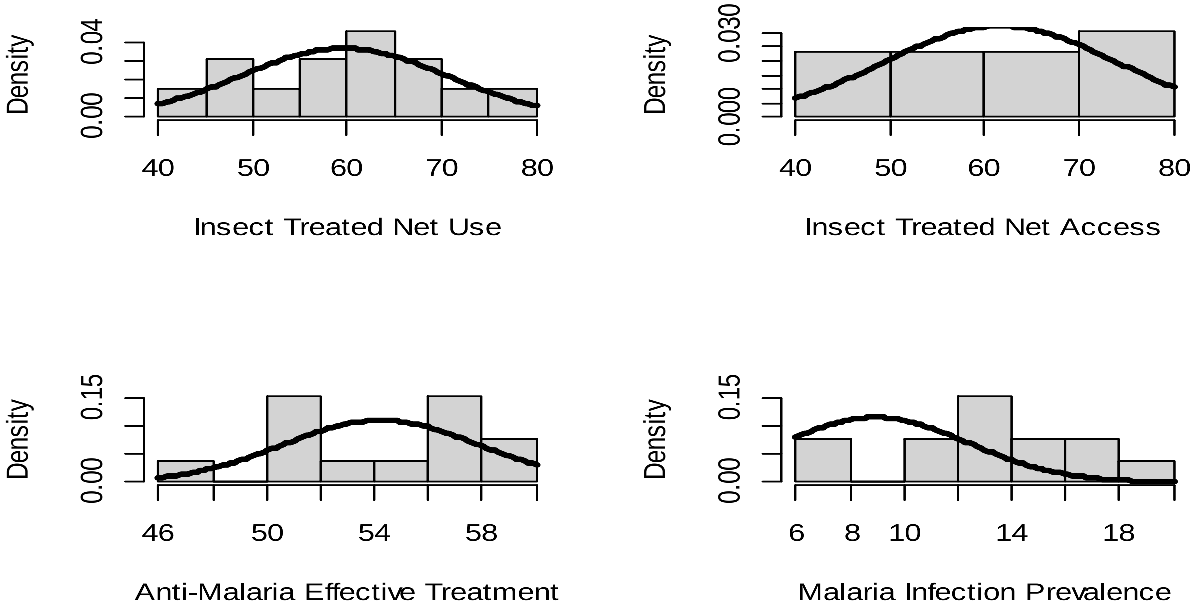


Figure 4.2: Histogram of Insect Treated Net Use, Insect Treated Net Access, Anti-Malaria Effective Treatment and Malaria Infection Prevalence.

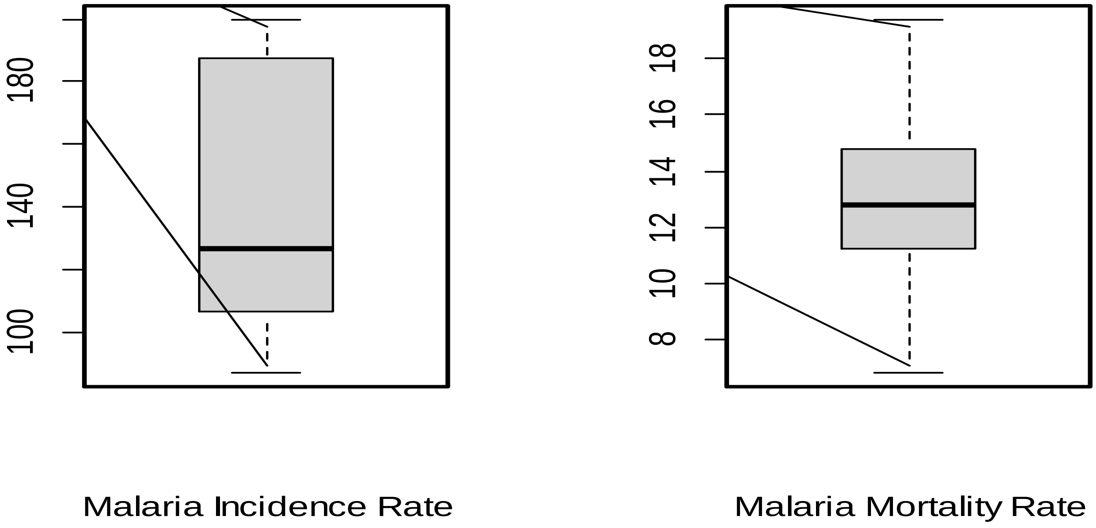


Figure 4.3: Boxplot of Malaria Incidence Rate and Mortality Rate

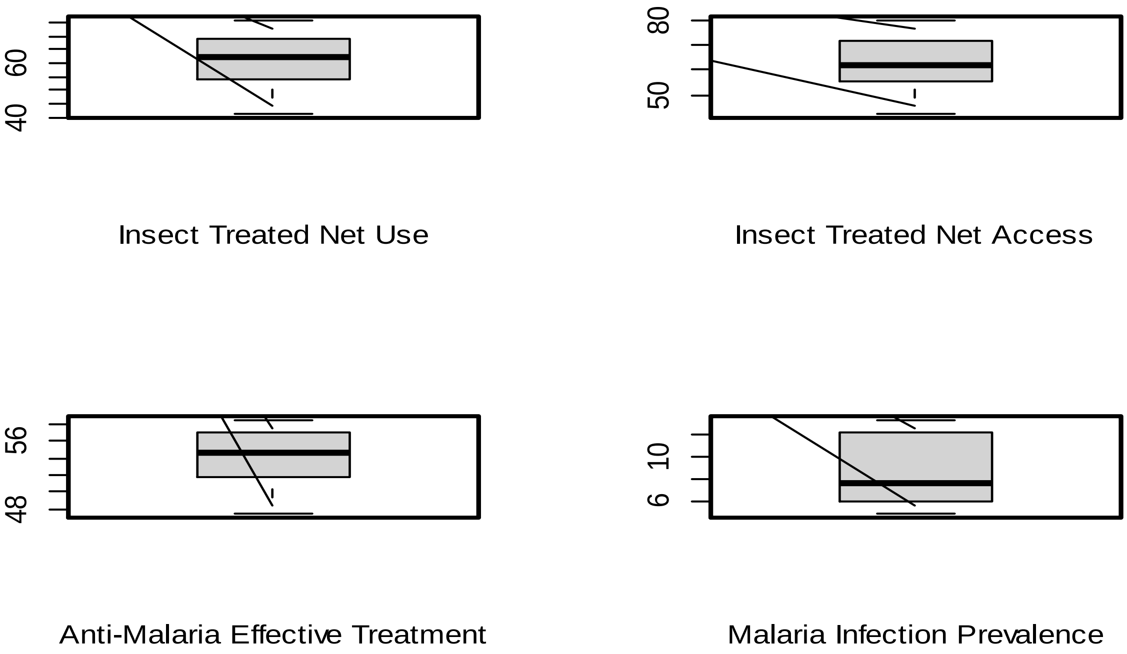


Figure 4.4: Boxplot of Insect Treated Net Use, Insect Treated Net Access, Anti-Malaria Effective Treatment and Malaria Infection Prevalence

4.2.2 Vector Autoregression Analysis

This study employed Vector Auto Regression Analysis. Initially, the Vector Autoregression Model was used without the inclusion of exogenous variables. Later on, exogenous variables were incorporated as well. Furthermore, a Bayesian estimation of the model was conducted, which included the incorporation of exogenous variables. The findings indicated a decrease in the Malaria Incidence Rate (Series 1) over time, whereas the Malaria Mortality Rate (Series 2) remained relatively constant. Notably, the Malaria Mortality Rate reached its peak value towards the end of the time series, as illustrated in figure 4.5.

Furthermore, the autocorrelation function (ACF) of the Malaria Incidence Rate (Series 1) and Malaria Mortality Rate (Series 2) indicated stationarity in the provided time series data, as shown in Figure 4.6. However, the partial autocorrelation function (PACF) of the Malaria Incidence Rate (Series 1) and Malaria Mortality Rate (Series 2) time series data revealed non-stationary states, as illustrated in Figure 4.7, respectively.

4.2.3 Vector Autoregression Model

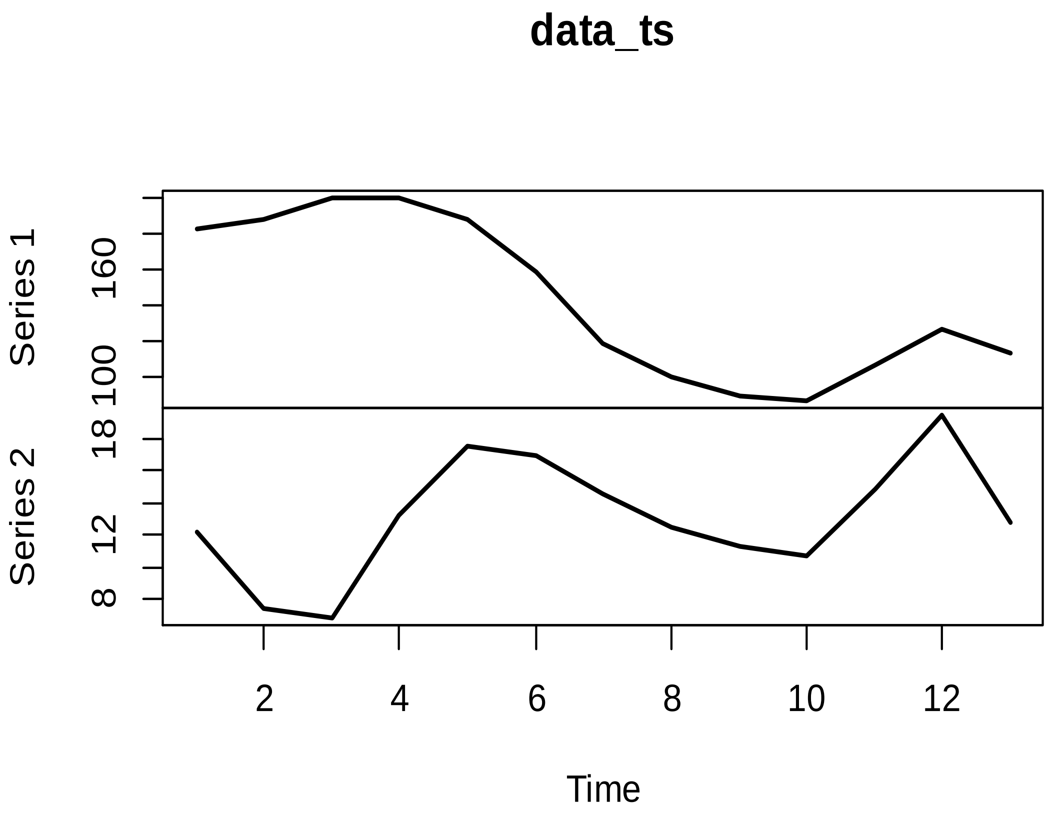


Figure 4.5: Plot of Time Series Data for Malaria Incidence Rate (Series 1) and Malaria Mortality Rate (Series 2)

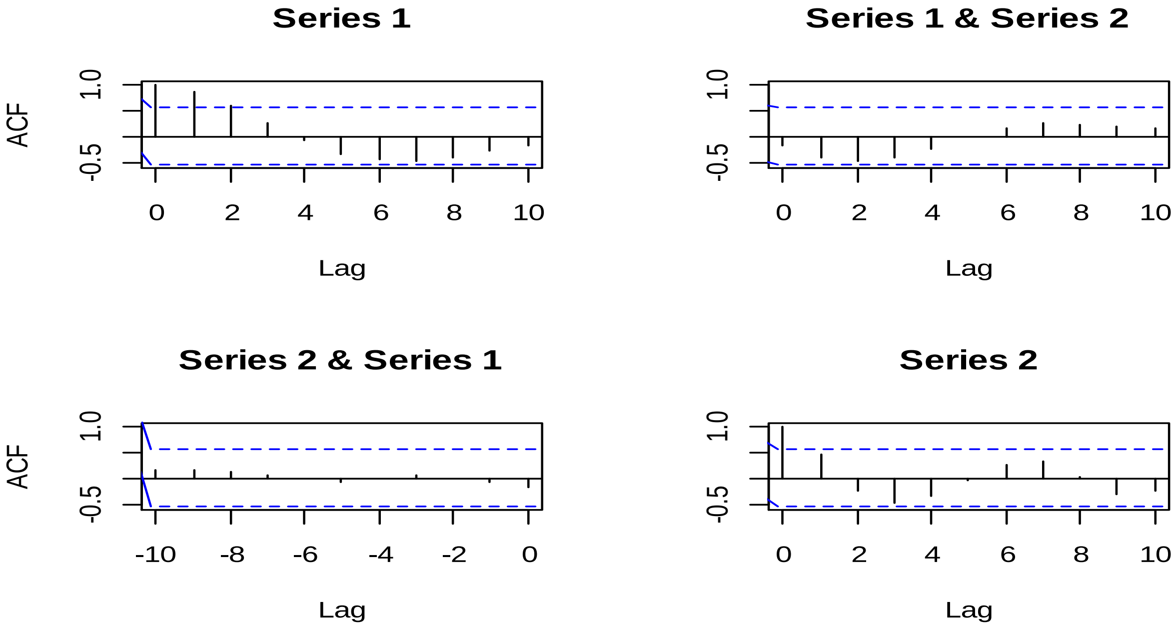


Figure 4.6: ACF of Malaria Incidence Rate (Series 1) and Malaria Mortality Rate (Series 2) Time Series Data

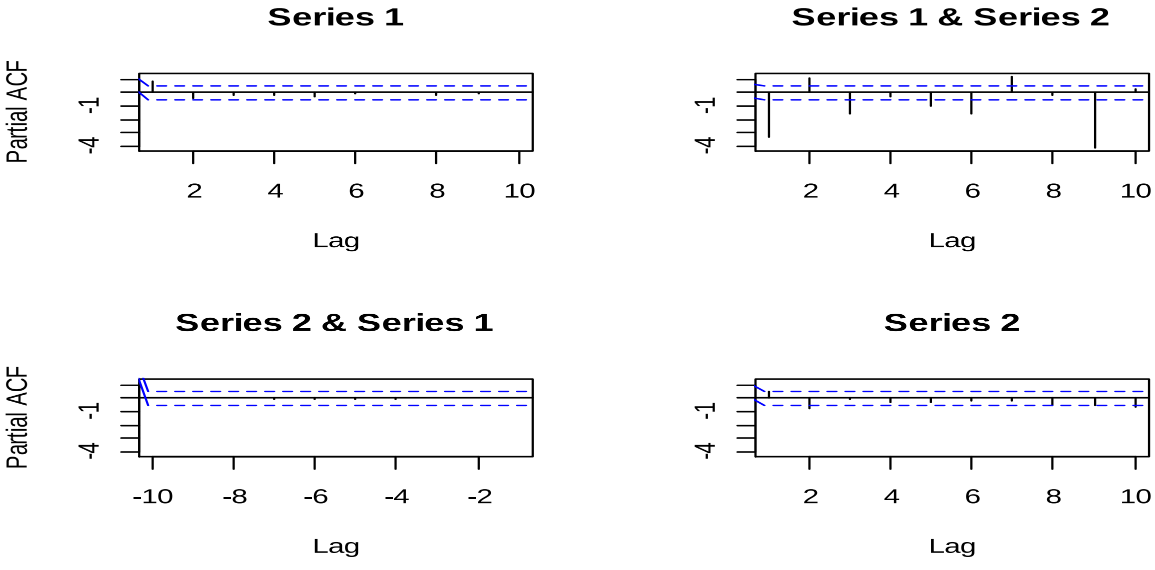


Figure 4.7: PACF of Malaria Incidence Rate (Series 1) and Malaria Mortality Rate (Series 2) Time Series Data

Table 4.3 displays the VAR estimation results for the model examining malaria incidence and mortality rates. Several noteworthy findings are evident. Firstly, the incidence series has a substantial positive effect on the outcome, as indicated by the estimate of 0.8537 and the highly significant p-value of 0.000422. These results suggest a strong relationship between past malaria incidence and current values, highlighting the predictive power of previous incidence rates for future incidence within the context of this model. The mortality series indicates a negative correlation with the outcome. The estimate is -3.2028 with a p-value of 0.081134. While this finding is only marginally significant at the 10% level, it suggests that higher past mortality rates may be linked to a decrease in future rates. The constant term in the model is positive and estimated at 57.3678, but it is not statistically significant ($p = 0.100137$). This means that the baseline level of the outcome variable, when other factors are not considered, does not significantly differ from zero. The other variables, labeled as Sd 1, Sd 2, and Sd 3, all had negative estimates (-3.7436, -2.9577, and -4.7234, respectively). However, they are not statistically significant as their p-values exceed the standard significance thresholds ($p > 0.75$). Therefore, it can be concluded that these standard deviations did not have a meaningful impact on the outcome variable within the context of the study VAR model.

Table 4.2: VAR Series 1 Model Estimates

	Estimate	Std Error	t-value	Pr(>|t|)
Incidence Series	0.8537	0.1219	7.006	0.000422 ***
Mortality Series	-3.2028	1.5294	-2.094	0.081134.
Constant	57.3678	29.5376	1.942	0.100137
Sd 1	-3.7436	15.1433	-0.247	0.812988
Sd 2	-2.9577	14.7426	-0.201	0.847620
Sd 3	-4.7234	14.4410	-0.327	0.754715

The model's residuals had a standard error of 17.67 with 6 degrees of freedom, indicating some unexplained variability. The multiple R-squared value is 0.9122, and the adjusted R-squared value is 0.8391, suggesting that the model explains a significant portion of the outcome variable's variability, although further improvement is possible. The F-statistics are 12.47 with a p-value of 0.004009, confirming the model's overall statistical significance and the strong influence of the predictors on the outcome variable.

In Series 2, the incidence series has a minimal positive estimate of 0.003748. However, this estimate is not statistically significant, as indicated by a p-value of 0.898 in Table 4.4. Therefore, the past incidence rates had little predictive value for the current outcome in this model. On the other hand, the mortality series shows a positive estimate of 0.527169, but it is also not statistically significant (p = 0.185). This suggests that while there is a positive relationship between past mortality rates and the outcome, this effect is not substantial enough to be considered statistically meaningful. Additionally, the estimate for the constant term is 5.699516, but it is not considered significant (p = 0.434). This suggests that the baseline level of the outcome variable, without taking other factors into account, does not differ significantly from zero.

Table 4.3: VAR Series 1 Model Estimates

	Estimate	Std Error	t-value	Pr(>|t|)
Incidence Series	0.003748	0.028064	0.134	0.898
Mortality Series	0.527169	0.352222	1.497	0.185
Constant	5.699516	6.802444	0.838	0.434
Sd 1	-2.780921	3.487461	-0.797	0.456
Sd 2	-4.253277	3.395186	-1.253	0.257
Sd 3	-2.801855	3.325735	-0.842	0.432

The additional variables, labeled as Sd 1, Sd 2, and Sd 3, had negative estimates (-2.780921, -4.253277, and -2.801855, respectively). However, these estimates are not statistically significant as their p-values exceed the significance level of 0.05. This suggests that these standard deviations did not have a meaningful impact on the outcome variable within this model. Furthermore, the model's residual standard error is 4.069 with 6 degrees of freedom, indicating a significant amount of unexplained variability in the outcome. The multiple R-squared value is 0.376, and the adjusted R-squared is -0.144, which is negative. This negative value indicates that the model does not fit the data well. Additionally, the F-statistic is 0.7231 with a p-value of 0.6304, further confirming that the model as a whole is not statistically significant. This also suggests that the predictors do not exert a strong influence on the outcome variable in this particular case.

The covariance matrix of residuals from the VAR model for malaria incidence and mortality rates reveals a significant disparity in the variability of the residuals between Series 1 and Series 2 (see Table 4.5). Series 1 displays a variance of 312.23, indicating a substantially higher level of unexplained variability compared to Series 2, which only exhibits a variance of 16.56. This discrepancy suggests that the model did not adequately capture the dynamics of Series 1 as effectively as it does for Series 2. Furthermore, there exists a covariance of 11.57 between the residuals of the two series. This indicates a moderate positive correlation between the unexplained variances of the two series. In other words, when the model under- or over-predicts one series, it tends to do the same for the other. However, the relatively low value of the covariance suggests that this relationship is not particularly strong.

Table 4.4: Covariance matrix of residuals from the VAR model

	Series 1	Series 2
Series 1	312.23	11.57
Series 2	11.57	16.56

The correlation matrix of residuals from the VAR model, as shown in Table 4.6, provides information about the relationship between the residuals of the two analyzed series. The diagonal elements of the matrix, which are both 1.0000, indicate a perfect correlation of each series with itself. This is expected. The off-diagonal elements reveal a correlation of 0.1609 between the residuals of Series 1 and Series 2. This positive but weak correlation suggests that there is a slight tendency for the residuals of the two series to move in the same direction. However, the value of 0.1609 is relatively low, indicating that the residuals of the two series are mostly independent. This means that the unexplained variance in one series is not strongly related to the unexplained variance in the other.

Table 4.5: Correlation matrix of residuals from the VAR model

	Series 1	Series 2
Series 1	1.0000	0.1609
Series 2	0.1609	1.0000

4.3 Estimating the Effects of Treatment Interventions (ITN Use, ITN Access, Effective Anti-Malaria Treatment, and Malaria Infection Prevalence) for Malaria Incidence and Mortality Rates in Migori County, Kenya.

4.3.1 Regression Analysis

Table 4.2 presents the regression results for the mortality rate as the response variable. The estimated intercept is -24.5374, but it lacks statistical significance (p = 0.4292). Conversely, the use of insecticide-treated nets (ITNs) shows a statistically significant negative impact on the mortality rate (estimate = -2.0500, p = 0.0175). This indicates that an increase in ITN use is linked to a decrease in malaria mortality. However, it is crucial to note that ITN access has a positive and significant relationship with mortality rates (estimate = 1.8128, p = 0.0169).

The treatment for malaria seemed to have some effectiveness, with a positive estimate of 0.7916. However, this effect is not statistically significant (p = 0.1470), meaning that while the treatment might reduce mortality, the relationship is not very strong in this specific model. Likewise, the prevalence of malaria infection also has a positive estimate of 0.5728, but it is also not statistically significant (p = 0.2871).

The overall model's residual standard error is 2.871 with 8 degrees of freedom. The multiple R-squared value is 0.5879, which means that the model explains approximately 58.79% of the variability in malaria mortality rates. The adjusted R-squared value is lower at 0.3819, indicating a moderate fit for the model. The F-statistic of 2.853 on 4 and 8 degrees of freedom has a p-value of 0.09664, suggesting that the model is marginally significant at the 10% level. This indicates that the predictors collectively contribute some explanatory power to the mortality rates, but it might be beneficial to enhance the model with additional or alternative variables.

Table 4.6: Regression Model Results for Mortality Rate Response

	Estimate	Std Error	t-value	Pr(>|t|)
Intercept	-24.5374	29.4675	-0.833	0.4292
Insect Treated Net Use	-2.0500	0.6870	-2.984	0.0175**
Insect Treated Net Access	1.8128	0.6033	3.005	0.0169**
Anti-Malaria Effective Treatment	0.7916	0.4929	1.606	0.1470
Malaria Infection Prevalence	0.5728	0.5022	1.140	0.2871

Significance level, **P-value <0.05

The regression analysis for the malaria incidence rate as the response variable is presented in Table 4.3 below. The predictive model yields highly significant results for several of the predictors. However, the intercept is not significant, as evidenced by its estimate of -1.8888 and a p-value of 0.91030. This suggests that the baseline incidence rate, without the influence of the predictor variables, is not statistically different from zero, as shown in Table 4.3.

The use of insecticide-treated nets (ITNs) significantly reduces the incidence rates of malaria (-1.4282, p-value = 0.00545), indicating a strong association between higher ITN use and a decrease in malaria cases. However, access to ITNs shows a significant positive relationship with incidence rates (1.3114, p-value = 0.00427). This implies that although more people have access to ITNs, it does not necessarily lead to a decrease in malaria cases, possibly due to issues related to proper usage or other influencing factors.

The estimate for the effectiveness of anti-malaria treatment is 0.5610, with a p-value of 0.07281. While this suggests that the treatment could help reduce malaria incidence, it is not as strong as other variables in the model. On the other hand, the prevalence of malaria infection has a very strong and highly significant positive effect on incidence rates. With an estimate of 13.2330 and a p-value of 4.06e-11, this result shows that as the prevalence of malaria increases, the incidence rates rise dramatically.

Furthermore, the model's residual standard error was 1.583 with 8 degrees of freedom, indicating a relatively low level of unexplained variability. The multiple R-squared value was 0.9991, and the adjusted R-squared was 0.9987, both suggesting a strong model fit in explaining variability in malaria incidence rates. The F-statistic is exceptionally high at 2296, with a p-value of 2.867e-12, confirming the overall significance of the model. These results indicate that the predictors used in this model had a very strong and significant influence on malaria incidence rates in Migori County.

Table 4.7: Regression Model Results for Malaria Incidence Rate Response

	Estimate	Std Error	t-valur	Pr(>|t|)
Intercept	-1.8888	16.2433	-0.116	0.91030
Insect Treated Net Use	-1.4282	0.3787	-3.772	0.00545 **
Insect Treated Net Access	1.3114	0.3326	3.944	0.00427 **
Anti-Malaria Effective Treatment	0.5610	0.2717	2.065	0.07281
Malaria Infection Prevalence	13.2330	0.2769	47.798	<0.0001 ***

Significance level, **P-value <0.05

4.3.2 Vector Autoregression with Exogenous Variables

The VAREX model results, presented in Table 4.7, provide insights into the relationship between malaria incidence and mortality rates. This analysis considers external factors such as insecticide-treated net (ITN) use, ITN access, effective anti-malaria treatment, and malaria infection prevalence. Regarding the incidence series, Series 1.11 and Series 1.12 had minimal impact, with estimates of -0.0157 and 0.0018, respectively. This suggests that past values of these series do not significantly affect current incidence rates. The constant term for the incidence series is positive (39.8727), indicating a baseline level of incidence in the absence of exogenous variables. Among the exogenous variables, ITN use negatively affects incidence (-0.8682), while ITN access and malaria infection prevalence positively influence incidence rates (0.7175 and 13.2048, respectively).

Table 4.8: Vector Autoregression with Exogenous Variables Model Results

	Incidence Series	Mortality Series
Series 1.11	-0.015748690	-0.4767374
Series 2.11	0.450001249	2.6487354
Series 1.12	0.001823247	-0.1241577
Series 2.12	0.114397189	-1.2180552
Constant	39.872684844	492.2120343
Insect Treated Net Use	-0.868202738	-1.2411085
Insect Treated Net Access	0.717548400	1.4198022
Anti-Malaria Effective Treatment	-0.222567827	-7.9697544
Malaria Infection Prevalence	13.204835786	1.6357221

The effects of the exogenous variables are more pronounced in the mortality series. Series 2.11 and Series 2.12 had positive and negative estimates (2.6487 and -1.2181, respectively), suggesting mixed influences of past values on current mortality rates. Notably, the constant term is large (492.2120), indicating a significant baseline level for mortality. ITN use has a negative impact on mortality (-1.2411), while ITN access and malaria infection prevalence had positive effects (1.4198 and 1.6357, respectively). Interestingly, anti-malaria effective treatment had a strong negative effect on mortality (-

7.9698). The VARX model suggests that while ITN use and anti-malaria treatments are critical in reducing both incidence and mortality, other factors such as ITN access and infection prevalence also play significant roles.

4.3.3 Bayesian Vector Autoregression with Exogenous Variables

The Bayesian VAR model is displayed in Table 4.8 below. The model results, which provide valuable insights, are also shown in Table 4.8. After optimizing the hyperparameter lambda to 0.15497, the model was run for 10,000 iterations, including 5,000 burn-in iterations and thinning every 1 iteration. The model achieved an acceptance rate of 97.4%, with 4,872 accepted draws.

According to the median coefficient values from the Bayesian VAR, the constant term for Var 1 is 36.684, while for Var 2 it is 3.530. When considering the lagged variables, Var 1-lag1 has a coefficient of 0.880, suggesting a strong positive impact from its previous value on the current value of Var 1.

On the other hand, the coefficient for Var 2-lag1 is 0.002, indicating a negligible effect of past values on the current value of Var 2. In terms of cross-variable relationships, the coefficient for Var 2-lag1 in Var 1 is -1.877, indicating a significant negative impact. Meanwhile, Var 2-lag1 has a coefficient of 0.706 on itself, revealing a moderate positive influence of past values on the current value of Var 2. These results highlight the dynamic relationships between the variables, with Var 1 showing notable persistence and Var 2 demonstrating varying degrees of influence.

Table 4.9: Median values coefficient values from a BVAR

	Var 1	Var 2
Constant	36.684	3.530
var1-lag1	0.880	0.002
var2-lag1	-1.877	0.706

Table 4.9 displays the median values for the variance-covariance matrix derived from the Bayesian VAR model. These values provide insights into the variability and relationship between the variables. Var 1 has a variance of 183.153, indicating high variability in its residuals. On the other hand, Var 2 has a lower variance of 10.482, suggesting less variability in its residuals. The covariance between Var 1 and Var 2 is 13.259, indicating a positive relationship between their residuals, although the magnitude is moderate. The log-likelihood value of -79.50488 serves as a measure of the model's fit, with more negative values indicating a less optimal fit. Overall, the results indicate that Var 1 exhibits considerable variability, while Var 2 is more stable. Additionally, there is a modest positive covariance between the residuals of the two variables.

Table 4.10: Median values for variance-covariance values from a BVAR

	Var 1	Var 2
Var 1	183.153	13.259
Var 2	13.259	10.482

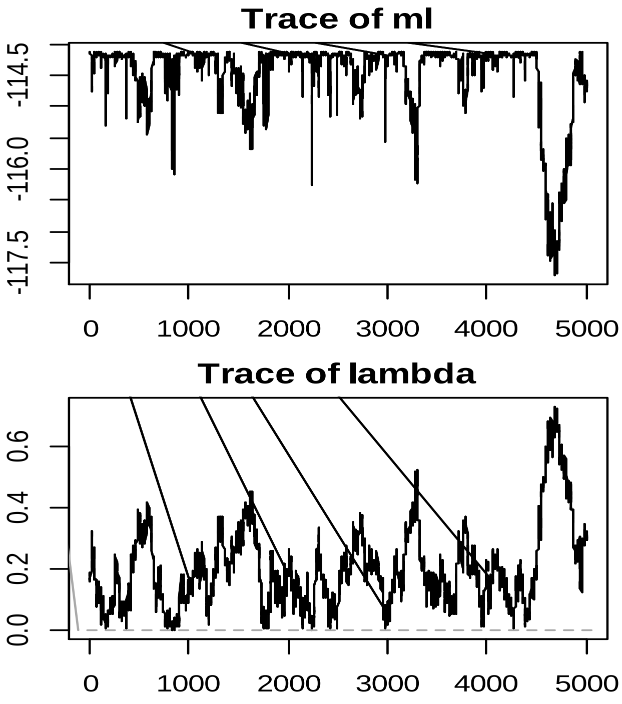


Figure 4.8: Convergence check of the hyperparameters with a trace and density plot

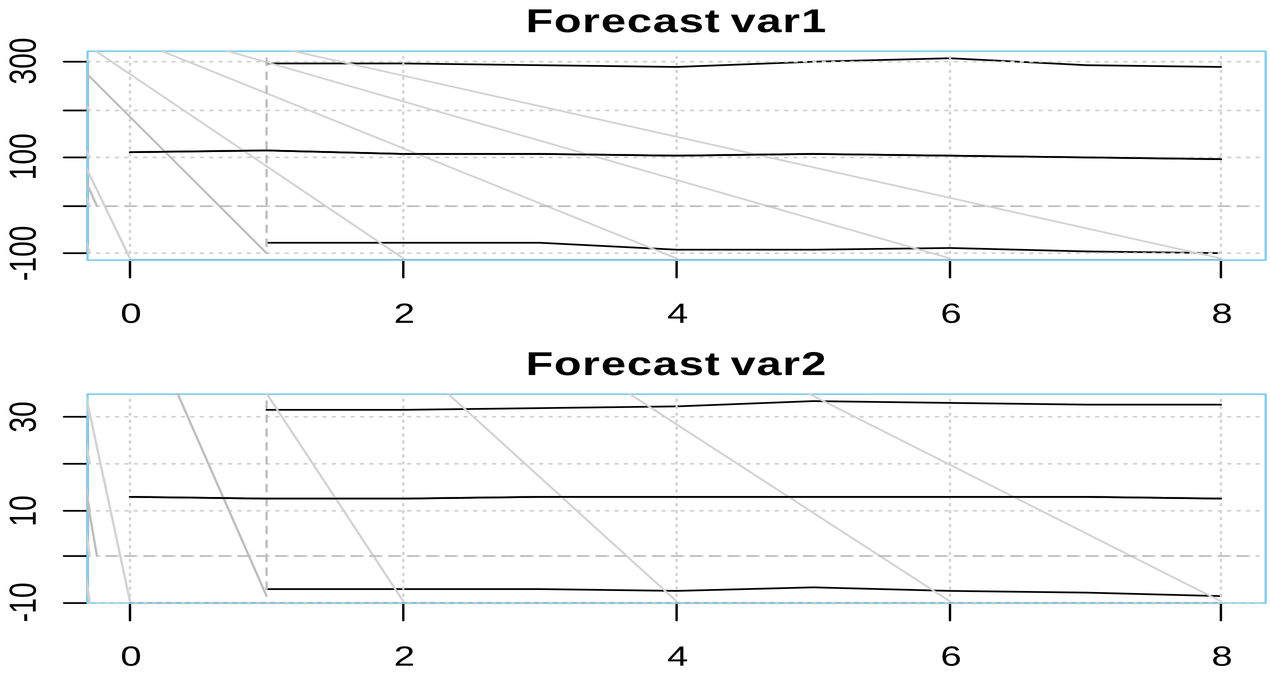


Figure 4.9: Plot forecasts

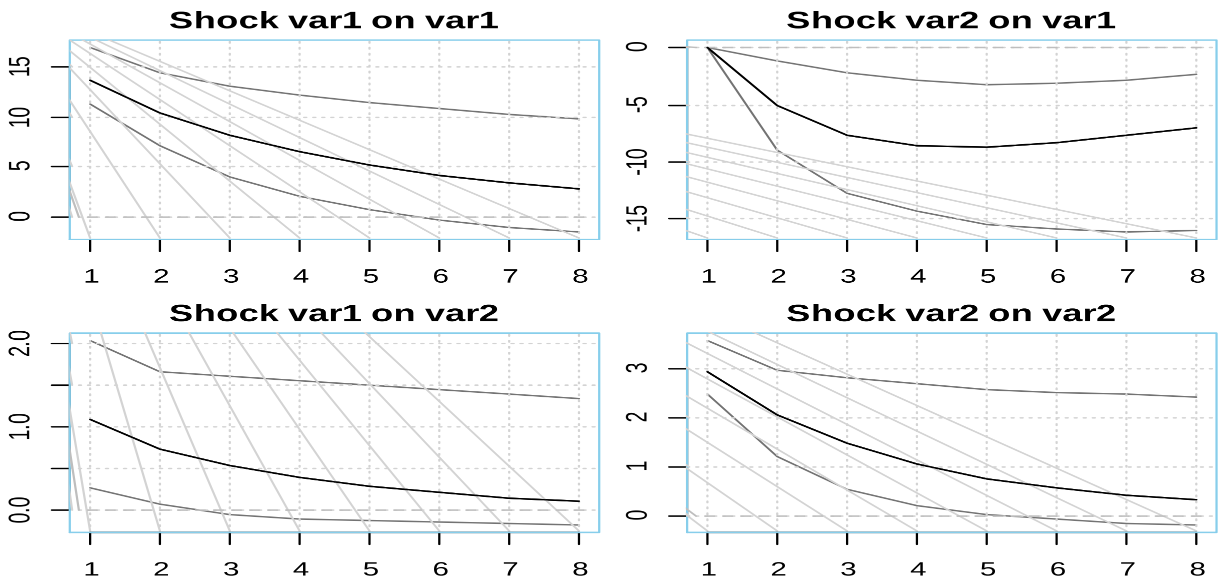


Figure 4.10: BVARX Model Impulse Responses

4.4 Residual Analysis

4.4.1 Regression Model Residuals

Table 4.10 presents a summary of the residuals for both the mortality and incidence rates in the regression model. The mortality rate residuals range from -3.4965 to 4.1801, with a median of -0.0283. This suggests that the residuals are centered around zero, but there are significant deviations. The interquartile range, which is from -2.1304 to 1.4765, shows that the middle 50% of the residuals include negative and positive values, indicating variability in the model fit.

For the incidence rate, the residuals range from -2.3257 to 2.7102, with a median of -0.0206. The interquartile range is narrower, ranging from -0.5273 to 0.5094. This indicates less dispersion in the residuals compared to the mortality rates. These findings suggest that the model fits the incidence rate data with relatively less variability. However, the wider range of deviations in the mortality rate residuals indicates potential areas for further refinement of the model.

Table 4:11: Regression Model Residuals

	Min	1^{ST} Q	Med	3^{RD} Q	Max
Malaria Mortality Rate	-3.4965	-2.1304	-0.0283	1.4765	4.1801
Malaria Incidence Rate	-2.32568	-0.52733	-0.02056	0.50942	2.71017

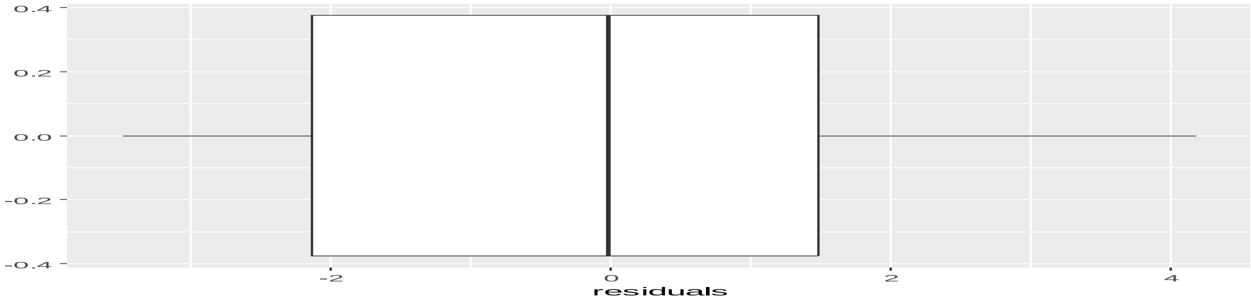


Figure 4.11: Boxplot of Residuals

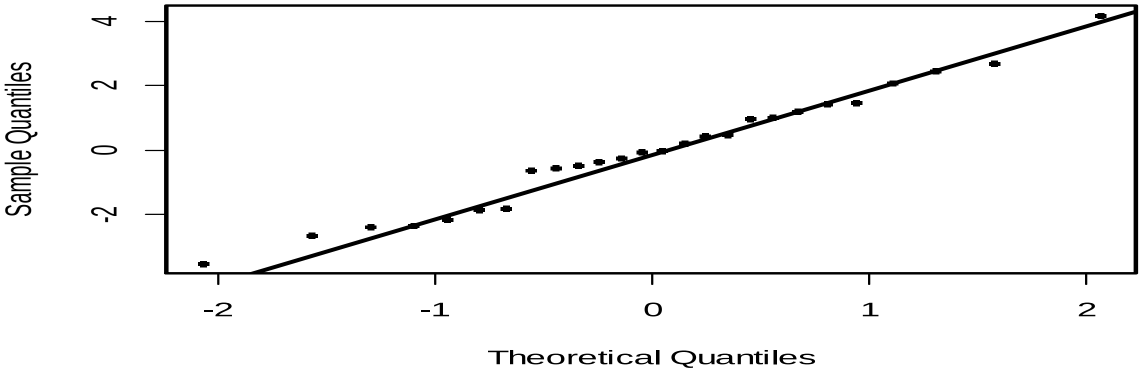


Figure 4.12: Q-Q plot of Residuals

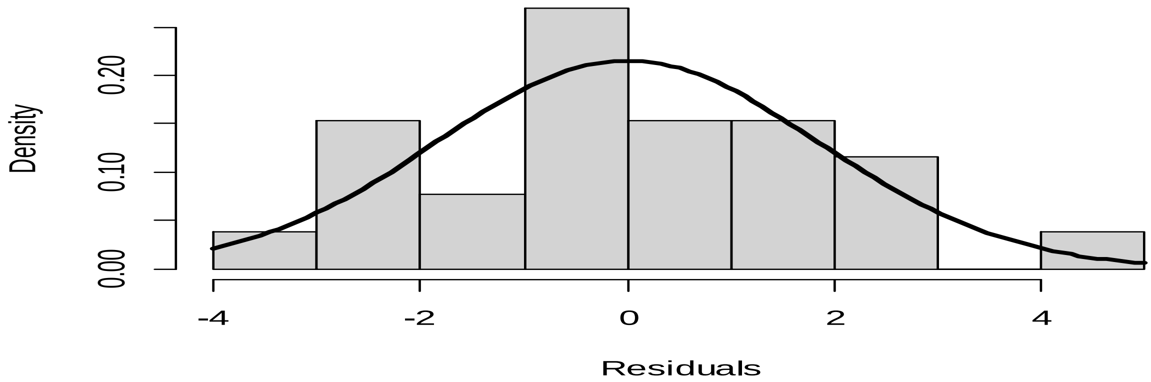


Figure 4.13: Histogram of Residuals

4.4.2 VAR Model Residuals

Table 4.11 presents the residuals from the Vector Autoregression (VAR) model for two series, highlighting their distribution and variability. For Series 1, the residuals range from a minimum of -17.524 to a maximum of 23.154, with a median of -0.253 and a mean of 0.000. The first quartile (1st Q) is -10.855 and the third quartile (3rd Q) is 10.088, indicating a substantial spread and considerable variability in the residuals. This wide range suggests that the VAR model did not fully capture the dynamics of Series 1, leading to large deviations.

In contrast, Series 2 has a narrower range of residuals, from -3.7386 to 4.3613, with a median of 0.06774 and a mean of 0.000. The interquartile range for Series 2 is from -3.1424 to 3.0428, reflecting less variability compared to Series 1. These results suggest that while Series 2 residuals are more tightly clustered around zero, indicating a potentially better model fit, Series 1 residuals exhibit more variability, pointing to potential areas for further model refinement.

Table 4:12: Vector Autoregression Model Residuals

	Min	1^{ST} Q	Med	Mean	3^{RD} Q	Max
Series 1	-17.524	-10.855	-0.253	0.000	10.088	23.154
Series 2	-3.73863	-3.14240	0.06774	0.000	3.04275	4.36129

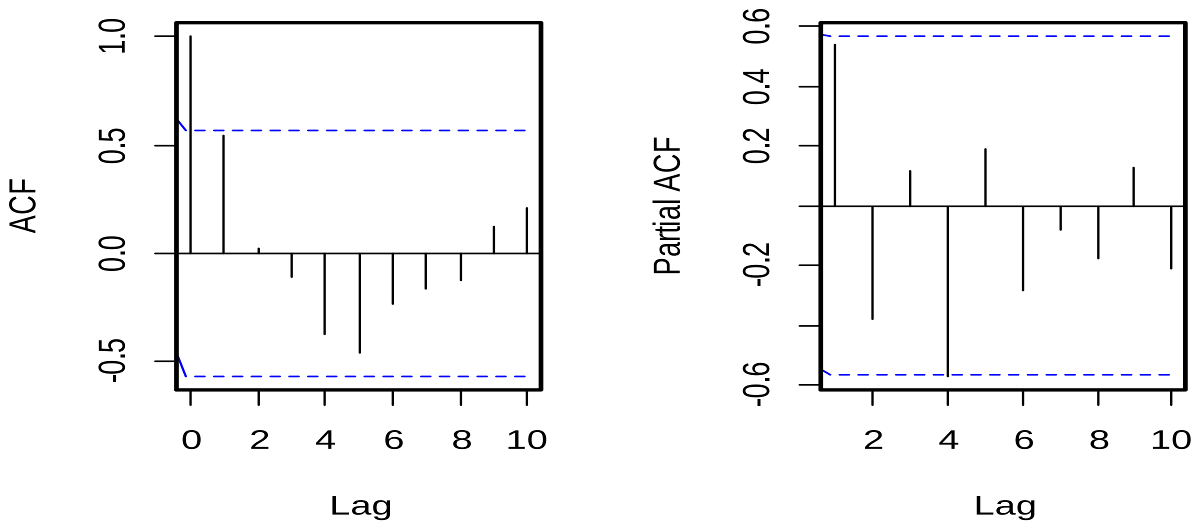


Figure 4.14: ACF and PACF plot for Malaria Incidence Rate residuals

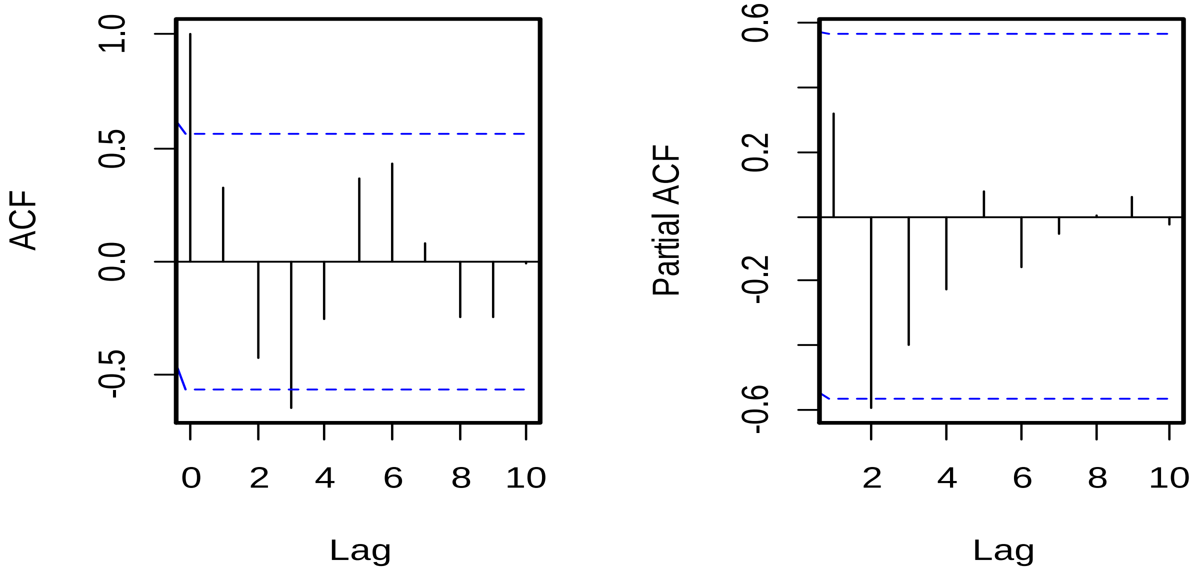


Figure 4.15: ACF and PACF plot for Malaria Mortality Rate residuals

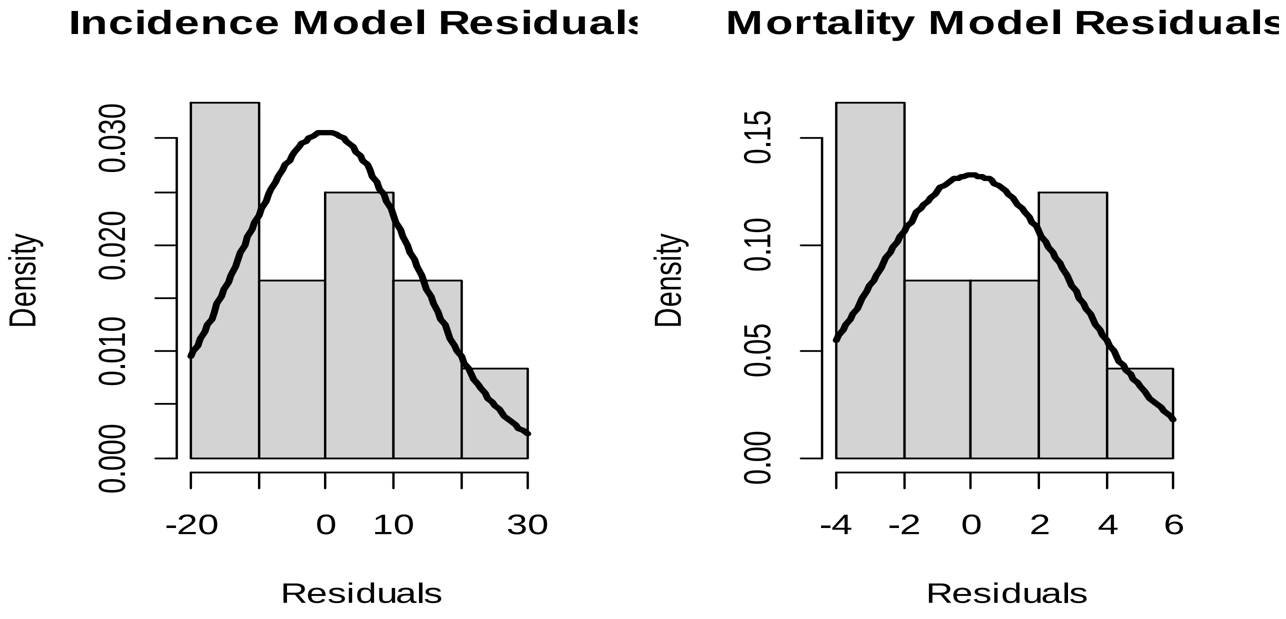


Figure 4.16: Histogram of VAR model Residuals

4.4.3 VARX Model Residuals

Table 4.12 displays the residuals from the Vector Autoregression with Exogenous Variables (VARX) model, indicating an improved fit compared to the previous VAR model. For Series 1, the residuals range from -0.77912 to 0.89667, with a median of 0.03469 and a mean of 0.000. The first quartile (1st Q) is -0.38513, and the third quartile (3rd Q) is 0.41159, suggesting a relatively narrow range of residuals and a better fit of the model to the data. The reduced spread in residuals implies that the inclusion of exogenous variables has enhanced the model's ability to capture the underlying dynamics of Series 1.

Similarly, for Series 2, the residuals range from -0.45035 to 0.61487, with a median of -0.02838 and a mean of 0.000. The first quartile is -0.12732, and the third quartile is 0.16708, indicating a tighter distribution of residuals around zero. These results reflect a more accurate fit for Series 2 as well, demonstrating that the VARX model effectively accounts for the variability in both series, leading to improved model performance.

Table 4:13: Vector Autoregression with Exogenous Variables Model Residuals

	Min	1ST Quartile	Median	Mean	3RD Quartile	Max
Series 1	-0.77912	-0.38513	0.03469	0.000	0.41159	0.89667
Series 2	-0.45035	-0.12732	-0.02838	0.000	0.16708	0.61487

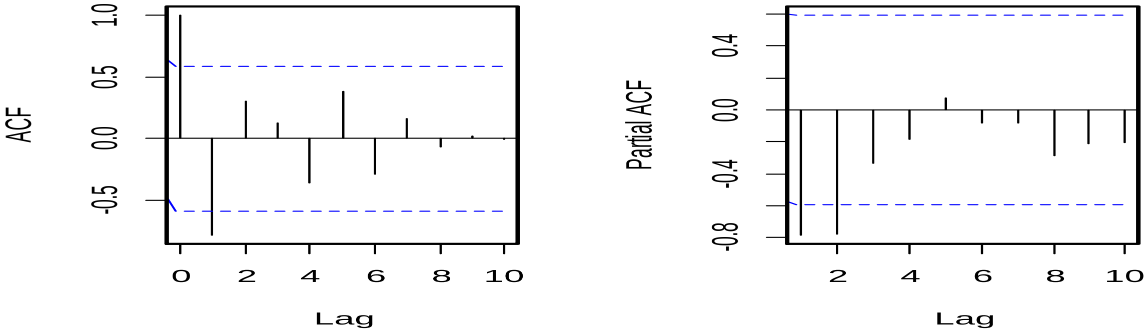


Figure 4.17: ACF and PACF Malaria Incidence Model residuals

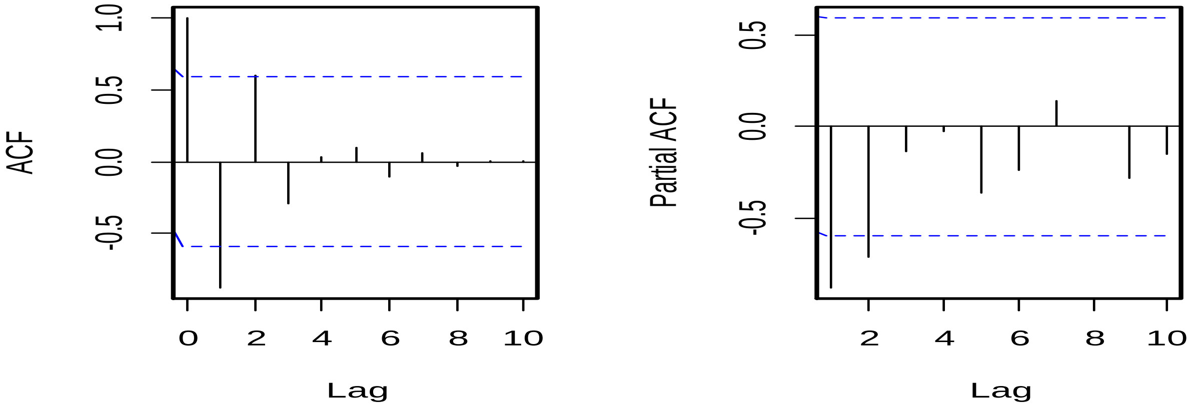


Figure 4.18: ACF and PACF plot for Malaria Mortality Rate Model residuals

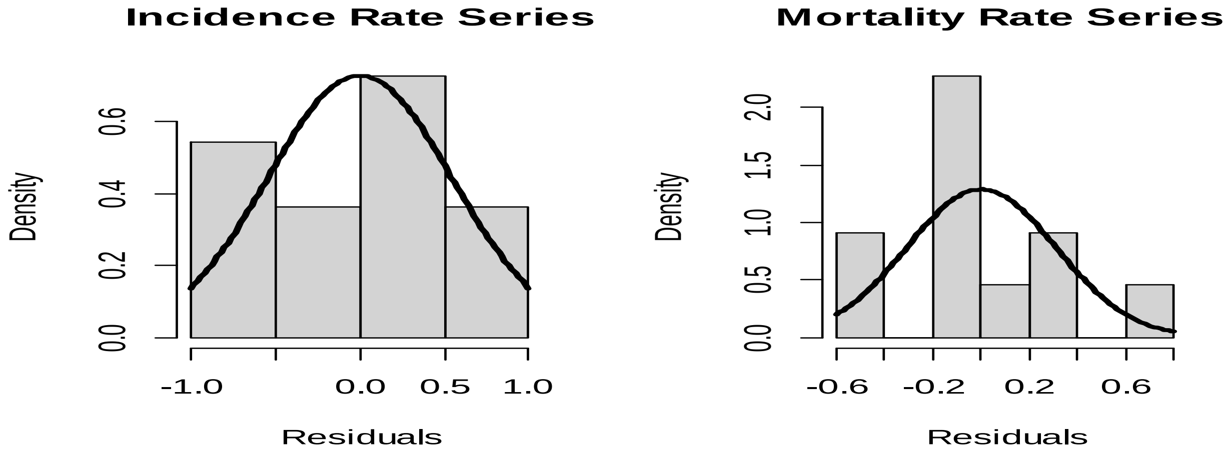


Figure 4.19: Histogram of Residuals for the VARX Model

4.5 Assessment of the goodness of fit of the model for malaria incidence and mortality rates in Migori County, Kenya

In evaluating the goodness of fit of the fitted models to evaluate the effect of treatment interventions on the malaria mortality and incidence rates, ANOVA model was employed and the results are illustrated below.

4.5.1 Analysis of Variance (ANOVA)

Table 4.13 presents the analysis of variance (ANOVA) for the regression model, indicating the significance of different predictors in explaining the variability in the response variables. The results demonstrate that the intercept is highly significant with a Pillai trace of 0.99992 and an approximate F-statistic of 46,355, resulting in a p-value of 3.739e-15. This suggests that the baseline model fit is statistically robust. Insect Treated Net Use also shows a significant impact on the response variable, with a Pillai trace of 0.96664 and an F-statistic of 101, yielding a p-value of 6.780e-06.

Similarly, Insect Treated Net Access has a Pillai trace of 0.75017 and an F-statistic of 11, indicating a statistically significant effect, although less pronounced compared to Insect Treated Net Use. Anti-Malaria Effective Treatment demonstrates a remarkably high Pillai trace of 0.99880 and an F-statistic of 2,910, highlighting its strong and significant influence on the response. Malaria Infection Prevalence also exhibits a significant effect with a Pillai trace of 0.99652, an F-statistic of 1,001, and a p-value of 2.496e-09. These collective results signify that all the predictors have a statistically significant impact on the response variables, with Insect Treated Net Use and Anti-Malaria Effective Treatment being particularly influential.

Table 4:14: Analysis of Variance for the Regression Model

	DF	Pillai	Approx F	Pr(>F)
Intercept	1	0.99992	46355	3.739e-15 ***
Insect Treated Net Use	1	0.96664	101	6.780e-06 ***
Insect Treated Net Access	1	0.75017	11	0.007794 **
Anti-Malaria Effective Treatment	1	0.99880	2910	6.008e-11 ***
Malaria Infection Prevalence	1	0.99652	1001	2.496e-09 ***

4.5.2 Normality Test

The Shapiro-Wilk normality test for the regression model residuals yields a W statistic of 0.98184 and a p-value of 0.9104. This high p-value indicates that the residuals do not significantly deviate from a normal distribution. In other words, the test fails to reject the null hypothesis that the residuals are normally distributed. This suggests that the assumption of normality for the residuals is reasonably satisfied, which is a positive indicator for the validity of the regression model's statistical inferences and assumptions.

The Lilliefors (Kolmogorov-Smirnov) normality test for the regression model residuals produces a D statistic of 0.10386 and a p-value of 0.6675. This p-value is quite high, indicating that there is no significant deviation from the normal distribution. Therefore, the test fails to reject the null hypothesis that the residuals are normally distributed. This result supports the assumption of normality for the residuals, suggesting that the residuals fit well with the normal distribution, which is beneficial for the robustness and reliability of the regression model's statistical assumptions and results. The results of the normality tests conducted on the residuals of the VAR model indicate that the residuals are approximately normally distributed.

Specifically, the Jarque-Bera (JB) test shows a chi-squared statistic of 1.8063 and a p-value of 0.7713, indicating no significant deviation from normality when considering both skewness and kurtosis. Further supporting this, the skewness test reveals a chi-squared statistic of 0.12676 with a p-value of 0.9386, suggesting that the residuals are not significantly skewed and are symmetrically distributed. The kurtosis test also strengthens the case for normality, with a chi-squared statistic of 1.6795 and a p-value of 0.4318, indicating normal kurtosis similar to the normal distribution's peakedness. Overall, these results provide further evidence for the assumption of normality of the residuals, ultimately validating the statistical assumptions and inferences of the VAR model.

The normality tests for the residuals of the VARX model provided strong evidence to support the assumption of normality. The Jarque-Bera (JB) test yielded a chi-squared statistic of 0.95548 with 4 degrees of freedom and a p-value of 0.9165. This indicates that the residuals do not significantly deviate from a normal distribution when considering both skewness and kurtosis. The skewness test resulted in a chi-squared statistic of 0.011777 with 2 degrees of freedom and a p-value of 0.9941, showing that the residuals are essentially symmetrically distributed with no significant skew. Similarly, the kurtosis test yielded a chi-squared statistic of 0.9437 with 2 degrees of freedom and a p-value of 0.6238, suggesting that the residuals have a distribution with normal kurtosis, indicating a typical level of peakedness and tail behavior. Overall, these findings confirm that the residuals of the VAR model1 align well with the assumption of normality, supporting the robustness of the model's statistical inferences.

4.5.3 Ljung-Box test

The Ljung-Box test results indicate that there is differing autocorrelation across the two series in the residuals of the VAR model. The first series has a p-value of 0.03465, suggesting significant autocorrelation, while the second series had a p-value of 0.2071, indicating no significant autocorrelation. This means that the second series is adequately modeled without significant serial correlation, but the first series did require some refinement in order to address the autocorrelation in its residuals. In summary, the VAR model captures the dynamics of the second series well, but improvements are needed to account for the autocorrelation in the residuals of the first series.

The results of the Ljung-Box test for the residuals of the VARX model reveal significant autocorrelation in both series. The Malaria Incidence series shows a X-squared statistic of 8.7119 with 1 degree of freedom and a p-value of 0.003161, indicating strong autocorrelation in the residuals. This suggests that the VARX model did not have fully captured the serial dependencies in the incidence data. Similarly, the Malaria Mortality series exhibits a X-squared statistic of 10.938 with 1 degree of freedom and a p-value of 0.0009423, which also indicates significant autocorrelation. This means that the model has not adequately addressed the serial correlation in the mortality data. These findings demonstrate that although the VARX model improves upon the basic VAR model, further refinement is necessary to better account for and capture the autocorrelation present in the residuals of both series.

4.6 The prediction of malaria incidence and mortality in Migori County, Kenya

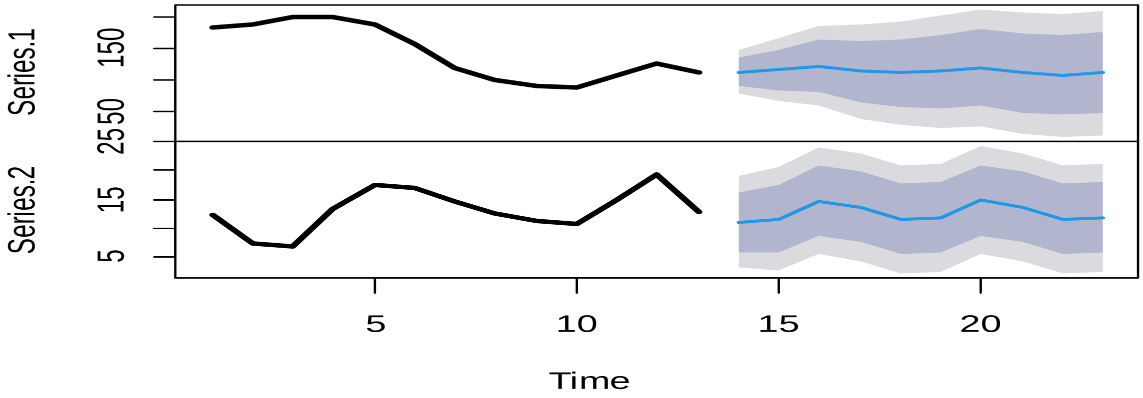


Figure 4.20: Fitted Model Forecasts

CHAPTER FIVE

DISCUSSIONS

5.1 Introduction

This chapter covers the summary of the findings, the conclusion and recommendations.

5.2 The analysis of the spread of malaria incidence and mortality using Vector Autoregression Models with Exogenous Variables in Migori County, Kenya.

This study applied statistical models to analyze malaria incidence and mortality rates in Migori County, Kenya. Results showed that the malaria mortality rates ranged from a minimum of 6.85 to a maximum of 19.37, with a mean of 13.07. Likewise, the malaria incidence rates varied between 87.14 and 199.58, with a mean of 142.57 respectively. The primary focus was on the analysis of malaria incidence and mortality with application of vector auto regression models

The findings indicated a decrease in the Malaria Incidence Rate under Series 1 over time, while the Malaria Mortality Rate under Series 2 remained relatively constant when using the Vector Autoregression model. However, the Autocorrelation Function (ACF) analysis of both the Malaria Incidence Rate (Series 1) and the Malaria Mortality Rate (Series 2) showed that the time series data were stationary over time. These results are similar to those of a study conducted in India, which observed a linear decline in malaria cases attributed to anti-malarial health interventions introduced in the country. These interventions led to a 14% reduction in malaria cases between 2011 and 2022, as estimated using ARIMA and Holt’s time series regression models (Singh *et al.*, 2024).

5.3 Estimating treatment intervention effects of ITN use, ITN access the model parameters and assessment the goodness of fit in analyzing malaria incidence and mortality rates in Migori County, Kenya.

The regression analysis revealed that Insecticide-Treated Net (ITN) use and effective anti-malaria treatment were significant predictors of both malaria mortality and incidence rates. The results show that ITN use had a substantial negative impact on mortality rates. Interestingly, effective anti-malaria treatment had a strong positive effect on mortality rates, underscoring its crucial role in reducing malaria-related deaths. A similar study found that ITNs and other anti-malaria prevention strategies were cost-effective in Bangladesh (Haque *et al.*, 2014) and Sri Lanka, which has been declared malaria-free since 2016 (Simac *et al.*, 2017). Furthermore, the multivariate effects of the exogenous variables on the mortality series 2.11 and series 2.12 showed both positive and negative effects (2.6487 and -1.2181, respectively), thus suggesting a mixed influences of past values on current malaria mortality rates

However, the overall fit of the regression model was somewhat mixed, with residuals showing a range of values that suggest some variability not fully captured by the model. Normality tests of residuals from this model supported that the residuals were approximately normally distributed, as indicated by high p-values in both the Shapiro-Wilk and Lilliefors tests. The results obtained are in line with (Hyndman and Athanasopoulos, 2018) on the principles of time series forecasting.

5.4 Predicting malaria incidence and mortality in Migori County, Kenya.

The VAR model results offered insights into the dynamics between the two series over time. The residuals from the VAR model showed considerable variability, especially for

Series 1, with a wide range of values and high standard deviations. This variability suggested potential issues with model fit and the presence of autocorrelation. The Ljung-Box test further confirmed significant autocorrelation in the residuals for Series 1, indicating that the model might not fully account for serial dependencies. In contrast, the residuals for Series 2 exhibited less variability and were more consistent with the normality assumption.

The VARX model, which incorporated exogenous variables, improved the fit of the model. The residuals from the VARX model exhibited a narrower range of values, indicating a better fit compared to the basic VAR model. However, the Ljung-Box test revealed that significant autocorrelation remained in both the malaria incidence and mortality series residuals, suggesting potential areas for further refinement. This indicates that while including exogenous variables enhanced the model, it did not completely resolve the issues with residual autocorrelation.

Additionally, Bayesian VAR (BVAR) modeling provided further insights into the data dynamics. The median coefficient values from the BVAR analysis revealed meaningful effects of lagged variables and exogenous predictors. However, the variance-covariance results indicated a high degree of residual variability. The normality tests and variance-covariance results from the BVAR model were consistent with those from the other models, supporting the robustness of the findings while also highlighting areas for improvement. Overall, these results emphasize the need for continued refinement and exploration of modeling approaches to better capture the complexities of malaria incidence and mortality dynamics in the study area.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusions

This study thoroughly analyzes the incidence and mortality rates of malaria in Migori County, Kenya. It employs multiple statistical modeling approaches to provide a comprehensive understanding of the data. The study's main findings are as follows:

- The modeling of malaria incidence and mortality rates dynamics remain a challenging task with presence of temporal variations utilizing time series data.
- The regression analysis demonstrated that factors like Insect Treated Net Use and Anti-Malaria Effective Treatment have a significant impact on both mortality and incidence rates. However, the residuals of the model showed variability and indicated a need for better fit. This was evident from the mixed results obtained from normality tests. Therefore, although the regression model provides valuable insights, it does not fully consider all the factors that influence malaria metrics. Hence, there is a need for improvement in the model.
- The use of Vector Autoregression (VAR) and Vector Autoregression with Exogenous Variables (VAREX) models has helped us gain a better understanding of the relationship between malaria incidence and mortality over time. While the VAR model showed that there were issues with model specification and parameterization due to significant autocorrelation in the residuals of the first series, the VARX model improved upon this by incorporating exogenous variables and providing a better fit. However, the Ljung-Box test results for the residuals of the VAREA model still indicated significant autocorrelation, suggesting that

further refinement is necessary. These findings highlight the complexity of capturing the temporal dynamics of malaria data and suggest the need for other modern modeling approaches.

- The Bayesian VAR (BVAR) model provided additional insights, but it also showed significant residual variability. The BVAR analysis confirmed the findings of other models, with normality tests and variance-covariance estimates aligning with previous results. The consistency across different models highlights the robustness of the study's findings, but it also highlights the ongoing difficulties in accurately modeling malaria incidence and mortality. In conclusion, this study underscores the need to improve statistical models to increase their explanatory power and accuracy, ultimately leading to more effective strategies for controlling malaria and improving public health outcomes in Migori County, Kenya.

6.2 Recommendations

Based on the findings of this investigation, the study proposes many methodological adjustments to increase the accuracy and robustness of future studies relating to this study.

- Addressing the identified autocorrelation in residuals, particularly for the VAR and VARX models, is crucial. This can be done by exploring alternative specifications or incorporating additional lagged terms to better capture the temporal dependencies within the data.
- In addition, using advanced model techniques such as Generalized Method of Moments (GMM) or Machine Learning approaches can offer more valuable insights and enhance the model's accuracy.

- Additionally, incorporating methods like residual diagnostics and robust standard error corrections can assist in refining the model and resolving any remaining issues related to autocorrelation and heteroskedasticity.

To enhance our understanding of malaria dynamics, future studies should consider incorporating additional variables. These variables should encompass environmental, socio-economic, and behavioral factors. For instance, local climate conditions, healthcare infrastructure, and socio-economic indicators such as poverty rates influence malaria incidence and mortality rates. Furthermore, exploring the effects of intervention programs and seasonal variations would provide valuable insights into their impact on malaria trends. Expanding the dataset to encompass more regions and a longer time span is also recommended to improve the generalizability of the findings and obtain a more comprehensive understanding of malaria dynamics in different contexts.

REFERENCES

Achieng, E., Otieno, V., & Mung'atu, J. (2020). Modeling the trend of reported malaria cases in Kisumu County, Kenya. *F1000Research*, *9*, 600. https://doi.org/10.12688/f1000research.23259.1

Amadi, M., & Erandi, K. K. W. H. (2024). Assessing the relationship between malaria incidence levels and meteorological factors using cluster-integrated regression. *BMC Infectious Diseases*, *24*, 664. https://doi.org/10.1186/s12879-024-09570-z

Ankamah, S., Nokoe, K. S., & Iddrisu, W. A. (2018). Modelling Trends of Climatic Variability and Malaria in Ghana Using Vector Autoregression. *Malaria Research and Treatment*, *2018*(1), 6124321. https://doi.org/10.1155/2018/6124321

Aregawi, M., Lynch, M., Bekele, W., Kebede, H., Jima, D., Taffese, H. S., Yenehun, M. A., Lilay, A., Williams, R., Thomson, M., Nafo-Traore, F., Admasu, K., Gebreyesus, T. A., & Coosemans, M. (2014). Time Series Analysis of Trends in Malaria Cases and Deaths at Hospitals and the Effect of Antimalarial Interventions, 2001–2011, Ethiopia. *PLOS ONE*, *9*(11), e106359. https://doi.org/10.1371/journal.pone.0106359

Badaoui, F., Bouhout, S., Amar, A., & Khomsi, K. (2023). Modelling of Leishmaniasis Infection Dynamics: A Comparative Time Series Analysis with VAR, VECM, Generalized Linear and Markov Switching Models. *Engineering Proceedings*, *39*(1), Article 1. https://doi.org/10.3390/engproc2023039038

Bayarbat, U., & Li, Y. (2020). Empirical Analysis of Relationship between Per Capita Health Expenditure and Economic Growth Based on Vector Autoregressive Model (VAR) in Mongolia. *Theoretical Economics Letters*, *10*(1), Article 1. https://doi.org/10.4236/tel.2020.101010

Chaves, L. F., Hashizume, M., Satake, A., & Minakawa, N. (2012). Regime shifts and heterogeneous trends in malaria time series from Western Kenya Highlands. *Parasitology*, *139*(1), 14–25. https://doi.org/10.1017/S0031182011001685

Chang, L., & Shi, Y. (2024). A discussion on the robust vector autoregressive models: Novel evidence from safe haven assets. *Annals of Operations Research*, *339*(3), 1725–1755. https://doi.org/10.1007/s10479-022-04919-6

Cordoni, F., Dorémus, N., & Moneta, A. (2024). Identification of Vector Autoregressive Models with Nonlinear Contemporaneous Structure. *Journal of Economic Dynamics and Control*, *162*, 104852. https://doi.org/10.1016/j.jedc.2024.104852

Darkoh, E. L., Larbi, J. A., & Lawer, E. A. (2017). A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Malaria Research and Treatment*, *2017*, 7820454. https://doi.org/10.1155/2017/7820454

Diao, O., Absil, P.-A., & Diallo, M. (2023). Generalized Linear Models to Forecast Malaria Incidence in Three Endemic Regions of Senegal. *International Journal of Environmental Research and Public Health*, *20*(13), 6303. https://doi.org/10.3390/ijerph20136303

Djihinto, O. Y., Medjigbodo, A. A., Gangbadja, A. R. A., Saizonou, H. M., Lagnika, H. O., Nanmede, D., Djossou, L., Bohounton, R., Sovegnon, P. M., Fanou, M.-J., Agonhossou, R., Akoton, R., Mousse, W., & Djogbénou, L. S. (2022). Malaria-Transmitting Vectors Microbiota: Overview and Interactions With Anopheles Mosquito Biology. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.891573

Eunice, A., Wanjoya, A., & Luboobi, L. (2017). Statistical Modeling of Malaria Incidences in Apac District, Uganda. *Open Journal of Statistics*, *7*(6), Article 6. https://doi.org/10.4236/ojs.2017.76063

Guefano, S., Tamba, J. G., Azong, T. E. W., & Monkam, L. (2021). Forecast of electricity consumption in the Cameroonian residential sector by Grey and vector autoregressive models. *Energy*, *214*, 118791. https://doi.org/10.1016/j.energy.2020.118791

Guissou, E., Da, D. F., Hien, D. F. de S., Yameogo, K. B., Yerbanga, S. R., Ouédraogo, G. A., Dabiré, K. R., Lefèvre, T., & Cohuet, A. (2023). Intervention reducing malaria parasite load in vector mosquitoes: No impact on Plasmodium falciparum extrinsic incubation period and the survival of Anopheles gambiae. *PLOS Pathogens*, *19*(5), e1011084. https://doi.org/10.1371/journal.ppat.1011084

Hoi, A. G., Greischar, M. A., & Mideo, N. (2024). Limited impact of within-vector ecology on the evolution of malaria parasite transmission investment. *Frontiers in Malaria*, *2*. https://doi.org/10.3389/fmala.2024.1392060

Hou, P. S., Fadzil, L. M., Manickam, S., & Al-Shareeda, M. A. (2023). Vector Autoregression Model-Based Forecasting of Reference Evapotranspiration in Malaysia. *Sustainability*, *15*(4), Article 4. https://doi.org/10.3390/su15043675

Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, *142*, 319–327. https://doi.org/10.1016/j.envres.2015.06.040

Karlsson, S., Mazur, S., & Nguyen, H. (2023). Vector autoregression models with skewness and heavy tails. *Journal of Economic Dynamics and Control*, *146*, 104580. https://doi.org/10.1016/j.jedc.2022.104580

Koné, I., & Atchadé, M. N. (2023). Spatial econometric modeling of malaria distribution in Burkina Faso. *Scientific African*, *21*, e01869. https://doi.org/10.1016/j.sciaf.2023.e01869

Laguna, F., Grillet, M., León, J., & Ludeña, C. (2017). Modelling malaria incidence by an autoregressive distributed lag model with spatial component. *Spatial and Spatio-Temporal Epidemiology*, *22*. https://doi.org/10.1016/j.sste.2017.05.001

Muriithi, D., Lumumba, V. W., & Okongo, M. (2024). A Machine Learning-Based Prediction of Malaria Occurrence in Kenya. *American Journal of Theoretical and Applied Statistics*, *13*(4), Article 4. https://doi.org/10.11648/j.ajtas.20241304.11

Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*, 100508. https://doi.org/10.1016/j.imu.2020.100508

Nsor-Ambala, R., & Bugri Anarfo, E. (2022). A vector autoregression (VAR) analysis of corruption, economic growth, and foreign direct investment in Ghana. *Cogent Economics & Finance*, *10*(1), 2146631. https://doi.org/10.1080/23322039.2022.2146631

Nyawanda, B. O., Beloconi, A., Khagayi, S., Bigogo, G., Obor, D., Otieno, N. A., Lange, S., Franke, J., Sauerborn, R., Utzinger, J., Kariuki, S., Munga, S., & Vounatsou, P. (2023). The relative effect of climate variability on malaria incidence after scale-up of interventions in western Kenya: A time-series analysis of monthly incidence data from 2008 to 2019. *Parasite Epidemiology and Control*, *21*, e00297. https://doi.org/10.1016/j.parepi.2023.e00297

Ogbuagada, S. O., Okolo, A., Torsen, E., & John, O. T. (2022). Multivariate time series analysis in modelling malaria cases in Jimeta metropolis of Adamawa state, Nigeria. *Fudma journal of sciences*, *6*(3), Article 3. https://doi.org/10.33003/fjs-2022-0603-970

Okiro, E. A., Alegana, V. A., Noor, A. M., & Snow, R. W. (2010). Changing malaria intervention coverage, transmission and hospitalization in Kenya. *Malaria Journal*, *9*(1), 285. https://doi.org/10.1186/1475-2875-9-285

Ototo, E. N., Ogutu, J. O., Githeko, A., Said, M. Y., Kamau, L., Namanya, D., Simiyu, S., & Mutimba, S. (2022). Forecasting the Potential Effects of Climate Change on Malaria in the Lake Victoria Basin Using Regionalized Climate Projections. *Acta Parasitologica*, *67*(4), 1535–1563. https://doi.org/10.1007/s11686-022-00588-4

Oyegoke, O. O., Adewumi, T. S., Aderoju, S. A., Tsundzukani, N., Mabunda, E., Adeleke, M. A., Maharaj, R., & Okpeku, M. (2023). Towards malaria elimination: Analysis of travel history and case forecasting using the SARIMA model in Limpopo Province. *Parasitology Research*, *122*(8), 1775–1785. https://doi.org/10.1007/s00436-023-07870-y

Rocha, E. M., Katak, R. de M., Campos de Oliveira, J., Araujo, M. da S., Carlos, B. C., Galizi, R., Tripet, F., Marinotti, O., & Souza-Neto, J. A. (2020). Vector-Focused Approaches to Curb Malaria Transmission in the Brazilian Amazon: An Overview of Current and Future Challenges and Strategies. *Tropical Medicine and Infectious Disease*, *5*(4), 161. https://doi.org/10.3390/tropicalmed5040161

Savi, M. K. (2023). An Overview of Malaria Transmission Mechanisms, Control, and Modeling. *Medical Sciences*, *11*(1), Article 1. https://doi.org/10.3390/medsci11010003

Shah, I., Mubassir, P., Ali, S., & Albalawi, O. (2024). A functional autoregressive approach for modeling and forecasting short-term air temperature. *Frontiers in Environmental Science*, *12*. https://doi.org/10.3389/fenvs.2024.1411237

Singh, M. P., Rajvanshi, H., Bharti, P. K., Anvikar, A. R., & Lal, A. A. (2024). Time series analysis of malaria cases to assess the impact of various interventions over the last three decades and forecasting malaria in India towards the 2030 elimination goals. *Malaria Journal*, *23*(1), 50. https://doi.org/10.1186/s12936-024-04872-8

Sugita, K. (2022). Forecasting with Bayesian vector autoregressive models: Comparison of direct and iterated multistep methods. *Asian Journal of Economics and Banking*, *6*(2), 142–154. https://doi.org/10.1108/AJEB-04-2022-0044

Sukono, S., Ibrahim, R. A., Riaman, R., Hertini, E., Hidayat, Y., & Saputra, J. (2023). Forecasting the Effectiveness of COVID-19 Vaccination Using Vector Autoregressive with an Exogenous Variable: On the Cases of COVID-19 in Indonesia. *Discrete Dynamics in Nature and Society*, *2023*(1), 6285328. https://doi.org/10.1155/2023/6285328

Thomas, A., Bakai, T. A., Atcha-Oubou, T., Tchadjobo, T., Rabilloud, M., & Voirin, N. (2024). Exploring malaria prediction models in Togo: A time series forecasting by health district and target group. *BMJ Open*, *14*(1), e066547. https://doi.org/10.1136/bmjopen-2022-066547

Virolainen, S. (2024). A Statistically Identified Structural Vector Autoregression with Endogenously Switching Volatility Regime. *Journal of Business & Economic Statistics*. https://www.tandfonline.com/doi/abs/10.1080/07350015.2024.2322090

Warsono, W., Russels, E., Wamiliana, W., Widiarti, W., & Usman, M. (2019). Vector Autoregressive with Exogenous Variable Model and its Application in Modeling and Forecasting Energy Data: Case Study of PTBA and HRUM Energy. *International Journal of Energy Economics and Policy*, *9*(2), Article 2.

Wiesen, T. F. P., & Beaumont, P. M. (2024). A joint impulse response function for vector autoregressive models. *Empirical Economics*, *66*(4), 1553–1585. https://doi.org/10.1007/s00181-023-02496-6

Yadav, S. K., & Akhter, Y. (2021). Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.645405

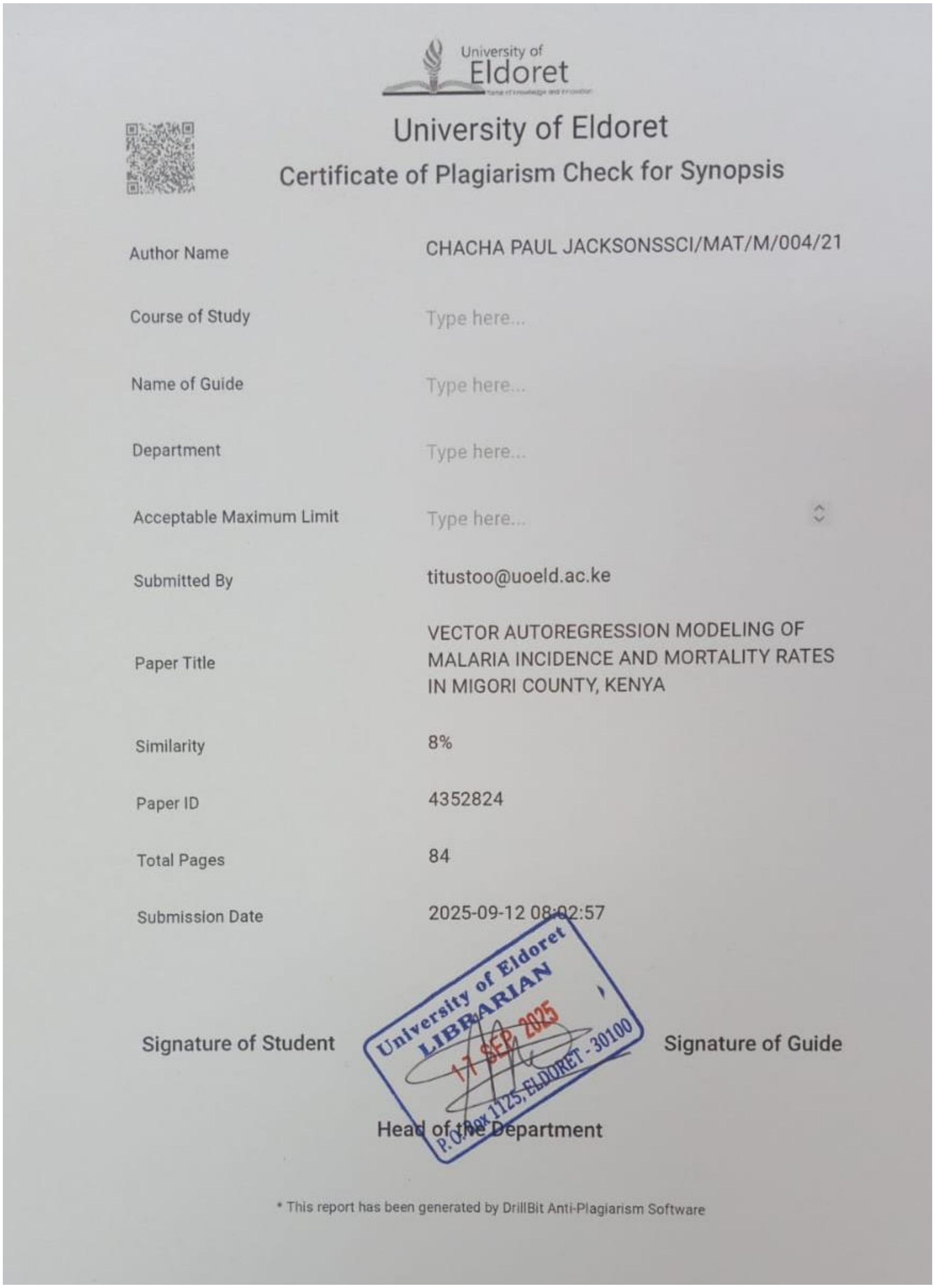
Zheng, Y. (2024). An Interpretable and Efficient Infinite-Order Vector Autoregressive Model for High-Dimensional Time Series. *Journal of the American Statistical Association*. https://www.tandfonline.com/doi/abs/10.1080/01621459.2024.2311365

Zhou, G., Githure, J., Lee, M.-C., Zhong, D., Wang, X., Atieli, H., Githeko, A. K., Kazura, J., & Yan, G. (2024). Malaria transmission heterogeneity in different eco-epidemiological areas of western Kenya: A region-wide observational and risk classification study for adaptive intervention planning. *Malaria Journal*, *23*(1), 74. https://doi.org/10.1186/s12936-024-04903-4

Zimmerman, R. H., Galardo, A. K. R., Lounibos, L. P., Galardo, C., Bahar, A. K., & van Santen, E. (2022). Vectorial capacities for malaria in eastern Amazonian Brazil depend on village, vector species, season, and parasite species. *Malaria Journal*, *21*(1), 237. https://doi.org/10.1186/s12936-022-04255-x

APPENDICES

Appendix I: Similarity Report



University of Eldoret

University of Eldoret

Certificate of Plagiarism Check for Synopsis

Author Name	CHACHA PAUL JACKSONSSCI/MAT/M/004/21
Course of Study	Type here...
Name of Guide	Type here...
Department	Type here...
Acceptable Maximum Limit	Type here...
Submitted By	titustoo@uoeld.ac.ke
Paper Title	VECTOR AUTOREGRESSION MODELING OF MALARIA INCIDENCE AND MORTALITY RATES IN MIGORI COUNTY, KENYA
Similarity	8%
Paper ID	4352824
Total Pages	84
Submission Date	2025-09-12 08:02:57

Signature of Student

Signature of Guide

Head of the Department

* This report has been generated by DrillBit Anti-Plagiarism Software