

**BINDURA UNIVERSITY OF SCIENCE EDUCATION
FACULTY OF SCIENCE AND ENGINEERING
DEPARTMENT OF STATISTICS AND MATHEMATICS**



**FORECASTING MALARIA CASES IN HARARE PROVINCE USING TIME SERIES
MODELS:**

**BY
RADWICK MUVHU
B202613B**

***A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS OF THE BACHELOR OF SCIENCE HONOURS DEGREE IN
STATISTICS AND FINANCIAL MATHEMATICS***

**SUPERVISOR: MS.J. PAGAN'A
JUNE 2024**

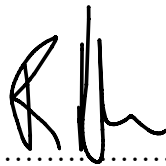
APPROVAL FORM

This is to certify that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgment. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.

I, Radwick Muvhu, the undersigned, hereby affirm that the work presented in this dissertation is my own original research, and I have duly acknowledged all sources of information and any assistance received. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University's regulations.

Student

RADWICK MUVHU



Signature.....Date...11/06/2024

Certified by

MS.J.HLUPO

Signature.....Date.....

Supervisor

MS.J.PAGAN'A



Signature

Date...11/06/2024

Chairperson

MR MAGODORA



Signature.....Date.....

DEDICATION

I dedicate this research project to my family members for their love, support, patience, encouragement and understanding. They gave me every reason to remain focused.

ACKNOWLEDGEMENTS

I am writing this acknowledgment with sincere gratitude, as I would never have been able to finish my thesis without the guidance of my supervisor, help from my friends, and support from my family. First and foremost, I wish to express my deepest thanks to my supervisor, Ms. J Pagana, for her consistent support in the development of my dissertation, her patience, motivation, and immense knowledge. Her guidance has been invaluable throughout the time of analysis and writing of this dissertation. I thank the Lord Jesus Christ, who has granted me the care, strength, knowledge, and the opportunity to pursue education up to this level. I would like to extend my gratitude to all the lecturers and staff of the Department of Mathematics and Statistics at Bindura University of Science Education for their support and academic knowledge imparted. Special attention and thanks go to my parents and siblings. Thank you all for your encouragement, financial support, advice, prayers, and patience throughout my studies. May the Lord Jesus Christ continue to bless you all abundantly.

Lastly, but not least, to all my colleagues: you are the best, and you are a family to me. I ask for God's guidance, mercies, and covering to be always with you.

ABSTRACT

Malaria remains a serious public health problem in Zimbabwe. Reliable forecasting of malaria is crucial for effective resource allocation, early warning systems, and implementing targeted intervention strategies. This research study aimed to explore the application of time-series models in forecasting malaria cases in Harare Province, Zimbabwe. The objectives were to establish the trend of malaria cases, predict the future trend using SARIMAX and deep learning LSTM models, and to determine the best-performing model in terms of mean absolute percentage error (MAPE). Historical weekly data on suspected and positive malaria cases from 2013 to 2023 were analyzed to develop reliable forecasting models. Two-time series models were developed, evaluated, and compared based on their predictive performance to identify the most appropriate model for accurate malaria case forecasting. The results showed an overall decreasing trend in malaria cases in Harare province. For suspected malaria case prediction, the SARIMAX (3,1,3)(0,0,0,13) model achieved a MAPE of 20%, while the LSTM model demonstrated exceptional performance with a MAPE of 0.099%. However, for positive malaria case forecasting, both the SARIMAX (6,0,6)(0,0,6,13) and LSTM models struggled, with the LSTM model having a MAPE of over 259 million percent. The findings suggest the LSTM model is superior for predicting suspected malaria cases, but faces challenges in accurately forecasting positive cases. The SARIMAX model provided reasonably accurate forecasts for both suspected and positive malaria cases. Further research is needed to improve positive case prediction capabilities, potentially by exploring alternative model architectures.

Table of Contents

| | |
|---|-------------------|
| APPROVAL FORM | <i>i</i> |
| DEDICATION | <i>ii</i> |
| ACKNOWLEDGEMENTS | <i>iii</i> |
| ABSTRACT | <i>iv</i> |
| List of Acronyms and Abbreviations | <i>ix</i> |
| CHAPTER 1: | <i>1</i> |
| 1.1 Introduction | <i>1</i> |
| 1.2 Background | <i>1</i> |
| 1.3 Problem Statement | <i>2</i> |
| 1.4 Aim of the study | <i>3</i> |
| 1.5 Research objectives | <i>3</i> |
| 1.6 Research questions | <i>3</i> |
| 1.7 Significance of the study | <i>3</i> |
| 1.8 Scope of the study | <i>4</i> |
| 1.9 Assumptions | <i>4</i> |
| 1.10 Limitations | <i>4</i> |
| 1.11 Definition of terms | <i>4</i> |
| 1.12 Conclusion | <i>5</i> |
| CHAPTER 2: LITERATURE REVIEW | <i>6</i> |
| 2.1 Introduction | <i>6</i> |
| 2.2 Theoretical literature review | <i>6</i> |
| 2.3 Time series analysis | <i>7</i> |
| 2.3.1 Components of time series | <i>7</i> |
| 2.3.2 Trend component..... | <i>7</i> |
| 2.3.3 Seasonality component..... | <i>7</i> |
| 2.3.4 Irregular component..... | <i>7</i> |
| 2.4 Seasonal Autoregressive Integrated Moving Average with exogenous variables | <i>7</i> |
| 2.5 Deep learning models | <i>13</i> |
| 2.6 Recurrent neural network | <i>13</i> |
| 2.7 Long Short-term Memory LSTM | <i>13</i> |
| 2.8 Empirical literature review | <i>14</i> |
| 2.9 Knowledge Gap | <i>17</i> |
| 2.10 Conceptual Framework | <i>17</i> |

| | |
|---|-----------|
| CHAPTER 3: RESEARCH METHODOLOGY..... | 18 |
| 3.1 Introduction..... | 18 |
| 3.2 Research Design..... | 19 |
| 3.3 Research instruments | 19 |
| 3.4 Data collection | 19 |
| 3.5 Target Population | 20 |
| 3.6 Description of variable..... | 20 |
| 3.7 Expected relation | 20 |
| 3.8 Data cleaning | 20 |
| 3.9 Diagnostics test..... | 21 |
| 3.9.1 Augmented Dickey Fuller Test..... | 21 |
| 3.9.2 Model diagnostics..... | 21 |
| 3.9.3 Ljung -Box test..... | 21 |
| 3.10 Models training..... | 21 |
| 3.11 Model validation..... | 22 |
| 3.12 Visual inspection..... | 22 |
| 3.13 Ethical Considerations..... | 22 |
| CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION | 23 |
| 4.1 Introduction..... | 23 |
| 4.2 Uploading the data into Jupyter note book | 24 |
| 4.3 Checking the dataset for completeness. | 24 |
| 4.4 Results of converting the week into date time index..... | 25 |
| 4.5 Descriptive Statistics..... | 26 |
| 4.6 Graphical Visualisation | 27 |
| 4.7 Stationarity test | 29 |
| 4.8 Decomposition graph for suspected malaria cases and positive | 31 |
| 4.8.1 Decomposed graph for suspected malaria cases | 32 |
| 4.9 ACF and PACF for weekly suspected | 33 |
| 4.9.1 ACF and PACF for weekly positive malaria cases..... | 33 |
| 4.10 Model identification for weekly suspected malaria cases..... | 34 |
| 4.10.1 summary of the best-fit model results..... | 34 |
| 4.11 Model identification for weekly positive malaria cases | 36 |
| 4.11.1 Summary statistics for the identified model | 36 |
| 4.12 Residual Analysis for weekly suspected malaria cases..... | 38 |
| 4.12.1 Residual analysis for weekly positive malaria cases | 39 |
| 4.13 Prediction for the SARIMAX on the test set | 40 |

| | |
|---|-----------|
| 4.14 Model validation..... | 40 |
| 4.15 Forecasting the Future Weekly Suspected Malaria Forecast | 42 |
| 4.16 Weekly positive malaria forecast | 43 |
| 4.17 Long Short Time Memory..... | 44 |
| 4.18 Model summary for Lstm for positive malaria cases..... | 46 |
| 4.18.1 Model summary for Lstm for positive malaria cases | 46 |
| 4.19 Validation metrics | 47 |
| 4.20 Forecast | 48 |
| 4.20.1 Forecasted values for weekly suspected malaria cases | 48 |
| 4.21 Forecast for weekly positive malaria cases | 49 |
| 4.22 Summary..... | 50 |
| CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS | 50 |
| 5.1 Introduction..... | 50 |
| 5.2 Summary of Research Findings | 51 |
| 5.3 Recommendations | 51 |
| 5.4 Conclusion | 52 |
| References | 52 |
| APPENDIX 1 | 55 |
| APPENDIX 2 | 78 |

TABLE OF FIGURE

| | |
|---|----|
| Figure 1 Conceptual framework for LSTM and SARIMAX | 18 |
| Figure 2 importing malaria data file into jupyter..... | 24 |
| Figure 3 Results for checking data completeness | 25 |
| Figure 4 Changing the week column into datetime index | 25 |
| Figure 5 Descriptive statistics for weekly suspected malaria cases..... | 26 |
| Figure 6 Descriptive statistics for weekly positive malaria cases | 27 |
| Figure 7 Graph for weekly suspected malaria cases | 28 |
| Figure 8 Trend for weekly positive malaria cases | 28 |
| Figure 9 ADF Test statics for weekly suspected malaria cases | 29 |
| Figure 10 ADF Test statistics for weekly positive malaria cases | 29 |
| Figure 11 ADF Test statistics for differenced weekly positive malaria cases | 29 |
| Figure 12 ADF Test statistics for exogenous variables | 30 |
| Figure 13 Decomposed plot for weekly suspected malaria cases..... | 31 |
| Figure 14 Decomposed plot for weekly positive malaria cases | 32 |
| Figure 15 ACF and PACF plot for weekly suspected malaria cases | 33 |
| Figure 16 ACF and PACF plot for weekly positive malaria cases | 33 |
| Figure 17 Model identification parameters with the lowest AIC | 34 |
| Figure 18 Summary Statistics for SARIMAX(3,1,3)(0,0,0,13) | 35 |
| Figure 19 Model identification parametres for weekly positive malaria cases | 36 |
| Figure 20 Summary statistics for SARIMAX(6,0,6)(0,0,6,13) | 37 |
| Figure 21 Model fit plot diagnostics for weekly suspected malaria cases | 38 |
| Figure 22 Model fit plot diagnostics for positive malaria cases | 39 |
| Figure 23 SARIMAX(3,1,3)(0,0,0,13) test set predictions | 40 |
| Figure 24 SARIMAX(2,0,3)(3,0,2,13) test set predictions | 40 |
| Figure 25 Forecasted results for weekly suspected malaria cases | 42 |
| Figure 26 Graph forecast for weekly suspected malaria cases | 42 |
| Figure 27 Forecasted results for weekly positive malaria cases | 43 |
| Figure 28 Graph for weekly positive malaria forecast..... | 43 |
| Figure 29 Normalised before being used for Lstm development | 44 |
| Figure 30 Training and Validation loss Graph for weekly suspected malaria cases | 45 |
| Figure 31 Training and validation loss graph for weekly positive malaria cases..... | 45 |
| Figure 32 Model summary for LSTM weekly positive malaria cases | 46 |
| Figure 33 Model summary for weekly positive malaria cases | 46 |
| Figure 34 Forecast for suspected malaria cases for 30 weeks | 48 |
| Figure 35 Graph for weekly suspected malaria forecast lstm | 48 |
| Figure 36 weekly positive malaria cases forecast..... | 49 |
| Figure 37 Graph for weekly positive malaria forecast..... | 49 |
| Figure 38 City of Harare approval form..... | 78 |
| | |
| Table 2 Description of variable | 20 |
| Table 3 Model validation for SARIMAX models | 40 |
| Table 4 Model validation for Long short-term memory results | 47 |

List of Acronyms and Abbreviations

| | |
|---------|--|
| DHIS | District Health Information |
| LSTM | Long Short Term Memory |
| SARIMA | Seasonal Autoregressive Integrated Moving Average |
| SARIMAX | Seasonal Autoregressive Integrated Moving Average with exogenous variables |
| ARIMA | Autoregressive Integrated Moving Average |
| ARMA | Autoregressive Moving Average |
| ARIMAX | Autoregressive Integrated Moving Average with exogenous variables |
| PACF | Partial Auto Correlation Function |
| ACF | Auto Correlation Function |
| WHO | World Health Organisation |

CHAPTER 1:

1.1 Introduction

Over one-third of the world's population is at risk of contracting malaria, which continues to be one of the leading causes of death for humans globally. The effect of this disease has sparked widespread debate and concern globally. In particular, malaria poses a significant threat in urban, peri-urban, and rural areas of developing countries, such as Zimbabwe (Tatem, 2013). Addressing and understanding the patterns and dynamics of malaria cases is crucial for effective disease management and prevention. Forecasting diseases, including malaria, is a valuable tool utilized by scientists in the field of public health. By employing techniques like time series analysis and deep learning models, researchers can understand the weekly patterns of malaria cases. This enables a deeper understanding of the disease's dynamics, facilitating more targeted interventions and resource allocation. In light of the importance of disease forecasting, this chapter aims to present a study focused on forecasting the weekly pattern of malaria cases. The study employs time series analysis and deep learning models to analyze historical data and predict future trends. By doing so, it aims to contribute to the body of knowledge surrounding malaria prevention and control strategies.

The chapter is structured as follows: first, the background of the study will provide an overview of the current understanding of malaria and the significance of forecasting its patterns. Next, the problem statement will outline the specific challenges and gaps in existing research. The aim of the study will be clearly defined, followed by research questions that guides the investigation. The significance of the study will be highlighted to emphasize the potential practical implications and benefits of the research findings. Additionally, the chapter will address the hypothesis and limitations of the study, and conclude by summarizing the key points and laying out the roadmap for the subsequent sections.

1.2 Background

Malaria, a curable yet life-threatening disease, has a long history dating back to ancient times. Early references to malaria can be found in Egyptian papyrus writings, and the renowned Greek physician Hippocrates provided detailed descriptions of the disease. Malaria has had a significant impact throughout history, even devastating invaders of the Roman Empire. The name "malaria"

itself originate from the Italian phrase "mal aria," meaning bad air, reflecting the belief that the disease was caused by noxious air in marshy areas.

Malaria is caused by plasmodium parasites and mainly spread to humans via the mosquito bites of female mosquitoes. For people who lack immunity, symptoms usually manifest in less than ten days. If untreated, it can progress to severe illness and have life-threatening consequences. The World Health Organization estimated that in 2019, recorded two hundred and twenty nine million malaria cases and forty hundred and nine thousand deaths ((WHO), 2020)).

In Zimbabwe, as in many other parts of Southern Africa, malaria is a major public health issue. The country is at high risk because of the presence of Plasmodium parasites and the Anopheles mosquito vector. While malaria is preventable and treatable, the disease continues to pose a substantial burden on the population. It is responsible for a significant number of deaths, especially among children under five years old. In 2015, malaria transmission was ongoing in Zimbabwe, highlighting the persistent threat the disease poses (WHO 2015).

The transmission of malaria is influenced by several factors, including present of parasite, the mosquito vector, and the environment. The female mosquitoes lay their eggs swampy area, and as they develop, they go through larval stages before emerging as adult mosquitoes. When a person is bitten by an infected mosquito, the parasites enter the blood circulation system and infect the blood cells and destroy. The leading characteristic symptoms of malaria include fever, headache fatigue, nausea and diarrhoea. If left untreated, malaria can result in severe complications and death. Globally, malaria remains a major concern, with increasing impact and devastating effects on human populations, particularly in developing countries like Zimbabwe.

1.3 Problem Statement

The escalating malaria cases transmission in Zimbabwe presents a significant public health challenge that affects both rural and urban areas, as well as peri-urban communities. Despite concerted efforts to combat the disease, the country continues to struggle with the burden of malaria cases and associated morbidity and mortality. The World Health Organization (WHO) African Region, including Zimbabwe, bears heaviest burden, accounting for over 95% of global malaria cases in 2021. The rising number of malaria cases poses severe implications for the health system, exacerbating resource constraints and straining healthcare services. The impact extends beyond the individual level, impeding socio-economic development and hindering progress toward achieving national and international health targets.

1.4 Aim of the study

The study aims to forecast the dynamics of Malaria cases in Harare Province using time series models.

1.5 Research objectives

The goal of the research is:

- ✓ To establish trend the of Malaria cases in Harare Province.
- ✓ To predict the future trend of Malaria cases using time series models.
- ✓ To determine the best-performing model between SARIMAX and Deep Learning models in terms of mean absolute percentage error.

1.6 Research questions

The study seeks to answer the following questions:

1. Is the trend in the Malaria cases increasing or decreasing?
2. How accurately can time series models forecast malaria cases?
3. Which models between SARIMAX and Deep Learning perform best in forecasting malaria cases?

1.7 Significance of the study

This study holds significant importance for various stakeholders, including the Government of Zimbabwe, Bindura University of Science Education, and the researcher involved. The significance of the study is outlined as follows:

Government of Zimbabwe:

The research findings from this study can be used as a valuable tool by the Government of Zimbabwe in forecasting the dynamic nature of malaria. By understanding the patterns and trends of malaria cases, the government can make informed decisions on resource allocation, develop effective prevention and control strategies, and improve management in hospitals and healthcare facilities. The study's insights can contribute to maximizing the utilization of available resources and improve decision-making steps related to the malaria control and prevention in Harare.

Bindura University of Science Education:

For university, the study adds to existing knowledge and becomes a valuable addition to the library materials. It serves as a reference for future researchers who wish to explore similar areas of study. The findings and methodologies employed in this research can provide a foundation for

further studies and contribute to the academic and scientific discourse surrounding malaria forecasting and control. The study's outcomes can potentially inspire and guide future research endeavour within the university and beyond.

Researcher:

Participating in this study provides the researcher with valuable opportunities for personal and professional growth. By working on forecasting malaria cases using programming languages and employing time series and deep learning models, the researcher's understanding of these techniques will be broadened. The researcher will gain practical experience in data analysis, modelling, and forecasting, enhancing their skills in these areas. The knowledge and expertise acquired through this study will serve as a valuable asset for future research endeavours, enabling the researcher to contribute further to the field of malaria prevention and control.

Literature:

The outcome of research will add to the of knowledge application and effectiveness of time series models in malaria forecasting

1.8 Scope of the study

The research focuses on using statistical and mathematical modelling in machine learning to forecast Malaria cases. The research used historical from City of Harare Health Information department (DHIS) Zimbabwe City Health Department and Harare Meteorological data sector of rainfall and temperature data, as these are conditions which influence malaria cases. The data is imported into the Python package and analyzed.

1.9 Assumptions

- ✓ The data utilized in this study comes from reputable sources and has not been tampered with.
- ✓ There are no missing data variables, and the weeks are defined as a time variable

1.10 Limitations

- ✓ The study focused on Harare Province exclusively.
- ✓ The study analyzed historical data from January 1, 2013 to December 31, 2023. The research did not account for each district's weekly temperature and rainfall
- ✓ The study used average weekly average temperature and weekly rainfall

1.11 Definition of terms

DHIS - District health information

WHO - World Health Organisation

Time series - is a collection of data points gathered consistently over a period, usually at equally spaced intervals of time. It represents the evolution of a variable or phenomenon over time, allowing for the analysis of patterns, trends, and dependencies in the data (Brockwell, 2016).

1.12 Conclusion

Chapter 1 sets the stage for the research study, establishing its relevance, significance, and scope.

It provides a solid foundation for subsequent chapters, laying the groundwork for the detailed analysis, methodology, and findings that follow

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This section focuses on number of key findings that are related to the study's available literature on the applicability of time series models in forecasting malaria cases, which will serve as a review of the concepts from earlier studies and associated works of literature to briefly. A research case framework is presented and covers the main focus of the dissertation. Many different models have been used to try and forecast malaria cases.

2.2 Theoretical literature review

Despite the Government of Zimbabwe's efforts to eliminate malaria and establish Harare as a non-malaria transmission zone, malaria transmission continues to persist in the country. Time series analysis has significantly impacted the health sector, particularly when forecasting various diseases such as typhoid (Nwakuwa Esther Promise, 2022), tuberculosis (Hellyon, 2022), and the recent coronavirus pandemic (Amasiri et al., 2022). Accurate forecasting of disease cases is crucial for effective public health planning and intervention strategies. The purpose of this study is to implement time series analysis to forecast malaria cases in Harare Province, where malaria is endemic. The study employs two models: SARIMAX and LSTM model, a traditional time-series model and a neural network-based model, respectively. SARIMAX models have been mostly used in time series forecasting, including prediction of infectious disease prevalence, as they can capture temporal dependencies and incorporate external factors such as climate variables. LSTM models, on the other hand, have gained influence in time series forecasting and disease prediction. By leveraging patterns and trends from historical malaria data, LSTM models can provide accurate predictions, particularly when the time series exhibits nonlinearity and long-term dependencies. Incorporating epidemiological information into the modelling process can enhance the accuracy and robustness of malaria case forecasts. Factors such as mosquito vectors, human behaviour, and environmental conditions are considered in the theory of malaria transmission and play significant roles in the dynamics and spread of the disease.

Theoretical Framework

2.3 Time series analysis

(Trirat, 2024) time series is an ordered collection of data points; the data are usually equally spaced. The data could be recorded in hourly, minute, and weekly intervals. Analysis of these data sets by data mining, pattern recognition, and predictive analysis is called time series analysis ((Mills, 2019)). It seeks to understand the underlying context of the relevant data points through the derivation of future values from past recorded values

2.3.1 Components of time series

Three key components of timeseries are: trend, seasonality, and irregularity

2.3.2 Trend component

A trend is gradual changes throughout time. The trend may be moving up or down, when it moves up, we have a positive trend and when it moves down, we have a negative trend

2.3.3 Seasonality component

(Zuo, 2022))The seasonal captures the seasonality variation, a repetition that keeps happening for a repeatedly set of time

2.3.4 Irregular component

Commonly referred to as the residual component. This is what is left after estimating the trend and seasonal component from the series. The residual component is the part that cannot be explained by the trend or seasonality component. It represents the part of the data that cannot be modelled.

2.4 Seasonal Autoregressive Integrated Moving Average with exogenous variables

(Papaioannou GP, 2016) stated that the autoregressive moving average (ARMA) model transformed into the autoregressive integrated moving average model. The ARMA is a fusion of the moving average MA(q) and the autoregressive AR(p). The order in which the auto regressive and moving average models are applied is determined by the ARMA model. These models are most suitable when the data exhibits normality properties. When the Auto-Correlation Function (ACF) steadily declines and the series does not exhibit a consistent pattern, the ARMA model is employed for modelling. However, if the time series data contains anomalies, the data is often transformed before being modelled by ARMA, which is then referred to auto regressive integrated moving average.

The autoregressive AR (p), also known as the average regression model, assumes that present time series information can be adequately described through its historical information. When there is a minor correlation between the present and past information, the current data changes to a white noise time series. The degree of reliance on the past determines strength of the

relationship, with a stronger dependency resulting in a random walk. To understand the properties of the target time series information, the AR(p) model examines its autocorrelation with the past information.

This means information from previous time t influences current results. In the autocorrelation function and partial auto-correlation function graphs show the autocorrelation function declines quickly, while the partial auto correlation function exhibit a cutoff limit. The auto correlation function calculates the correlation between data points that are separated by a certain number of periods, indicating order of correlation based on time differences. The PACF, on the other hand, represents the correlation coefficient between two variables after controlling for any other intermediate values. A typical regression model for AR(p) shows

$$Y_t = C + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t$$

In this equation:

When forecasting malaria cases Y_t shows the number of malaria cases at a certain period, t , which can be daily weekly or monthly cases, $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$ are the lagged values of the malaria cases where p is the order of the AR model. lagged values capture impact of previous malaria cases, $\phi_1, \phi_2, \dots, \phi_p$ are the auto regressive coefficients that determine the influence of the lagged values on the current malaria cases they quantify how much the previous malaria cases contribute to the current cases, ε_t represents is error or residual terms at time t which capture for factors other than the lagged values. In context of malaria forecasting this could be factors such as change in climate patterns

The ARMA form simplifies model interpretation by centering lagged values around the mean (μ). It helps to observe how each lagged value affects the current value relative to the time series mean. The present value on the time series is a weighted average of past residuals, while the moving average reflects a moving average process. Given that the residual term is white noise, the current value is calculated as the mean of the previous white noise. The moving average model, which is based on the summing of these terms, presents an average regression feature because of the high normality and average regression properties of white noise. Unlike the autoregressive model AR(p), the moving average model MA(q) uses a weighted linear combination with white noise ε_t . The general form of the MA(q) model is given as:

$$Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_p\varepsilon_{t-p}$$

Where:

Y_t is the time series being modelled

μ is mean or intercept of the time series.

ε_t indicates the error term or residual at time t , which accounts for unexplained fluctuation in the time series.

$\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-p}$ are the lagged prediction errors

$\theta_1; \theta_2; \theta_p$ represents moving average coefficients that determine the impact of previous prediction errors on the current time series

Estimating time series data solely with auto regressive and moving average models challenging. To address this, the autoregressive moving average model joins the strengths of two models. It assumes that previous time series data and error term determine the present data, exhibiting an average regression characteristic. The ARMA model is suitable for time series analysis because of its normality and efficiency in approximation with fewer parameters.

The equation for ARMA

$$Y_t = c + \varphi_1Y_{t-1} + \varphi_2Y_{t-2} + \dots + \varphi_pY_{t-p} + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_p\varepsilon_{t-p} + \varepsilon_t$$

In this equation:

Y_t is the time series being modelled.

$Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$ represents the lagged values in the time series, p representing the order of the autoregressive model.

c represents the intercept.

$\varphi_1, \varphi_2, \dots, \varphi_p$ represents the autoregressive coefficients that determine impact of previous values on the present value of the time series.

$\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-p}$ represent prediction errors or residuals at time t and the lagged prediction errors, where q is order of moving average model.

$\theta_1, \theta_2, \dots, \theta_p$ represent moving average coefficients that determine impact of previous prediction errors on the present value of time series

For non-normal or unstable time series data with rising trends or variance, normalization techniques such as log transformation, differencing, and seasonal differencing are required before applying the ARIMA model for analysis.

Several strategies can be used to analyze these seasonal time series models. These include regression models with indicator and trigonometric functions, as well as Winters' seasonal exponential smoothing. However, these approaches presume that the seasonal time series data are independent, which may not be the case as time series data frequently display correlation. In such cases the most appropriate is to use the ARIMA model. Even though the original data lacked normality or average regression qualities, the differenced data could have these properties. The ARIMA model utilizes different time series and is essentially an ARMA model. The auto regressive moving average and auto integrated moving average models with a difference of 0 are equal.

ARIMA model is consist of three orders: p, d, and q. ARIMA(p,d,q), p), p are autoregressive terms number, d is the non-seasonal differences to make the series stationarity, and q represent the number of lagged forecasted errors in the prediction equation.

The equation for an ARIMA (p, d, q)

$$(1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - B)^d Y_t = c + (1 + \theta_1 B + \dots + \theta_p B^p) \varepsilon_t$$

When forecasting malaria cases, Y_t represents the the number of malaria cases over time. $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$ are the time series lagged values. These represent the past values of malaria cases, up to a certain lag order p, by utilizing these lag values as predictors, we are able to capture relationship between past and current malaria cases. c represent the intercept or constant. It represents an average level of malaria cases when other variables are equal zero. $\varphi_1, \varphi_2, \dots, \varphi_p$ are the autoregressive values. These coefficients determine impact of previous lagged value on the present value of malaria cases. They represent the direction and strength of the connection between past cases with present cases. $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-p}$ are the prediction errors or residuals at time t and the lagged prediction error. They represent the difference between actual and predicted values of malaria cases. By including these residuals as predictors, they can capture any remaining patterns or information in the data. $\theta_1, \theta_2, \dots, \theta_p$ are moving average coefficients. The coefficients determine the effect of past prediction errors on the present value of malaria cases. They represent the influence of past error in adjusting the forecasted values.

When time series data exhibit seasonality trends, the seasonal autoregressive integrated moving average (SARIMA) model is commonly used (Dabral, 2017). SARIMA combines the seasonal, autoregressive (AR), integrated (I), and moving average (MA) components. It assumes that data contain trends, seasonal components, and irregular terms. Developing SARIMA models typically involves several steps.

First, (Hassani, 2017) OCSB test is used to find the order of the seasonal differencing. Osborn, Chui, Smith, and Birchenhall (OCSB) test statistical determine order of seasonal differencing in a time series. It identifies appropriate number of seasonal differences required to make the series stationary. The OCSB test examines the autocorrelation structure of time series at different seasonal lags to determine the appropriate order of seasonal differencing. After seasonal order differencing is determined, the KPSS unit-root test is employed to establish order of non-seasonal differencing. The KPSS test checks whether time series unit root or is stationary. If time series found to have unit root, it implies that differencing is required to make it stationary. After determining the orders of seasonal and non-seasonal differencing, the model space is explored using stepwise processes. This involves fitting different models with various combinations of autoregressive (AR), moving average (MA), seasonal autoregressive (SAR), and seasonal moving average (SMA) terms. Stepwise process evaluates the performance of each model based on goodness-of-fit tests such as Akaike Information Criterion and examination of estimated residuals

The formula for SARIMA (p,d,q)(P, D, Q) where (P, D, Q)s represents additional seasonal. The specific values of P, D, and Q depend on order seasonality of data.

The equation for a SARIMA (p, d, q)(P, D, Q)m model is represented as follows:

$$\begin{aligned} (1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - B)^d (1 - B^m)^D (Y_t - \mu) \\ = (1 + \theta_1 B + \dots + \theta_p B^p)(1 + \theta_1 B^m + \dots + \theta_p B^m) \varepsilon_t \end{aligned}$$

When forecasting malaria cases Y_t represents the number of malaria cases at time t. It refers to the observed values of malaria cases over time. B represents the backshift operator. When B is applied to the malaria cases time series, it shifts the values backward by one time period. For example, $B^m Y_t$ represents the number of malaria cases at time t-m. μ represents the average number of malaria cases. It is a constant value that represents the baseline level of malaria cases. $\varphi_1, \varphi_2, \dots, \varphi_p$ are the autoregressive (AR) coefficients. They represent the effect of previous p

lagged values of the malaria cases on the present values. The AR(p) captures relationship between the current cases and past cases. d represents order of differencing applied to the malaria cases time series. Differencing is used to remove trends and seasonality from the data, making it stationary. $\theta_1, \theta_2, \dots, \theta_p$ are the moving average (MA) coefficients. They represent effect of previous q lagged error terms on the present value. The MA(q) component captures the effect of the previous errors on the present number of cases. ε_t represents error term at time t. It captures the part of the observed number of cases that is not explained by the model. P represents order of seasonal autoregressive (SAR) terms. These terms capture the seasonal patterns in malaria cases. The SAR(P) component represents the effect of the previous P lagged seasonal values on the present value. D represents order of seasonal differencing applied malaria cases time series. It is similar to the non-seasonal differencing (d), but it is applied to a lagged seasonal value. $\theta_1, \theta_2, \dots, \theta_p$ these are the seasonal moving average (SMA) coefficients. They represent the effect of the previous Q lagged seasonal error terms on the current value. m represents the seasonal period. In the case of malaria forecasting, it represents the length of the seasonal pattern in the data, such as the annual or seasonal fluctuations in malaria cases.

If the order of the seasonal time series model is zero, it is equivalent to the ARIMA model.

The SARIMAX is an extension model for SARIMA model, it incorporating the influence of exogenous factors. Exogenous factors are external factors that can impact the time series being analyzed. By including these variables in the modelling process, the SARIMAX model can enhance the accuracy and predictive power of future value estimations. These variables include economic indicators, weather data, demographic information, or any other relevant factors that are believed to impact the time series under investigation (Marco Peixeiro 2022).

The SARIMAX model is a more comprehensive version of the SARIMA model that includes exogenous variables. The equation for a SARIMAX (p, d, q)(P, D, Q)m model can be expressed as follows:

$$(1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - \Phi_1 B^m - \dots - \Phi_p B^m) (1 - B)^d (1 - B^m)^D (Y_t - \mu - \beta_1 X_{1t} - \dots - \beta_k X_{kt}) = (1 + \theta_1 B + \dots + \theta_p B^p)(1 + \theta_1 B^m + \dots + \theta_p B^m)\varepsilon_t$$

In this equation

When forecasting malaria cases, the SARIMAX model adds $\beta_1, \beta_2, \dots, \beta_k$ which represents the coefficients associated with the exogenous variables. These coefficients capture the relationship between exogenous variables and outcome of malaria cases

2.5 Deep learning models

Deep learning, is part of machine learning models, which is specifically designed for neural network models which is highly effective for time series analysis, particularly when dealing with large datasets that possess numerous features and exhibit non-linear relationships. Deep learning offers various architectures, such as LSTM, CNN, and auto-regressive deep neural network models. Three types of deep learning models that are suitable for times series forecasting are single model, multi-step model, and multi-output model, (Peixeiro, 2022).

2.6 Recurrent neural network

A recurrent neural network (RNN) is a specific type of deep learning structure tailored for processing sequential information. RNNs form a family of networks that share a common architecture. Within the RNN family, there are different model such as long short-term memory and gated recurrent unit (GRU). RNNs incorporate a hidden state that is fed back into the network, enabling it to use past information as input when processing the next element in a sequence. This mechanism allows the network to mimic the concept of memory. However, RNNs are subject to a limitation known as short-term memory, which means that the influence of past information decreases as the sequence progresses (Goodfellow, I., Bengio, Y., & Courville, A., 2016) This is a key aspect of RNNs that is important to consider when using them for modeling sequential data.

2.7 Long Short-term Memory LSTM

The Long Short-Term Memory (LSTM) architecture is an enhancement of the Recurrent Neural Network (RNN) architecture, specifically designed to tackle the diminishing gradient problem. This problem arises when the influence of past information on the network fades over time. In LSTM, a cell state is incorporated into the RNN to preserve crucial information for extended periods. The LSTM architecture is composed of 3 gates: forget gate, input gate, output gate. Forget gate decides what information from the past and current sequence values should be kept or discarded. Input gate identifies pertinent information from the current sequence and updates the cell state accordingly. Output gate uses the past information stored in the cell state to process

the current sequence element. It can either produce a result to the output layer or generate new information for the next sequence element. LSTM can be implemented in several ways, as single-step model, multi-step model, and multi-output model. In the single-step model, a single value representing the prediction is outputted. The multi-step model, the output is a sequence of predictions for multiple future time steps. The multi-output model produces predictions for more than one target. (Peixeiro, 2022).

2.8 Empirical literature review

Numerous studies have been conducted to analyze and compare time series models for forecasting future estimates of time series data. Many of these studies aim to determine the best-performing model among the thousands available. This section reviews a few studies that focus on LSTM and SARIMAX models and other related models.

(Gondwe, 2021) utilized the SARIMA model in the Nsanje district hospital of Malawi to examine trends in malaria cases among children aged 5 years and under. The study found that the SARIMA model $(0, 1, 2)(0, 1, 1)_{12}$ was the most suitable for forecasting malaria incidence in the Nsanje district. The research concluded that malaria cases were increasing at a unknown rate and predicted values from the model closely matched the actual values, demonstrating its adequacy for monthly malaria case forecasts.

(Wang M, 2019) compared the predictive performance of several models, including ARIMA, STL + ARIMA, BP ANN, and LSTM network models, in Yunnan province, China. The study employed auto ARIMA to build a SARIMA $(0, 1, 2)(0, 1, 1)_{12}$ model. The BP ANN model was established with optimal parameters NNARR $(12, 15)_{12}$, and the LSTM model had a learning rate of 0.001, MSE as the loss function, and 2 hidden layers. The results indicated that the LSTM model outperformed the other three models in terms of prediction performance for malaria cases.

Anwar and Malar (2016) used an ARIMA model to predict future trends in malaria incidence in Afghanistan. They developed two ARIMA models, ARIMA $(4, 1, 1)(1, 0, 1)_{12}$ and ARIMA $(1, 1, 1)(1, 0, 1)_{12}$. ARIMA $(4, 1, 1)(1, 0, 1)_{12}$, was identified as the best fit model and was used in estimating the number of malaria cases in a given month based on cases occurring in the 1st, 2nd, 3rd, 4th, and 12th months prior, after adjusting for negative seasonal moving averages. While the model performed well for short-term one-step-ahead predictions, its long-term prediction performance was not satisfactory. However, the model ARIMA $(1, 1, 1)(1, 0, 1)_{12}$ demonstrated better long-term predictive power, with estimates remaining close to the

actual data although the model did not have a good fit to the actual data. The study suggested that ARIMA can be applied to forecast malaria patterns in Afghanistan.

(Thomas Schincariol, 2021) conducted a study to evaluate the feasibility of predicting malaria cases in the French Guiana-Brazil cross-border area and developing an early warning system. The study compared LSTM and ARIMAX model. The LSTM model yielded good prediction results, with a slight increase of only 1 per cent in RMSE for Brazilian data and a 12 per cent lower MAE. The LSTM model predicted a lower number of cases compared to the ARIMAX model, which had higher predicted cases. The study concluded that LSTM model outperformed ARIMAX model in terms of predictive errors, temporality prediction of malaria peaks, and prediction of low cases. However, the study covered a relatively short period from 2014 to 2019, limiting its ability to conclude long-term prediction performance.

(Sakubu, 2023) employed deep learning models to study dynamics of malaria in Burundi. The research developed two types of deep learning model univariate LSTM and multivariate LSTM. After running and fitting the data, the models were tested using RMSE. The univariate LSTM model achieved a smaller RMSE than the multivariate LSTM model. The estimated malaria cases in the country during the study period were 12,959,182.46, with the univariate model predicting 12,841,653.9 (approximately one million cases less than the observed cases) and the multivariate LSTM model predicting 15,215,766.15 cases (around two and a half million more than the actual reported cases).

(Zinszer, 2014) and his team used ARIMAX to predict future cases of malaria in Bhutan. They looked at monthly malaria cases reported by health centers from 1994 to 2008, as well as weather data like temperature and rainfall. They made different ARIMA models for each district in the province. The best models for each district were different, but the best overall model was ARIMA (2, 1, 1)(0, 1, 1)₁₂. This model predicted that there would be between 15 to 82 cases in 2009 and 67 to 149 cases in 2010. The population in 2009 was 285,375, and they expected it to be 289,085 in 2010. The ARIMAX model, which included monthly cases and weather factors, showed different results for different districts. A one-month delay in the highest temperature was a strong sign of more malaria cases in four districts. The number of malaria cases in previous months was also a good predictor in one district, but no factor could predict malaria cases in two districts. The ARIMA models were useful in predicting the number of cases in areas where

malaria is common in Bhutan. However, the factors that predicted malaria cases were not the same when using the ARIMAX model with selected lag times and weather predictors.

(Mohamed, 2022)The results showed that the ANN model performed better than the other models, with the lowest values for RMSE (39.4044), MAE (29.1615), MAPE (31.3611), and MASE (0.6618). Furthermore, the researcher explored the inclusion of three meteorological variables, such as humidity, in the ANN model. The findings indicated that incorporating these climatological variables enhanced the model's predictive ability for malaria incidence data. It is important to note that this study focused solely on the Marodijeh region of Somaliland only.

Multiple research efforts in disease burden prediction have also shown that ANN models outperform classic approaches such as SARIMA. However, It remains difficult to select a single method for predicting malaria, as no single approach has proved consistent superiority over others.

(Adeola, (2019), predicted malaria cases in Nkomazi, South Africa using environmental variables sensed remotely. These variables, including vegetation indices, water index, land surface temperature, and rainfall, were evaluated monthly using SARIMA models. Predictions were made for 56 months of the remain information and compared with actual malaria cases. All environmental variables, except for land surface temperature, were significant in predicting malaria transmission. Rainfall showed the highest correlation with malaria cases. The SARIMA model without environmental variables could explain 41% of the variation in malaria cases. However, when these environmental variables were included, the model explained about 65% of the variation. The study concluded that the predicted number of malaria cases closely matched the observed cases, suggesting the model's effectiveness in predicting malaria cases in the study area.

(Nwakuwa Esther Promise, 2022) developed a Hybrid SARIMA-LSTM Model was used to predict malaria cases in Nigeria. The study used two time series statistical methods, SARIMA and LSTM, to analyze and forecast malaria prevalence. Data from the World Bank, covering 2003 to 2019, was used. The study found that malaria cases in Nigeria are low during the dry season and increase during the rainy season. The SARIMA-LSTM model was more accurate in forecasting than the standalone SARIMA and LSTM models. The SARIMA-LSTM model outperformed the SARIMA model because it could capture both linear and nonlinear

characteristics in the data. The study concluded that the Hybrid SARIMA-LSTM Model was effective in predicting malaria prevalence in Nigeria.

2.9 Knowledge Gap

These studies provide valuable empirical results on the analysis of predictive models for malaria cases. However, there is gap in literature regarding predictive models for malaria cases in Harare province. This gap calls for further research to study both SARIMAX and LSTM models in forecasting malaria cases in Harare. By employing multiple predictive models, the accuracy of predictions can potentially be enhanced. It is assumed that increasing the number of forecasting methods used will improve the accuracy of the study's predictions

2.10 Conceptual Framework

(Brown, 2016) described conceptual model as a high-level representation or framework that describes the key components, relationships, and processes of a system or phenomenon. They elaborate that it provides a conceptual understanding of how different elements interact and work together to achieve a specific goal or outcome and conceptual models help to clarify and communicate complex ideas, theories, or systems in a simplified and visual manner. A diagram is used in this study give stages of occurrence of the study

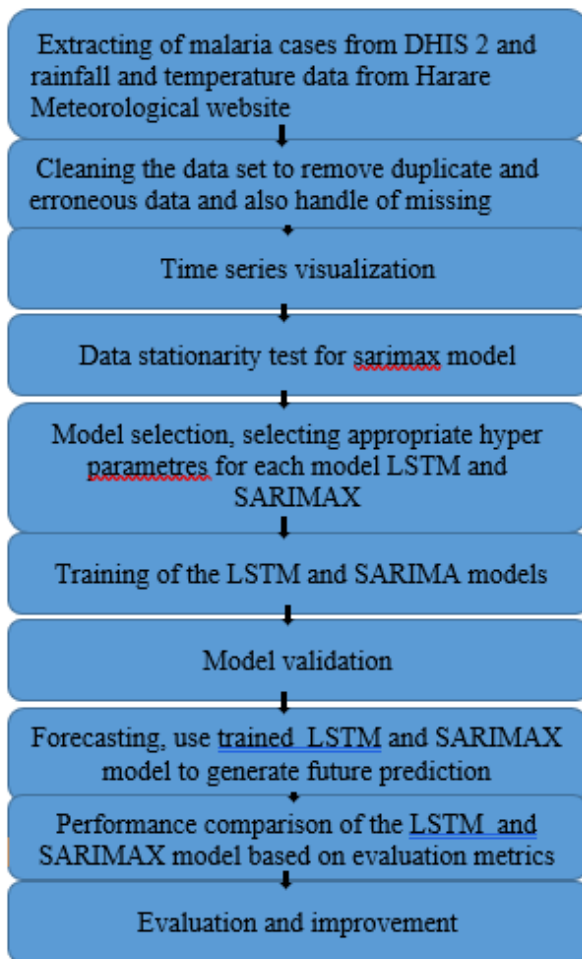


Figure 1 Conceptual framework for LSTM and SARIMAX

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

A research methodology is a fundamental component in any research as it delineates systematic steps and procedures employed by researchers to conduct their investigation. It serves as a roadmap that keeps researchers focused, making sure that the study is conducted efficiently, effectively, and manageable manner. A well-crafted research methodology is crucial in ensuring that the study produces valid and reliable results that can effectively address the research

question (Ali, 2023). This section of the research study examines and discusses key elements of research methodology of forecasting malaria cases using timeseries models

3.2 Research Design

The research design pertains to the blueprint outlining how data was collected, measured, and analyzed. It encompasses selection of a comprehensive strategy that harmoniously integrates all the research components, ensuring a cohesive and logical approach to effectively address the research problem (Creswell, 2014). In this research project, a predictive research design was utilized. A predictive research design is a methodology employed to create models or algorithms that has the ability to predict or forecast using past data (Hair et al., 2017). This design allows researchers to leverage past data patterns to generate insights and projections about future outcomes or trends. By developing predictive models or algorithms, researchers can have information into the phenomena under investigation and potentially make informed decisions or recommendations based on the forecasted results.

3.3 Research instruments

The researcher analyzed epidemiological reports and statistical data from the DHIS system and also analyzed weekly rainfall and temperature records from Kutsaga Tobacco Research Institute website. To gather relevant data on malaria cases. The researcher also conducted interviews with a health department staff member to have information about the current development of malaria programme with current measures taken by the city health department to combat malaria in Harare.

3.4 Data collection

The research utilised secondary data from the City Health Department's online database, known as DHIS 2, and the website of the Kutsaga Tobacco Research Institute. Additionally, interviews were conducted with four health workers to gain insights into the background and the life cycle of malaria in Harare, and measures that the government has put in place to combat the transmission of malaria in Harare Province and current control programmes of malaria cases in Harare Province. The data collected spanned from January 1, 2013, to December 31, 2023, and included weekly suspected malaria cases, weekly positive malaria cases, daily rainfall received, daily maximum temperature, and daily minimum temperature

3.5 Target Population

Donald and Pamela (2013) defined a target population as the overall organization of individuals whom the researcher was interested in researching and analysing. The research's target population is the general population of Harare, which includes residents and visitors, who are at risk of contracting malaria within the city.

3.6 Description of variable

| Variable | Description of the variable |
|--------------------------------------|---|
| Weekly received rainfall (WRmm) | Amount of rainfall occurred in the week expressed in millimetres |
| Average Maximum temperature (WATM) | Average maximum temperature recorded in week expressed in degrees Celsius |
| Average Minimum temperature (WATm) | Average minimum temperature recorded in a week expressed in degrees Celsius |
| Weekly positive malaria cases (WPm) | Number of malaria cases reported on weekly interval |
| Weekly suspected malaria cases (WSm) | Number of weekly suspected malaria cases |
| Week (wk) | Week number in a year |

Table 1 Description of variable

3.7 Expected relation

The expected relation is that higher rainfall and higher temperatures has positively correlate with higher malaria cases

3.8 Data cleaning

For temperature, last observation carried forward technique was used to fill empty columns for daily minimum and maximum temperature data. Last observation carried forward imputation replaces missing values with the most recent observed number in the dataset, the assumption is that the missing values follow the same pattern as the most recent observed values. For rainfall missing values, the mean imputation method was used to calculate the daily missing received rainfall, the technique assumes that missing values are similar to the observed values in the dataset (Kelleher,2015). After the data was cleaned, a weekly aggregation process was

conducted. This entailed grouping the daily data by week and computing the weekly average rainfall received, weekly average maximum temperature and weekly average minimum temperatures for each week. For positive malaria cases and suspected malaria cases whatsapp artificial intelligence programme was used to estimate missing values from week from week 2 to week 40 of 2013, week 44 and week 45 of 2013.

3.9 Diagnostics test

3.9.1 Augmented Dickey Fuller Test

The test was used to check positive malaria cases and suspected malaria cases data stationarity and their exogenous factors data of weekly received rainfall, weekly average temperature and weekly minimum temperature. At 5 percent significance level. For positive malaria cases ADF Test Statistic was -4.913 with a p-value of 3.191e-05 which concluded stationary data. For malaria suspected cases it showed an augmented dickey fuller test statistic: -2.722 p-value: 0.0701 which concluded that the was not stationery and differencing was applied to make the suspected malaria data cases data stationary with ADF Statistic: -11.789 p-value: 9.897e-22 to make data stationary. For the exogenous factors

ADF Test for weekly rainfall Received showed ADF Statistic: -7.117 with p-value of 3.804e-10, ADF Test for average weekly max temp ADF Statistic: -7.708 with p-value:1.281e-11, ADF Test for average weekly min temp: ADF Statistic: -8.570 with p-value: 8.202e-14. The results concluded that the data was stationary for exogenous factors. In time series analysis, checking for stationarity is important because stationarity or non-stationarity can affect the behaviour or parts of a trend, and therefore the results.

3.9.2 Model diagnostics

Model fit plot diagnostics was used to assess the normality of residuals on the models for SARIMAX

3.9.3 Ljung -Box test

The test was used to check for significance of residuals at 5 percent significant level

3.10 Models training

The data was divided into train and test sets both the suspected and positive malaria cases test sets were set to 0.01 of the total data, and a random state of 42. SARIMAX models with a seasonality cycle of 13 weeks were developed. The model that best fits the data was chosen based on the smallest value of the Akaike Information Criterion (AIC). The data for weekly

suspected malaria cases, positive malaria cases, average weekly rainfall, average weekly maximum temperature, and average maximum temperature was normalized before being trained to build the LSTM model. The data for both positive and suspected malaria cases was also divided into train and test sets, including the exogenous factors. The test size was set to 0.01, and the random state was set to 42. The model was trained for 10 epochs for weekly positive malaria and batch size 5 and 100 epochs for weekly suspected malaria cases, batch size 3 to predict future steps of 30 weeks.

3.11 Model validation

After training the models, the researcher evaluated the models to assess their validity on the test sets based on their MAPE, MAE, and MSE on the test set data, to compare the models and determine the most suitable model for predicting positive malaria cases and positive malaria cases, based on the MAPE, MSE, and MAE metrics.

3.12 Visual inspection

Visual inspection was used to determine the trend of positive malaria cases and suspected malaria cases in Harare province, visual inspection was used to check for a monotonic trend for positive malaria cases in the data and suspected malaria cases

3.13 Ethical Considerations

The researcher observed research principles when the research was conducted. Permission was required from the City of Harare, city health department to legitimize the study. The researcher explained to the respondent's issues on confidentiality matters and also on the intention of the study

CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION

4.1 Introduction

This chapter provides a complete overview of forecasting malaria cases in Harare Province using time series models analysis. The objective is to establish trend of malaria cases in Harare Province, predict the future trend of malaria cases using time series models, and determine the best-performing model in terms MAPE. This chapter contains an in-depth exploration of SARIMAX and Lstm model, in forecast positive malaria cases and suspected malaria cases. The analysis includes preprocessing and visualization dataset, model selection, validation, and evaluation of forecasting performance. The results will provide valuable information into the

effectiveness of time series models in predicting malaria cases, contributing to the development of effective strategies for malaria control and prevention in Harare Province.

4.2 Uploading the data into Jupyter note book

The weekly suspected malaria cases and weekly positive malaria cases data was imported into jupyternote book as spread sheet file

```
import pandas as pd
df = pd.read_excel ("C:\\Users\\hp\\documents\\malaria2015.xlsx")
df.head ()
```

| | week | weekly suspected malaria | weekly positive malaria cases | weekly rainfall Received | average weekly max temp | average weekly min temp |
|---|------|--------------------------|-------------------------------|--------------------------|-------------------------|-------------------------|
| 0 | 1 | 41 | 17 | 281.9 | 27.685714 | 17.100000 |
| 1 | 2 | 168 | 25 | 138.1 | 28.700000 | 16.442857 |
| 2 | 3 | 199 | 22 | 41.2 | 29.533333 | 17.142857 |
| 3 | 4 | 215 | 23 | 7.1 | 28.942857 | 17.942857 |
| 4 | 5 | 221 | 30 | 6.9 | 26.171429 | 18.042857 |

Figure 2 importing malaria data file into jupyter

4.3 Checking the dataset for completeness.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 572 entries, 0 to 571
Data columns (total 6 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   week                                    572 non-null    int64
1   weekly suspected malaria                572 non-null    int64
2   weekly positive malaria cases          572 non-null    int64
3   weekly rainfall Received                572 non-null    float64
4   average weekly max temp                 572 non-null    float64
5   average weekly min temp                 572 non-null    float64
dtypes: float64(3), int64(3)
memory usage: 26.9 KB

```

Figure 3 Results for checking data completeness

From the result in Figure 3 the data had no missing value gaps. The data consisted for a sequency 572 data and 5 variables. This process ensures the integrity of the dataset and forms the foundation for subsequent analyses and forecasting models.

4.4 Results of converting the week into date time index

```

datetime64[ns]
date      week  weekly suspected malaria  weekly positive malaria cases  \
date
2013-01-08    1                41                17
2013-01-15    2               168                25
2013-01-22    3               199                22
2013-01-29    4               215                23
2013-02-05    5               221                30

date      weekly rainfall Received  average weekly max temp  \
date
2013-01-08                281.9                27.685714
2013-01-15                138.1                28.700000
2013-01-22                 41.2                29.533333
2013-01-29                 7.1                28.942857
2013-02-05                 6.9                26.171429

date      average weekly min temp
date
2013-01-08                17.100000
2013-01-15                16.442857
2013-01-22                17.142857
2013-01-29                17.942857
2013-02-05                18.042857

```

Figure 4 Changing the week column into datetime index

Figure 4 shows the conversion of week column into datetime column which was set as the index and frequency for the timeseries forecast with frequency = 1 week of (w – TUE)

4.5 Descriptive Statistics

```
count    572.000000
mean     203.143357
std      143.568983
min       4.000000
25%      77.000000
50%     184.000000
75%     289.000000
max      762.000000
Name: weekly suspected malaria, dtype: float64
```

Figure 5 Descriptive statistics for weekly suspected malaria cases

Figure 5 presents the descriptive statistics for the weekly suspected malaria cases. The data shows an average of 203.14 suspected malaria cases per week, with a large standard deviation of 143.5 indicating substantial variability and spread in the weekly case counts. At the lower end of the distribution, the minimum cases of suspected malaria recorded in a single week was 4. Looking at the percentiles, 25% of weeks had 77 or fewer suspected cases, the median (50th percentile) was 184 or fewer suspected cases per week, and 75% of weeks saw 289 or fewer suspected cases. At the high end, the maximum number of suspected malaria cases recorded in a single week was 762. The weekly suspected malaria cases exhibited a wide range, from a minimum of 4 up to peak of 762, with average suspected malaria cases of about 203 cases per week. However, the data was highly variable, as indicated by the large standard deviation of 143.5.

```
count    572.000000
mean     31.664336
std      53.678741
min      0.000000
25%      4.000000
50%     12.000000
75%     33.000000
max     329.000000
Name: weekly positive malaria cases, dtype: float64
```

Figure 6 Descriptive statistics for weekly positive malaria cases

Figure 6 presents the descriptive statistics for the weekly positive malaria cases. On average, the data shows 31.664 positive malaria cases per week. However, the standard deviation is quite high at 53.67, indicating significant variability and a wide spread in the weekly case counts. At the low end, the minimum number of positive malaria cases recorded in a single week was 0. Moving up the distribution, the 25th percentile shows that 25% of weeks had 4 or fewer positive cases. The median, or 50th percentile, was 12 positive cases per week. 75% of weeks saw 33 or fewer positive cases. At the high end, the maximum number of positive malaria cases recorded in a single week was 362. The average positive case count per week was around 32, the data exhibited a large range, from weeks with no positive cases up to an extreme high of 362 positive cases in a single week, as indicated by the high standard deviation.

4.6 Graphical Visualisation

Figure 7 and figure 8 show the graphs of suspected malaria cases and positive malaria cases, the two graphs both exhibits a diminishing trend

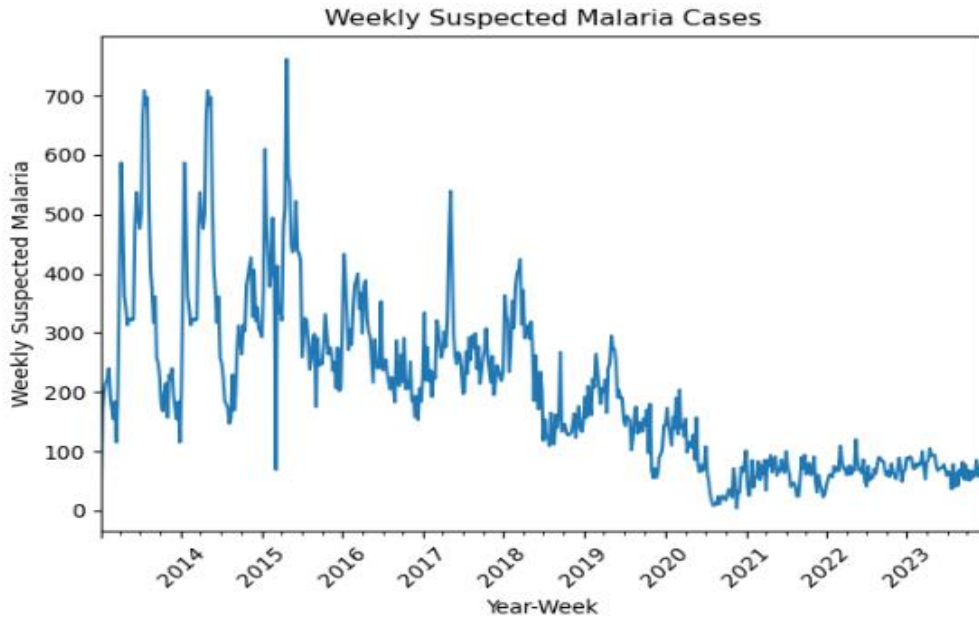


Figure 7 Graph for weekly suspected malaria cases

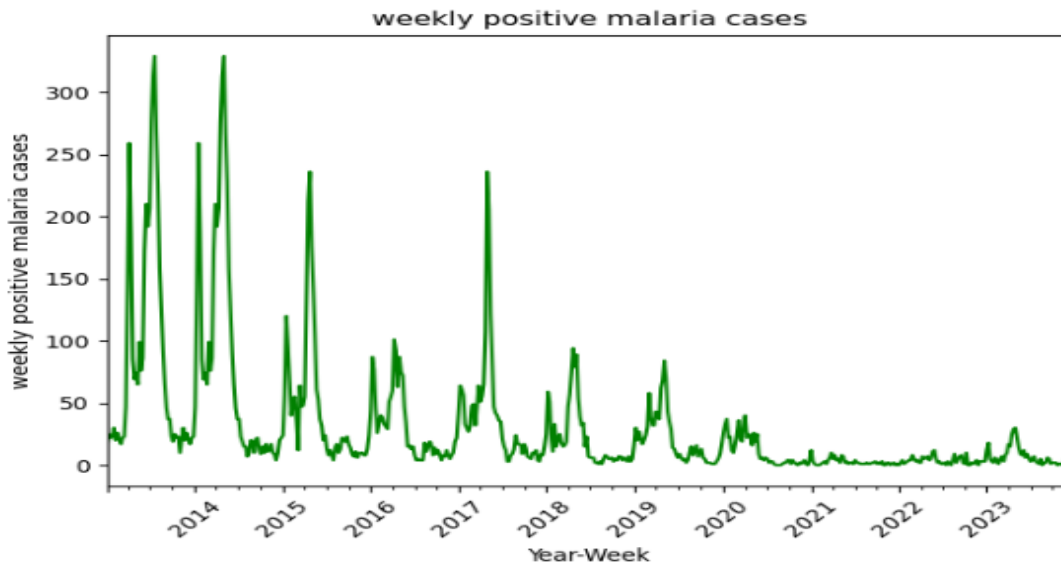


Figure 8 Trend for weekly positive malaria cases

4.7 Stationarity test

ADF Test Statistic: -2.722702760258099

p-value: 0.07019141584117579

The data is non-stationary (fail to reject the null hypothesis)

Figure 9 ADF Test statistics for weekly suspected malaria cases

ADF Test Statistic: -4.252697636119653

p-value: 0.0005358845782033955

The data is stationary (reject the null hypothesis)

Figure 10 ADF Test statistics for weekly positive malaria cases

ADF Statistic: -11.789627307404183

p-value: 9.897306614947402e-22

Critical Values:

1%: -3.4419977165341673

5%: -2.866678179017994

10%: -2.5695064902419396

The data is stationary (reject the null hypothesis)

Figure 11 ADF Test statistics for differenced weekly positive malaria cases

ADF Test for weekly rainfall Received:

ADF Statistic: -7.123908303378895

p-value: 3.662083893875041e-10

Critical Values:

1%: -3.4422521197633187

5%: -2.866790184232015

10%: -2.569566175304558

ADF Test for average weekly max temp:

ADF Statistic: -7.703814049416904

p-value: 1.317756674752786e-11

Critical Values:

1%: -3.442102384299813

5%: -2.8667242618524233

10%: -2.569531046591633

ADF Test for average weekly min temp:

ADF Statistic: -8.72149541113173

p-value: 3.3759289130733035e-14

Critical Values:

1%: -3.44218748274498

5%: -2.8667617276005006

10%: -2.569551011281552

The data is stationary (reject the null hypothesis)

Figure 12 ADF Test statistics for exogenous variables

The figures (9, 10, 11 and 12) shows the results of the ADF test performed for weekly positive malaria cases, weekly suspected malaria cases, weekly received rainfall, and average weekly maximum temperature variables. For the weekly positive malaria cases, the argument dickey fuller test statistic is -4.919, p-value of 3.198e-05, indicating that the data is stationary. For the weekly suspected malaria cases, the initial ADF test statistic is -2.722 with p-value 0.0701, suggesting that the data is not stationary. After performing differencing, the augmented dickey fuller test statistic for the differenced data -11.789 with a p-value of 9.897e-22, indicating stationary. For exogenous factors, the data shows stationary as well: the augmented dickey fuller test statistic for weekly rainfall received: -7.117 and p-value of 3.804e-10, the ADF test statistic for average weekly maximum temperature is -7.708 with p-value 1.281e-11, and the ADF test

statistic for average weekly minimum temperature is -8.570 with a p-value of $8.202e-14$. At a significance level of 5 percent significant level.

4.8 Decomposition graph for suspected malaria cases and positive

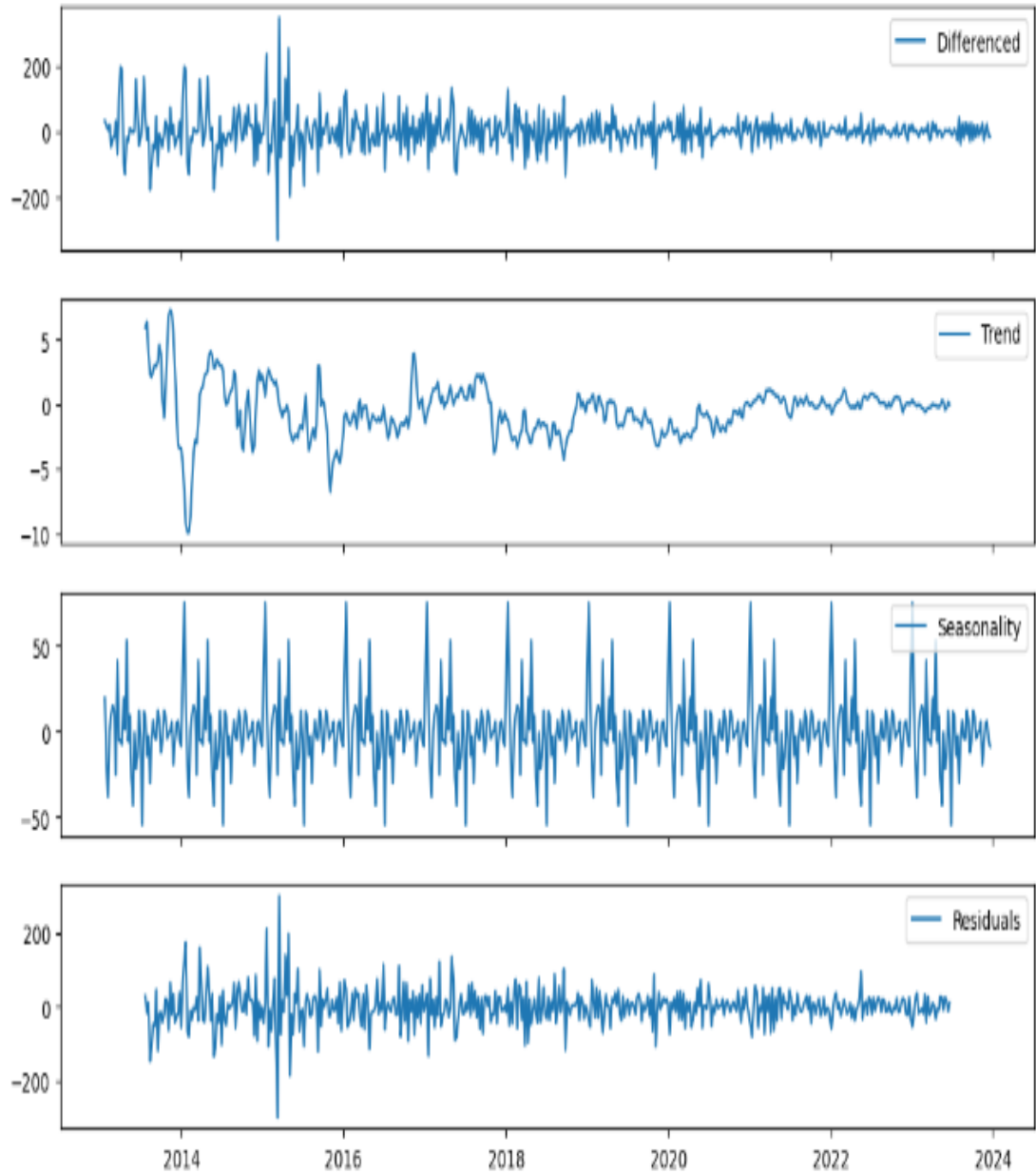


Figure 13 Decomposed plot for weekly suspected malaria cases

The weekly suspected malaria cases' time series breakdown is displayed in Figure 13. The observed data, or the real weekly time series of suspected malaria cases, is shown in the first graphic. The trend component is seen in the second plot. It shows a gradual decline in the total number of suspected cases of

malaria. The seasonal component is displayed in the third plot. The recurring pattern in this graphic over time indicates that there is seasonality in the data. The residuals are shown in the final plot. These are the data variances that the observed trend and seasonal components are unable to explain. The unexplained fluctuations or irregular causes impacting the suspected numbers of malaria cases are represented by the residuals.

4.8.1 Decomposed graph for suspected malaria cases

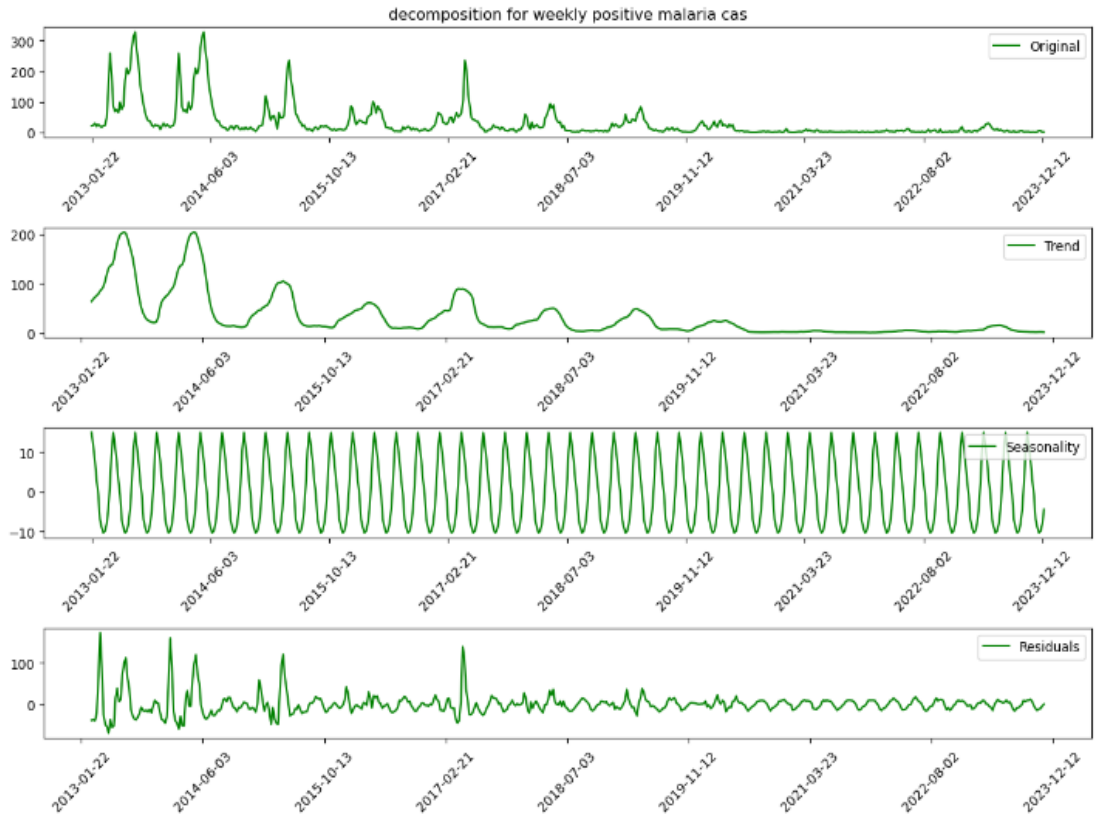


Figure 14 Decomposed plot for weekly positive malaria cases

Figure 14 shows the decomposition plots for weekly positive malaria cases. First plot displays the observed data - the actual weekly time series of positive malaria cases. Second plot reveals the trend component. The trend indicates a decrease in overall number of weekly positive malaria cases from 2013 up to 2023. Third plot shows the seasonal component. The plot exhibits a recurrent pattern over time, demonstrating seasonality in the positive malaria case data. Last plot show residuals. These are the variations in the data that are not explained by the identified trend and seasonal components. The residuals represent the unexplained fluctuations or irregular factors influencing the weekly positive malaria case counts

4.9 ACF and PACF for weekly suspected

Fig 15 shows an ACF and PACF plots portrayed below reflected about how observations in time series are being indexed over time and how well are they related to each other for suspected malaria cases up lag 20. The two plots clarify the order concept in Autoregressive and Moving Averages in time series analysis. The PACF shows 4 significant lags

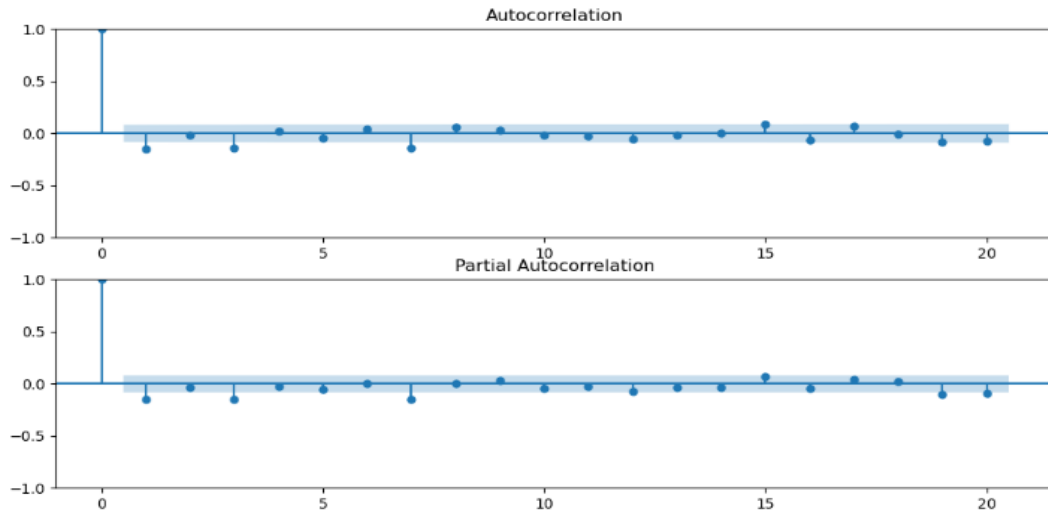


Figure 15 ACF and PACF plot for weekly suspected malaria cases

4.9.1 ACF and PACF for weekly positive malaria cases

Figure 16 shows ACF and PACF plots for weekly positive malaria cases with ACF plot showing a slowly decay from lag 1 up to lag 20. The ACF shows that 10 significant lag excluding the lag 0 which shows the relationship between the current lag and its self

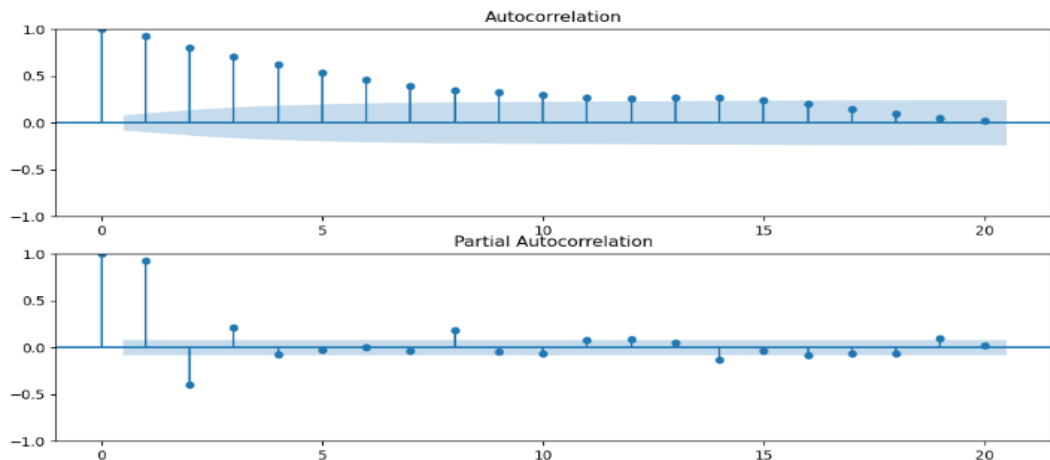


Figure 16 ACF and PACF plot for weekly positive malaria cases

4.10 Model identification for weekly suspected malaria cases

| 0% | 0/16 [00:00<?, ?it/s] | |
|----|-----------------------|-------------|
| | (p,q,P,Q) | AIC |
| 0 | (3, 3, 0, 0) | 6097.689632 |
| 1 | (3, 3, 3, 3) | 6098.539082 |
| 2 | (3, 3, 3, 0) | 6099.335329 |
| 3 | (3, 3, 0, 3) | 6100.113663 |
| 4 | (0, 3, 3, 3) | 6114.423674 |
| 5 | (3, 0, 3, 3) | 6117.474990 |
| 6 | (0, 3, 3, 0) | 6117.863725 |
| 7 | (0, 3, 0, 0) | 6118.341819 |
| 8 | (0, 3, 0, 3) | 6119.153702 |
| 9 | (3, 0, 3, 0) | 6119.700084 |
| 10 | (3, 0, 0, 0) | 6121.096403 |
| 11 | (3, 0, 0, 3) | 6121.308351 |
| 12 | (0, 0, 3, 0) | 6137.794763 |
| 13 | (0, 0, 0, 3) | 6139.218723 |
| 14 | (0, 0, 0, 0) | 6140.072208 |
| 15 | (0, 0, 3, 3) | 6145.190555 |

Figure 17 Model identification parameters with the lowest AIC

Figure 17 shows all AIC values with their respective to the p, q, P, Q parameters used to select the best-fit model. For the weekly suspected malaria cases, the data required differencing once to reach stationarity, so $D=1$ and $d=0$. The best-fit model for the weekly suspected malaria cases was the SARIMAX (3,0,3)(0,0,0,13) model. .

4.10.1 summary of the best-fit model results

SARIMAX Results

```

=====
Dep. Variable:   weekly suspected malaria   No. Observations:   565
Model:          SARIMAX(3, 1, 3)           Log Likelihood      -3038.845
Date:           Mon, 27 May 2024          AIC                 6097.690
Time:           13:02:39                   BIC                 6141.040
Sample:         0                           HQIC                6114.612
Covariance Type: opg
=====

```

| | coef | std err | z | P> z | [0.025 | 0.975] |
|--------------------------|-----------|---------|---------|-------|----------|----------|
| weekly rainfall Received | -0.1345 | 0.071 | -1.898 | 0.058 | -0.273 | 0.004 |
| average weekly max temp | 1.1803 | 1.279 | 0.923 | 0.356 | -1.325 | 3.686 |
| average weekly min temp | -0.4134 | 1.100 | -0.376 | 0.707 | -2.570 | 1.743 |
| ar.L1 | -1.0128 | 0.051 | -19.803 | 0.000 | -1.113 | -0.913 |
| ar.L2 | 0.5928 | 0.050 | 11.765 | 0.000 | 0.494 | 0.692 |
| ar.L3 | 0.7513 | 0.048 | 15.780 | 0.000 | 0.658 | 0.845 |
| ma.L1 | 0.8548 | 0.049 | 17.305 | 0.000 | 0.758 | 0.952 |
| ma.L2 | -0.8538 | 0.033 | -25.911 | 0.000 | -0.918 | -0.789 |
| ma.L3 | -0.9059 | 0.045 | -20.173 | 0.000 | -0.994 | -0.818 |
| sigma2 | 2803.1966 | 96.215 | 29.135 | 0.000 | 2614.619 | 2991.775 |

```

=====
Ljung-Box (L1) (Q):      0.52   Jarque-Bera (JB):      947.16
Prob(Q):                 0.47   Prob(JB):              0.00
Heteroskedasticity (H): 0.08   Skew:                  0.87
Prob(H) (two-sided):    0.00   Kurtosis:              9.10
=====

```

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 18 Summary Statistics for SARIMAX(3,1,3)(0,0,0,13)

Figure 18 shows SARIMAX(3,1,3)(0,0,0,13) and coefficient for weekly suspected malaria cases

Model Fit:

The log-likelihood coefficient is -3038.845, the Akaike Information Criterion (AIC) is 6097.690, the Bayesian Information Criterion (BIC) is 6141.040, and the Hannan-Quinn Information Criterion (HQIC) is 6114.612. This information criteria suggest that the model provides a reasonable fit to the data, as lower values indicate better model fit.

Coefficient Estimates:

The coefficient estimates for the model's terms are provided, along with their standard errors, z-statistics, and p-values. Coefficients for weather variables (weekly rainfall received, average weekly maximum temperature, and average weekly minimum temperature) are not statistically significant at the 5% level, indicating that these variables may not significant impact on weekly

suspected malaria cases. AR and MA coefficients are all statistically significant, suggesting these terms are important in capturing the temporal dynamics of the time series.

Residual Diagnostics:

The Ljung-Box (L1) test for autocorrelation in the residuals has a p-value of 0.47, indicating that there is no significant autocorrelation in the residuals.

The Jarque-Bera (JB) test for normality of the residuals has a p-value of 0.00, suggesting that the residuals do not follow a normal distribution. The heteroskedasticity (H) test has a p-value of 0.00, indicating the presence of heteroskedasticity (non-constant variance) in the residuals.

4.11 Model identification for weekly positive malaria cases

Figure 19 To select the best fit model, the researchers used (AIC). The model with the lowest AIC was chosen as best fit model for the weekly positive malaria cases, the data was stationary, so no differencing was required ($d=0$ and $D=0$). The best fit model for the weekly suspected malaria cases was the SARIMAX(6,0,6)(0,0,6,13). The summary of the results for best fit model is presented in Figure 20

| Order | (p, q, P, Q) | AIC |
|-------|--------------|-------------|
| 0 | (6, 6, 0, 6) | 4856.243859 |
| 1 | (6, 6, 6, 0) | 4860.339899 |
| 2 | (6, 6, 6, 6) | 4873.668450 |
| 3 | (6, 6, 0, 0) | 4877.200515 |
| 4 | (6, 0, 0, 6) | 4877.445818 |
| 5 | (6, 0, 6, 0) | 4879.387166 |
| 6 | (6, 0, 6, 6) | 4890.432174 |
| 7 | (6, 0, 0, 0) | 4892.897437 |
| 8 | (0, 6, 0, 6) | 4910.586440 |
| 9 | (0, 6, 6, 0) | 4912.408197 |
| 10 | (0, 6, 0, 0) | 4932.854910 |
| 11 | (0, 6, 6, 6) | 4933.189798 |
| 12 | (0, 0, 6, 0) | 5853.054918 |
| 13 | (0, 0, 0, 6) | 5974.166184 |
| 14 | (0, 0, 6, 6) | 6080.294322 |
| 15 | (0, 0, 0, 0) | 6129.774553 |

Figure 19 Model identification parameters for weekly positive malaria cases

4.11.1 Summary statistics for the identified model

```

=====
SARIMAX Results
=====
Dep. Variable:    weekly positive malaria cases    No. Observations:    565
Model:           SARIMAX(6, 0, 6)x(0, 0, 6, 13)    Log Likelihood       -2406.122
Date:           Fri, 24 May 2024                  AIC                  4856.244
Time:           10:35:54                          BIC                  4951.654
Sample:         0                                  HQIC                 4893.484

Covariance Type:    opg
=====
              coef    std err          z      P>|z|      [0.025    0.975]
-----
weekly rainfall Received    -0.0345    0.021    -1.648    0.099    -0.076    0.007
average weekly max temp     0.1074    0.283     0.380    0.704    -0.446    0.661
average weekly min temp     0.1682    0.213     0.790    0.430    -0.249    0.586
ar.L1                       0.8940    0.318     2.813    0.005     0.271    1.517
ar.L2                       -0.6616    0.172    -3.840    0.000    -0.999   -0.324
ar.L3                        0.7066    0.250     2.821    0.005     0.216    1.198
ar.L4                       -0.1559    0.201    -0.777    0.437    -0.549    0.238
ar.L5                        0.5595    0.186     3.010    0.003     0.195    0.924
ar.L6                       -0.4927    0.271    -1.819    0.069    -1.024    0.038
ma.L1                        0.4965    0.319     1.556    0.120    -0.129    1.122
ma.L2                        0.6083    0.447     1.362    0.173    -0.267    1.484
ma.L3                        0.1578    0.584     0.270    0.787    -0.986    1.302
ma.L4                       -0.0008    0.487    -0.002    0.999    -0.956    0.954
ma.L5                       -0.5492    0.418    -1.313    0.189    -1.369    0.271
ma.L6                       -0.1062    0.143    -0.741    0.459    -0.387    0.175
ma.S.L13                     0.1254    0.045     2.795    0.005     0.037    0.213
ma.S.L26                     0.1045    0.052     2.027    0.043     0.003    0.206
ma.S.L39                     0.0844    0.040     2.113    0.035     0.006    0.163
ma.S.L52                     0.2672    0.037     7.236    0.000     0.195    0.340
ma.S.L65                     0.0850    0.048     1.780    0.075    -0.009    0.179
ma.S.L78                     0.0650    0.066     0.991    0.322    -0.064    0.193
sigma2                       287.1170    10.606    27.072    0.000    266.330    307.904

=====
Ljung-Box (L1) (Q):           0.00    Jarque-Bera (JB):           2082.21
Prob(Q):                      0.95    Prob(JB):                   0.00
Heteroskedasticity (H):       0.04    Skew:                       1.12
Prob(H) (two-sided):          0.00    Kurtosis:                   12.14
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```

Figure 20 Summary statistics for SARIMAX(6,0,6)(0,0,6,13)

Figure 20 provides a comprehensive summary of the model's performance and relationships between response variable and explanatory variables.

The coefficient for weekly rainfall received is -0.0345, indicating a negative relationship between rainfall and weekly positive malaria cases, though this relationship is not statistically significant at the 5% level (p-value = 0.099).

The coefficients for average weekly maximum and minimum temperatures are 0.1074 and 0.168, respectively, suggesting a positive relationship between temperatures and malaria cases, but these relationships are also not statistically significant (p-values > 0.05).

The autoregressive (AR) terms show a mix of positive and negative coefficients, with significant coefficients for lags 1, 2, 3, 5, and 6, indicating a complex autoregressive structure in the data.

The moving average (MA) terms also show a mix of positive and negative coefficients, but only the seasonal MA terms at lags 13, 26, 39, 52, and 65 are statistically significant, suggesting the presence of a strong seasonal component in the data.

The model has a log-likelihood of -2406.122, an AIC of 4856.244, and a BIC of 4951.654, indicating a reasonably good fit.

The Ljung-Box test for autocorrelation in the residuals is not significant (p-value = 0.95), suggesting the model has adequately captured the temporal structure in the data.

The Jarque-Bera test rejects the null hypothesis of normality in the residuals (p-value < 0.01), indicating potential non-normality. The Heteroskedasticity (H) test also suggests the presence of heteroskedasticity in the residuals (p-value < 0.01).

4.12 Residual Analysis for weekly suspected malaria cases

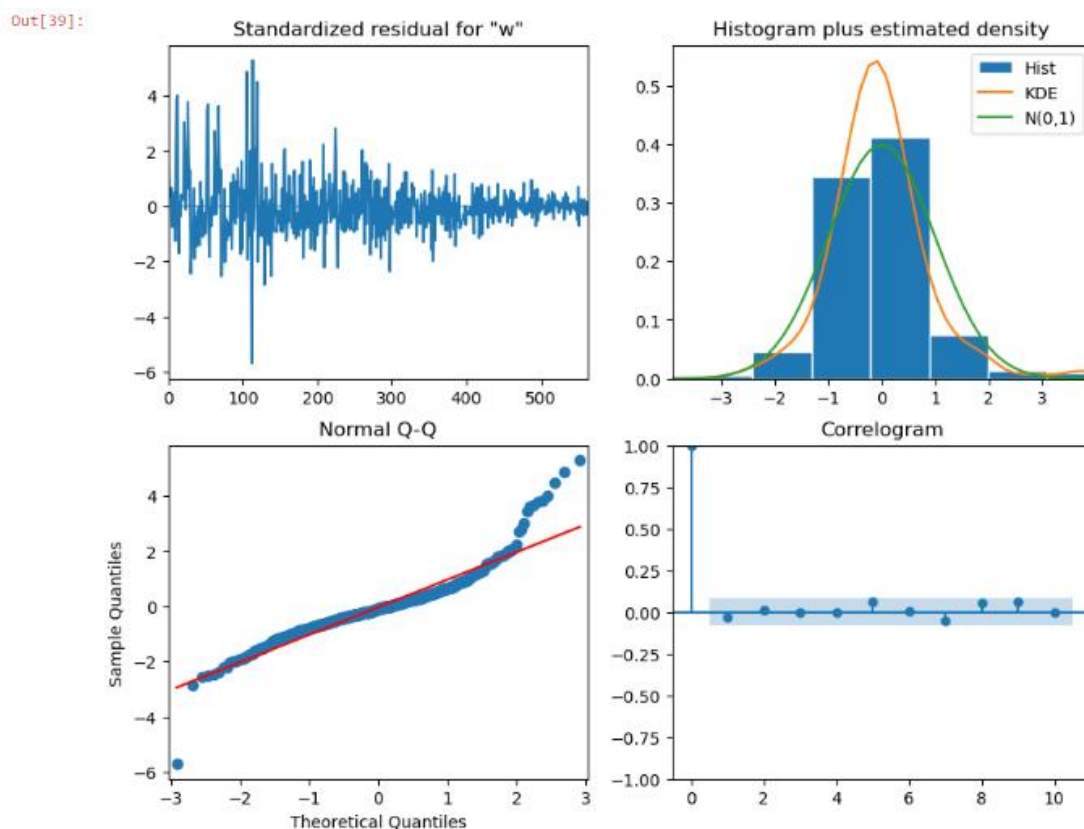


Figure 21 Model fit plot diagnostics for weekly suspected malaria cases

The residual analysis of the chosen model, SARIMAX(3,1,3)(0,0,0,13), is displayed in Figure 21. Similar to white noise, the residuals show no trend and a relatively steady variance over time.

The residual distribution in the top-right plot closely resembles a normal distribution. The Q-Q

plot in the bottom left, which displays a comparatively straight line that sits on $y = x$, lends additional credence to this. Lastly, similar to white noise, the correlogram displays no significant coefficients after lag 0 with no significant latency. As a result, the residuals of this model resemble white noise when viewed graphically.

4.12.1 Residual analysis for weekly positive malaria cases

Out[42]:

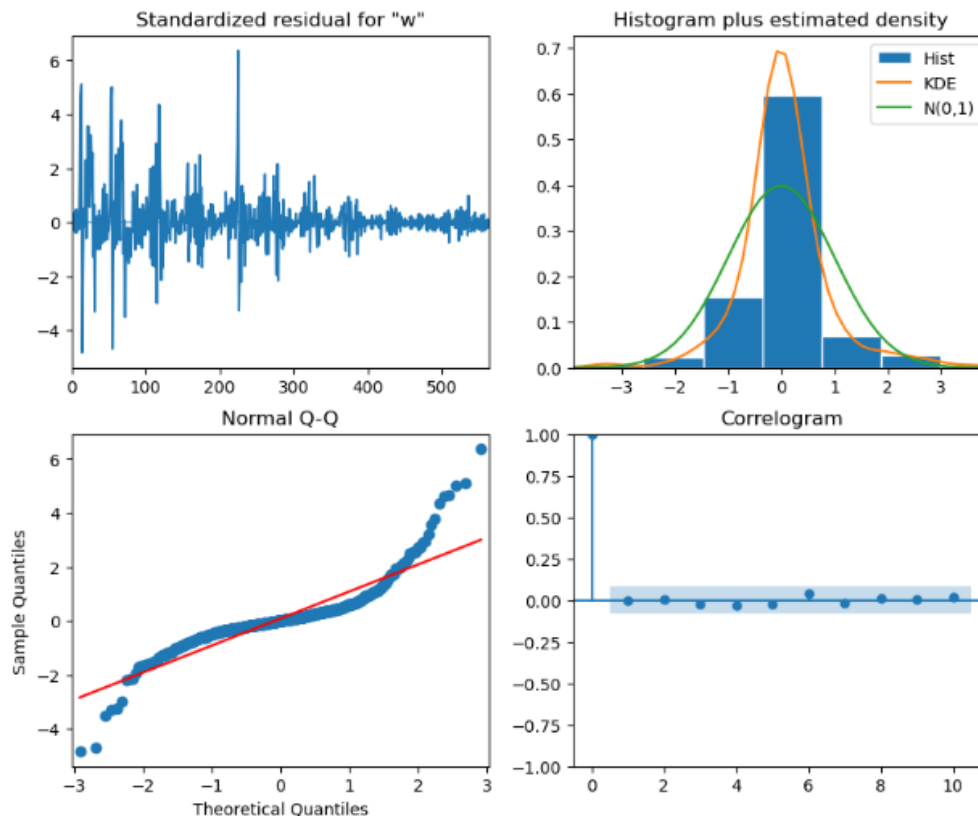


Figure 22 Model fit plot diagnostics for positive malaria cases

Figure 22 shows the residual analysis of the selected model $SARIMAX(6,0,6)(0,0,6,13)$. Compared to white noise, the residuals show no trend and a relatively steady variance over time. The residual distribution in the top-right plot resembles a normal distribution quite a bit. The Q-Q plot in the bottom left, which displays a comparatively straight line that sits on $y = x$, lends additional credence to this. Lastly, following lag 0, the correlogram displays no significant coefficients, much like white noise. As a result, the residuals of this model resemble white noise when viewed graphically.

4.13 Prediction for the SARIMAX on the test set

| | actual | pred_last_value | pred_SARIMAX |
|-----|--------|-----------------|--------------|
| 561 | 68 | 52 | 66.920709 |
| 562 | 51 | 68 | 58.544307 |
| 563 | 67 | 51 | 67.437900 |
| 564 | 56 | 67 | 59.491986 |
| 565 | 61 | 56 | 64.188726 |
| 566 | 85 | 61 | 61.114217 |
| 567 | 58 | 85 | 54.930200 |
| 568 | 57 | 58 | 85.842959 |
| 569 | 79 | 57 | 62.130782 |
| 570 | 64 | 79 | 52.628372 |
| 571 | 49 | 64 | 82.509127 |

Figure 23 SARIMAX(3,1,3)(0,0,0,13) test set predictions

Figure 23 shows SARIMAX(3,1,3)(0,0,0,13) predicated values for suspected malaria cases from week 561 to week 571 values on the test values

| | actual | pred_last_value | pred_SARIMAX |
|-----|--------|-----------------|--------------|
| 561 | 2 | 2 | 1.326553 |
| 562 | 2 | 2 | 5.788167 |
| 563 | 0 | 2 | 0.353447 |
| 564 | 1 | 0 | 3.430311 |
| 565 | 1 | 1 | -0.231840 |
| 566 | 0 | 1 | 0.437880 |
| 567 | 2 | 0 | 1.182551 |
| 568 | 6 | 2 | 0.783054 |
| 569 | 5 | 6 | 2.063203 |
| 570 | 1 | 5 | 7.676716 |
| 571 | 1 | 1 | 4.357066 |

Figure 24 SARIMAX(2,0,3)(3,0,2,13) test set predictions

Figure 24 shows SARIMAX(6,0,6)(0,0,6,13) predicated values for weekly positive malaria cases from week 561 to week 571 on the test set

4.14 Model validation

| | MAE | MAPE | MSE |
|---|-------|--------|--------|
| Suspected malaria cases SARIMAX(3,1,3)(0,0,0,13) | 12.12 | 20.00% | 275.73 |
| Positive malaria cases SARIMAX(6,0,6)(0,0,6,13) | 2.54 | Inf% | 10.45 |

Table 2 Model validation for SARIMAX models

Table 3 shows results for the SARIMAX models, the SARIMAX (3,1,3)(0,0,0,13) MAPE for prediction 20.00%. The Mean Absolute Percentage Error (MAPE) is 20.00%, indicating that on average, the model's predictions differ from the actual values by 20% in relative terms. A MAPE of 20% suggests the SARIMAX model is making reasonably accurate percentage-based forecasts, with a moderate level of error. The mean squared error is 275.39, which measures the average squared difference between the predicted with actual values. This relatively high MSE value suggests the SARIMAX model is making predictions with a larger degree of absolute error, on average. The mean absolute error is 12.12, indicating the average absolute difference between the predicted and actual values. The MAE with 12.12 indicates that the SARIMAX model is making predictions with an average absolute error of approximately 12 units for the weekly suspected malaria cases. The above table shows model performance metrics for the SARIMAX (6,0,6)(0,0,6,13) forecast of the weekly positive malaria cases. The MAPE (Mean Absolute Percentage Error) for the SARIMAX (6,0,6)(0,0,6,13) prediction is reported as inf%, indicating that the model is unable to make accurate percentage-based forecasts. This is likely due to the presence of zero or near-zero values in the actual time series, which can cause the percentage error to become extremely large or undefined. The MSE (Mean Squared Error) for the SARIMAX (6,0,6)(0,0,6,13)prediction is 10.45, suggesting the model is making reasonably accurate predictions, on average, with a moderate degree of error. The MAE (Mean Absolute Error) for the SARIMAX (6,0,6)(0,0,6,13) prediction is 2.54, indicating the model is making predictions with an average absolute error of 2.54 units.

4.15 Forecasting the Future Weekly Suspected Malaria Forecast

| | actual | pred_last_value | pred_SARIMAX |
|----|--------|-----------------|--------------|
| 0 | 68.0 | 52 | 68.818557 |
| 1 | 51.0 | 68 | 69.196579 |
| 2 | 67.0 | 51 | 68.488238 |
| 3 | 56.0 | 67 | 61.254996 |
| 4 | 61.0 | 56 | 60.794478 |
| 5 | 85.0 | 61 | 64.238223 |
| 6 | 58.0 | 85 | 53.541036 |
| 7 | 57.0 | 58 | 85.416874 |
| 8 | 79.0 | 57 | 63.703904 |
| 9 | 64.0 | 79 | 53.326564 |
| 10 | 49.0 | 64 | 81.366124 |
| 11 | NaN | 49 | 76.609152 |
| 12 | NaN | 49 | 61.925162 |
| 13 | NaN | 49 | 53.391400 |
| 14 | NaN | 49 | 76.971536 |
| 15 | NaN | 49 | 64.156057 |
| 16 | NaN | 49 | 57.653869 |
| 17 | NaN | 49 | 73.805849 |
| 18 | NaN | 49 | 58.126338 |
| 19 | NaN | 49 | 57.535170 |
| 20 | NaN | 49 | 74.977392 |
| 21 | NaN | 49 | 59.583807 |
| 22 | NaN | 49 | 56.130168 |
| 23 | NaN | 49 | 77.395753 |
| 24 | NaN | 49 | 59.935334 |
| 25 | NaN | 49 | 57.435622 |
| 26 | NaN | 49 | 73.577465 |
| 27 | NaN | 49 | 60.841836 |
| 28 | NaN | 49 | 60.894145 |
| 29 | NaN | 49 | 76.642143 |
| 30 | NaN | 49 | 57.594513 |
| 31 | NaN | 49 | 56.155879 |
| 32 | NaN | 49 | 75.421567 |
| 33 | NaN | 49 | 58.681207 |
| 34 | NaN | 49 | 58.721912 |
| 35 | NaN | 49 | 72.985005 |
| 36 | NaN | 49 | 61.030703 |
| 37 | NaN | 49 | 60.141327 |
| 38 | NaN | 49 | 72.976017 |
| 39 | NaN | 49 | 60.171646 |
| 40 | NaN | 49 | 60.355821 |

Figure 25 Forecasted results for weekly suspected malaria cases

Fig 25 shows the forecasted results for weekly suspected ma; for the next 30 weeks for weekly suspected malaria cases.

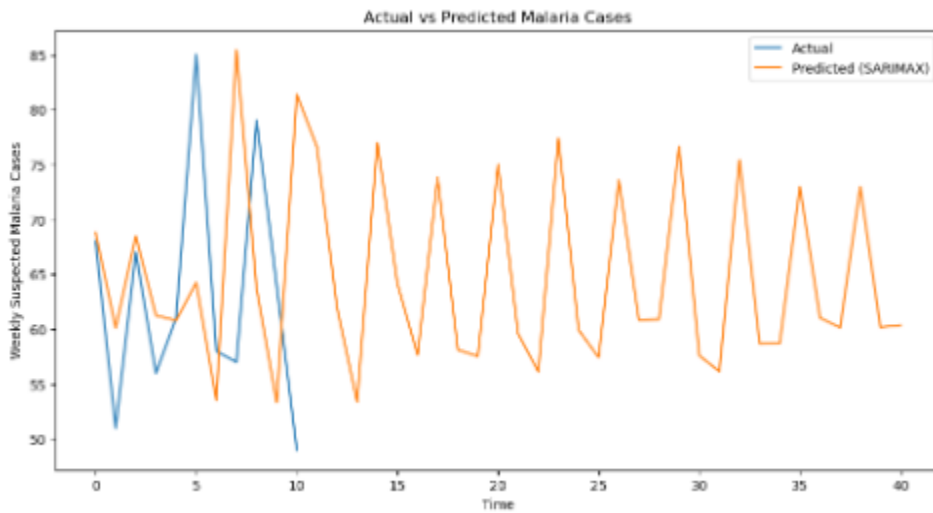


Figure 26 Graph forecast for weekly suspected malaria cases

Figure 26 shows the graph for weekly suspected malaria cases from 0 – 10 time it shows predicted SARIMAX versus actual weekly suspected malaria cases in dataset or the test set. From 11- 40 time it shows the predicted suspected malaria cases

4.16 Weekly positive malaria forecast

| | actual | pred_last_value | pred_SARIMAX |
|----|--------|-----------------|--------------|
| 0 | 2.0 | 2 | -0.359775 |
| 1 | 2.0 | 2 | 5.141341 |
| 2 | 0.0 | 2 | 0.809185 |
| 3 | 1.0 | 0 | 3.094283 |
| 4 | 1.0 | 1 | -0.408399 |
| 5 | 0.0 | 1 | 1.487610 |
| 6 | 2.0 | 0 | 0.263123 |
| 7 | 6.0 | 2 | 0.629804 |
| 8 | 5.0 | 6 | 3.257232 |
| 9 | 1.0 | 1 | 6.751317 |
| 10 | 1.0 | 1 | 4.032482 |
| 11 | NaN | 1 | 1.583622 |
| 12 | NaN | 1 | 0.812316 |
| 13 | NaN | 1 | 1.993390 |
| 14 | NaN | 1 | 1.995545 |
| 15 | NaN | 1 | 0.012971 |
| 16 | NaN | 1 | 1.003913 |
| 17 | NaN | 1 | 1.003449 |
| 18 | NaN | 1 | 0.012761 |
| 19 | NaN | 1 | 1.993695 |
| 20 | NaN | 1 | 5.956150 |
| 21 | NaN | 1 | 4.965038 |
| 22 | NaN | 1 | 1.002933 |
| 23 | NaN | 1 | 1.001478 |
| 24 | NaN | 1 | 1.582117 |
| 25 | NaN | 1 | 0.818569 |
| 26 | NaN | 1 | 1.983499 |
| 27 | NaN | 1 | 1.985861 |
| 28 | NaN | 1 | 0.028147 |
| 29 | NaN | 1 | 1.006377 |
| 30 | NaN | 1 | 1.005689 |
| 31 | NaN | 1 | 0.027460 |
| 32 | NaN | 1 | 1.982969 |
| 33 | NaN | 1 | 5.894801 |
| 34 | NaN | 1 | 4.916102 |
| 35 | NaN | 1 | 1.004352 |
| 36 | NaN | 1 | 1.002601 |
| 37 | NaN | 1 | 1.575982 |
| 38 | NaN | 1 | 0.822073 |
| 39 | NaN | 1 | 1.971446 |
| 40 | NaN | 1 | 1.973662 |

Figure 27 Forecasted results for weekly positive malaria cases

Figure 27 shows SARIMAX (6,0,6)(0,0,6,13) forecast for the 30 weeks of 2024 for weekly positive malaria cases

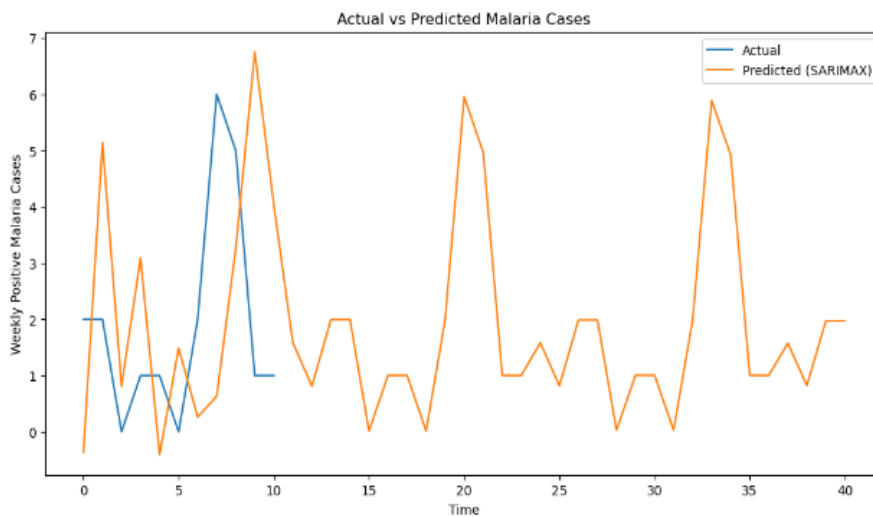


Figure 28 Graph for weekly positive malaria forecast

Figure 28 shows a graph of weekly positive malaria cases from 0 to 10, displaying the predicted SARIMAX and the actual weekly positive malaria cases in the dataset or test set. From 11 to 40, it displays the future forecast of expected malaria cases.

4.17 Long Short Time Memory

| | week | weekly suspected malaria | weekly positive malaria cases \ |
|-------|------------|--------------------------|---------------------------------|
| count | 572.000000 | 572.000000 | 572.000000 |
| mean | 286.500000 | 0.262722 | 0.096244 |
| std | 165.266452 | 0.189405 | 0.163157 |
| min | 1.000000 | 0.000000 | 0.000000 |
| 25% | 143.750000 | 0.096306 | 0.012158 |
| 50% | 286.500000 | 0.237467 | 0.036474 |
| 75% | 429.250000 | 0.375989 | 0.100304 |
| max | 572.000000 | 1.000000 | 1.000000 |

| | weekly rainfall Received | average weekly max temp \ |
|-------|--------------------------|---------------------------|
| count | 572.000000 | 572.000000 |
| mean | 0.065961 | 0.619968 |
| std | 0.123346 | 0.149829 |
| min | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.526020 |
| 50% | 0.002838 | 0.635075 |
| 75% | 0.074317 | 0.718067 |
| max | 1.000000 | 1.000000 |

| | average weekly min temp |
|-------|-------------------------|
| count | 572.000000 |
| mean | 0.313060 |
| std | 0.146904 |
| min | 0.000000 |
| 25% | 0.184009 |
| 50% | 0.341715 |
| 75% | 0.440445 |
| max | 1.000000 |

Figure 29 Normalised before being used for Lstm development

Fig 29 shows the normalised data results for an LSTM model before being split for training and testing data.

Out[9]: <matplotlib.legend.Legend at 0x1dbefee6a10>

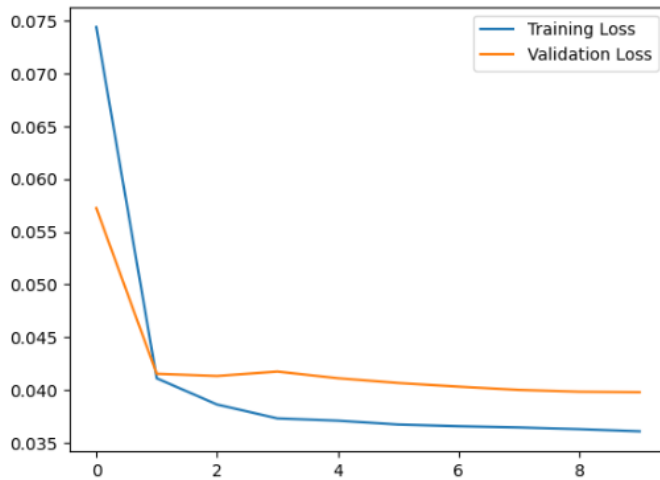


Figure 30 Training and Validation loss Graph for weekly suspected malaria cases

Fig 30 shows the training and validation loss graphs for weekly suspected malaria cases, the graph shows a narrow space between training loss function and validation loss function implying a difference in performance between model ability to fit the training data and its ability to extrapolate to new, unseen data. The small gaps between training loss and validation indicate a good generalization

Out[15]: <matplotlib.legend.Legend at 0x1dbf3dc2f50>

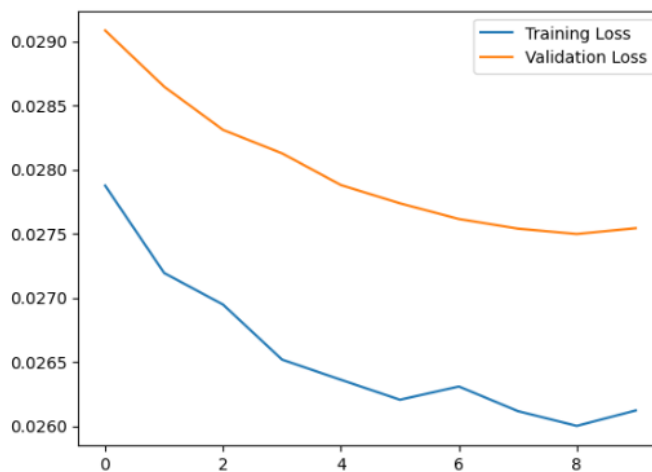


Figure 31 Training and validation loss graph for weekly positive malaria cases

Fig 31 shows the training and validation loss function graphs for weekly suspected malaria cases, the graph shows a major difference between the training loss and validation loss. function implying the difference in performance between model's ability to fit the training data and its

ability to generalise to new future data, the large gaps between train loss and validation indicate suggest an overfit

4.18 Model summary for Lstm for positive malaria cases

Model: "sequential_22"

| Layer (type) | Output Shape | Param # |
|------------------|--------------|---------|
| lstm_26 (LSTM) | (None, 10) | 480 |
| dense_22 (Dense) | (None, 1) | 11 |

Total params: 1,475 (5.77 KB)

Trainable params: 491 (1.92 KB)

Non-trainable params: 0 (0.00 B)

Optimizer params: 984 (3.85 KB)

Figure 32 Model summary for LSTM weekly positive malaria cases

4.18.1 Model summary for Lstm for positive malaria cases

Model: "sequential_26"

| Layer (type) | Output Shape | Param # |
|----------------------|---------------|---------|
| lstm_33 (LSTM) | (None, 3, 50) | 10,400 |
| dropout_14 (Dropout) | (None, 3, 50) | 0 |
| lstm_34 (LSTM) | (None, 50) | 20,200 |
| dropout_15 (Dropout) | (None, 50) | 0 |
| dense_26 (Dense) | (None, 1) | 51 |

Total params: 91,955 (359.20 KB)

Trainable params: 30,651 (119.73 KB)

Non-trainable params: 0 (0.00 B)

Optimizer params: 61,304 (239.47 KB)

Figure 33 Model summary for weekly positive malaria cases

4.19 Validation metrics

| | MAPE | MAE | MSE |
|----------------------------------|-------------------|---------|----------|
| Lstm for positive malaria cases | 259759251806283.3 | 35.680 | 3787.530 |
| Lstm for suspected malaria cases | 0.999 | 224.497 | 77.357 |

Table 3 Model validation for Long short-term memory results

Table 4 shows the results of validation metric for LSTM model

For weekly suspected malaria cases the Mean Squared Error (MSE) of 77,357.87 indicates the model is making predictions with relatively small average squared errors. This indicates that the model is effectively capturing the pattern. The mean absolute error (MAE) of 224.50 provides further confirmation - the average absolute difference between projected and actual results is around 224 suspected malaria cases. This is a fairly low error margin, especially for a public health metric like this. The mean absolute percentage error (MAPE) of 0.999 (or 99.99%) is exceptionally low. This implies the model's percentage-based prediction errors are very small, meaning it is making highly accurate forecasts relative to the true suspected malaria case counts. The Mean Squared Error (MSE) of 3,787.53 suggests the model is making predictions with fairly large average squared errors. This points to substantial differences between the predicted and actual positive malaria case counts.

4.20 Forecast

4.20.1 Forecasted values for weekly suspected malaria cases

Forecasted Weekly Suspected Malaria Cases:

Week 1: 233.16
Week 2: 186.73
Week 3: 85.74
Week 4: 37.82
Week 5: 68.46
Week 6: 188.29
Week 7: 132.83
Week 8: 68.66
Week 9: 68.51
Week 10: 77.33
Week 11: 119.77
Week 12: 85.43
Week 13: 65.43
Week 14: 68.23
Week 15: 94.13
Week 16: 97.11
Week 17: 77.17
Week 18: 67.93
Week 19: 79.88
Week 20: 92.97
Week 21: 87.32
Week 22: 73.12
Week 23: 74.58
Week 24: 84.51
Week 25: 89.94
Week 26: 88.85
Week 27: 74.83
Week 28: 78.87
Week 29: 86.66
Week 30: 84.56

Figure 34 Forecast for suspected malaria cases for 30 weeks

Fig 34 shows the forecasting result for long-short memory for the first 30 weeks of 2024

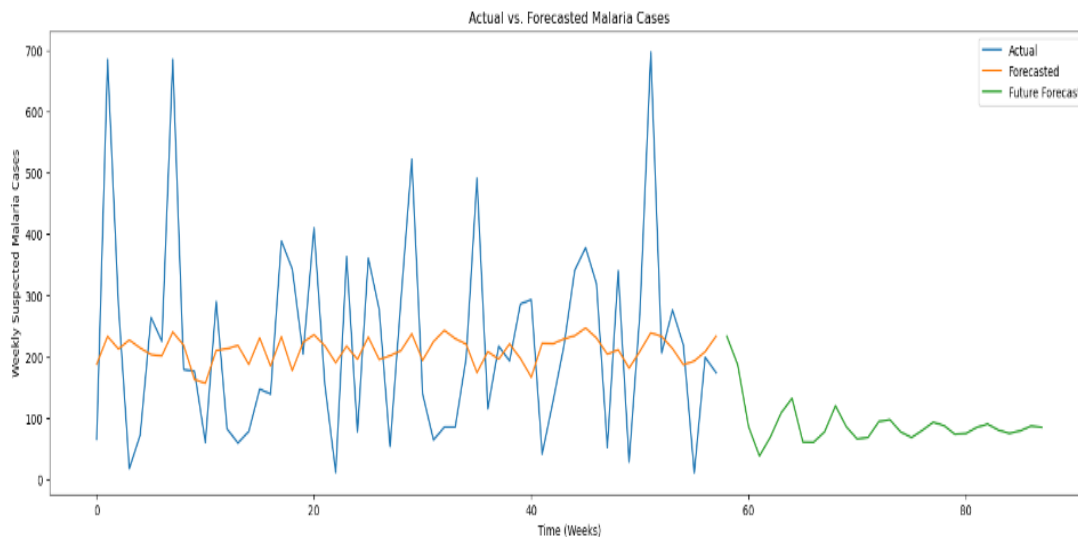


Figure 35 Graph for weekly suspected malaria forecast lstm

Figure 35 shows the actual weekly suspected malaria cases, forecasted malaria and future forecast made by the long short-term memory

4.21 Forecast for weekly positive malaria cases

```

Forecasted weekly positive malaria cases for week 1: 36.67
1/1 ----- 0s 37ms/step
Forecasted weekly positive malaria cases for week 2: 19.47
1/1 ----- 0s 28ms/step
Forecasted weekly positive malaria cases for week 3: 7.05
1/1 ----- 0s 30ms/step
Forecasted weekly positive malaria cases for week 4: 2.15
1/1 ----- 0s 44ms/step
Forecasted weekly positive malaria cases for week 5: 3.82
1/1 ----- 0s 46ms/step
Forecasted weekly positive malaria cases for week 6: 6.99
1/1 ----- 0s 36ms/step
Forecasted weekly positive malaria cases for week 7: 7.60
1/1 ----- 0s 29ms/step
Forecasted weekly positive malaria cases for week 8: 3.45
1/1 ----- 0s 38ms/step
Forecasted weekly positive malaria cases for week 9: 2.66
1/1 ----- 0s 33ms/step
Forecasted weekly positive malaria cases for week 10: 5.84
1/1 ----- 0s 36ms/step
Forecasted weekly positive malaria cases for week 11: 7.72
1/1 ----- 0s 33ms/step
Forecasted weekly positive malaria cases for week 12: 5.09
1/1 ----- 0s 36ms/step
Forecasted weekly positive malaria cases for week 13: 2.37
1/1 ----- 0s 16ms/step
Forecasted weekly positive malaria cases for week 14: 4.55
1/1 ----- 0s 16ms/step
Forecasted weekly positive malaria cases for week 15: 7.26
1/1 ----- 0s 23ms/step
Forecasted weekly positive malaria cases for week 16: 6.46
1/1 ----- 0s 29ms/step
Forecasted weekly positive malaria cases for week 17: 2.92
1/1 ----- 0s 33ms/step
Forecasted weekly positive malaria cases for week 18: 3.37
1/1 ----- 0s 35ms/step
Forecasted weekly positive malaria cases for week 19: 6.36
1/1 ----- 0s 33ms/step
Forecasted weekly positive malaria cases for week 20: 7.17
1/1 ----- 0s 33ms/step
Forecasted weekly positive malaria cases for week 21: 4.18
1/1 ----- 0s 34ms/step
Forecasted weekly positive malaria cases for week 22: 2.67
1/1 ----- 0s 31ms/step
Forecasted weekly positive malaria cases for week 23: 5.27
1/1 ----- 0s 28ms/step
Forecasted weekly positive malaria cases for week 24: 7.23
1/1 ----- 0s 16ms/step
Forecasted weekly positive malaria cases for week 25: 5.48
1/1 ----- 0s 27ms/step
Forecasted weekly positive malaria cases for week 26: 2.68
1/1 ----- 0s 26ms/step
Forecasted weekly positive malaria cases for week 27: 4.12
1/1 ----- 0s 44ms/step
Forecasted weekly positive malaria cases for week 28: 6.74
1/1 ----- 0s 37ms/step
Forecasted weekly positive malaria cases for week 29: 6.53
1/1 ----- 0s 38ms/step
Forecasted weekly positive malaria cases for week 30: 3.38
    
```

Figure 36 weekly positive malaria cases forecast

Fig 36 shows the forecast results for the first 30 week of 2024 of Long short term model

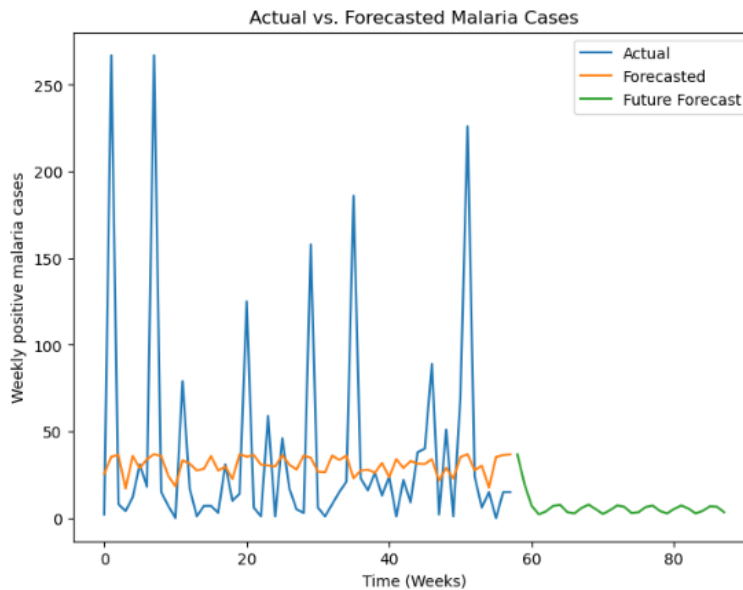


Figure 37 Graph for weekly positive malaria forecast

Figure 37 shows the actual weekly positive malaria cases, forecasted malaria and future forecast made by the long short-term memory.

4.22 Summary

This chapter presents the results of the data analysis and modelling process which aimed to identify the most appropriate time-series models for forecasting malaria cases in Harare province. Through a rigorous diagnostics testing process. The developed models were validated for their accuracy and reliability. The models were then used to forecast future the number of weekly positive malaria cases and weekly suspected malaria cases for the next 30 weeks of 2024. Chapter 5 below presence the research findings in detail

CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a summary of the key findings from the study, along with recommendations for future research and a conclusion. The study aimed to forecast the dynamics

of Malaria cases in Harare Province using time series models, including SARIMAX and deep learning approach

5.2 Summary of Research Findings

For suspected malaria cases the LSTM model achieved excellent performance, with an MSE of 77,357.87, MAE of 224.50, and an exceptionally low MAPE of 0.999 (99.99%). These metrics indicate the LSTM model is making highly accurate, low-error predictions for suspected malaria cases, capturing the underlying patterns in the data very well. The model would likely be very useful for real-world malaria surveillance and planning purposes given its strong predictive capabilities for suspected cases. For weekly positive malaria cases the LSTM model struggled significantly with predicting positive malaria cases, with an MSE of 3,787.53 and a catastrophically high MAPE of over 259 trillion percent. The results suggested the LSTM model has fundamental issues accurately forecasting positive malaria cases, likely due to shortcomings in the model architecture, feature engineering, or training data quality. The model would not be suitable for real-world applications targeting positive malaria case prediction in its current state. For the SARIMAX model performance for weekly suspected malaria case prediction, the SARIMAX (3,1,3)(0,0,0,13) model achieved a MAPE of 20%, indicating reasonably accurate percentage-based forecasts with a moderate level of error. The MSE of 275.39 and MAE of 12.12 for the SARIMAX model suggest it is making predictions with a larger degree of absolute error compared to the LSTM model's performance on suspected cases. For positive malaria case prediction, the SARIMAX (6,0,6)(0,0,6,13) model had an undefined MAPE due to issues with zero/near-zero values, but a more reasonable MSE of 10.45 and MAE of 2.54. The LSTM model appears to be the superior performer for suspected malaria case prediction, while both the LSTM and SARIMAX models struggle to some degree with accurately forecasting positive malaria cases. Further model refinement and exploration of alternative approaches may be necessary to improve positive case prediction capabilities.

The research also noted a decrease in the general trend for malaria cases in Harare provinces

5.3 Recommendations

The provincial health authorities should maintain a robust system for monitoring and recording Malaria cases, as this data is crucial for accurate forecasting and effective intervention planning. Future research could investigate the potential benefits of combining SARIMAX and LSTM models, or other time series techniques, to further improve the accuracy and reliability of Malaria case forecasts. The inclusion of other relevant variables, such as socioeconomic factors,

vegetation index and public health interventions routines data, to enhance the predictive power of the time series models.

5.4 Conclusion

This study has demonstrated the effectiveness of time series models, particularly the LSTM approach, in forecasting the dynamics of Malaria cases in Harare Province. The findings provide valuable insights for public health authorities and policymakers in their efforts to monitor, predict, and respond to the Malaria burden in Harare Province. The recommendations outlined in this chapter suggest a background for further research and improvements to the forecasting tools, ultimately supporting the goal of reducing the impact of Malaria in Zimbabwe.

References

- (, J. P. D., 202). Evaluation of prediction models for the malaria incidence in Marodijeh Region. *Somaliland* 46(2):395–408.
- (WHO), W. H. O., 2020). World Malaria Report 2020. Geneva, Switzerland: WHO.
- Aaltio, I., & Heilmann, P., 2013. *Data Collection. In Handbook of Qualitative Research Methods on Human Resource Management: Innovative Techniques.* s.l.:Edward Elgar Publishing.
- Adeola, A. M. e. a., (2019. "Predicting malaria cases using remotely sensed environmental variables in Nkomazi, South Africa.. *Geospatial Health* 14.1 .

- Ali, M. A. & K. M., 2023. Effective Strategies for Crafting Research Proposals in Higher Education.. *International Journal of Business and Management Research*, 11(4), 107-120..
- Anon., 1997. "On Airs, Waters, and Places" (Peri Aerōn, Hydatōn, Topōn) Article.
- Anon., n.d. s.l.:s.n.
- Brockwell, P. J. & D. R. A., 2016. *Introduction to Time Series and Forecasting*. Springer. s.l.:s.n.
- Brown, B. & J. M., 2016. Principals' technology leadership: How a conceptual framework shaped a mixed methods study.. *Journal of School Leadership*, 26(5), 811-836..
- Caimo, M. C. A., n.d. *Time series Analysis [Research Report] ESPON Inspire Public policy marking with territorial Evidence 2012 bal 03609303*, s.l.: s.n.
- Chennai SS, D. Z., 2018. Development of artificial intelligence approach to forecasting oyster norovirus outbreaks along Gulf of Mexico coast.. *Environ Int* 111:212–223..
- Creswell, J. W., 2014. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.. s.l.:s.n.
- Dabral, P. P. a. M. Z. M., 2017. Modelling and forecasting of rainfall time series using SARIMA. *Environmental Processes* .
- Elvis Adam Alhassan I, A. M. I. A. E., 2017. Time Series Analysis of Malaria Cases in Kasena Nankana Municipality International Journal of Statistics and Applications ,. 43-56DOI: 10.5923/j.statistics.20170702.01.
- Gondwe, T. et al., 2021. Epidemiological Trends of Malaria in Five Years and under Children of Nsanje District in Malawi, 2015–2019. *Int. J. Environ. Res. Public Health* .
- Gondwe, T. Y. Y. Y. S. K. M. M. G. L. M. P. .. & M. T., 2021. Epidemiological trends of malaria in five years and under children of Nsanje district in Malawi, 2015–2019.. *International journal of environmental researc*.
- Gondwe, T. Y. Y. Y. S. K. M. M. G. L. M. P. .. & M. T., 2021. Epidemiological trends of malaria in five years and under children of Nsanje district in Malawi, 2015–2019..
- Gondwe, T. Y. Y. Y. S. K. M. M. G. L. M. P. .. & M. T., International journal of environmental researc. Epidemiological trends of malaria in five years and under children of Nsanje district in Malawi,. 2021.
- Goodfellow, I., Bengio, Y., & Courville, A., 2016. *Deep learning*. Cambridge: MIT press.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E., 2017. *Multivariate data analysis (8th ed.)*. Cengage Learning. s.l.:s.n.
- Hassani, H. S. E. S. A. N. F. G. & G. R., 2017. Forecasting accuracy evaluation of tourist arrivals... *Annals of Tourism Research*, .
- Hochreiter, S. & S. J., 1997. Long short-term memory. *Neural Computation*.
- Hyndman RJ, K. A., 2006. Another look at measures of forecast accuracy.. *Int J Forecast* 22:679–688.
- J. S. Adeyeye1, E. B. N., 2023. Predicting Malaria Incident Using Hybrid SARIMA-LSTM.Model Vol. 9, No. 1,.
- Kelleher, C., Mac Namee, B., & D'Arcy, A , 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press. s.l.:s.n.
- L), C., 2017. , *Instruments for gathering data*. In E Moore & Dooly (Eds) ,*Qualitative approaches to research on plurilingual education*. s.l.:Research publishing net.
- Makridakis S, W. S. H. R., 2008. *Forecasting methods and applications*. Wiley,. New York: s.n.
- Makridakis, S. H. M., n.d. ARMA models and the Box-Jenkins methodology. *journal of Forecasting*. 1997.

- Mills, T. C., 2019. *Applied time series analysis: A practical guide to modeling and forecasting*. Academic press. s.l.:s.n.
- Mohamed, J. M. A. I. & D. E. I., 2022. *Evaluation of prediction models for the malaria incidence in Marodijeh Region*,. Somaliland. *Journal of Parasitic Diseases*, 1-14.: s.n.
- Nwakuwa Esther Promise, N., 2022. Time series analysis of typhoid fever incidence in Nigeria using SARIMA and LSTM models..
- Orjuela canoon AD, J. A. D. G. M. A. G. C. V. E. P. M., 2022. Time series forecasting for tuberculosis incidence neural network models networks models..
- Papioannou GP, D. C. D. A., 2016. Analysis and modelling for short term to medium term load . Forecasting using a Hybrid Learning Principal Component and comparison with classical statistical models SARIMAX exponential smoothing and artificial. *The case of greek electricity market energies* .
- Peixeiro, M., 2022. <https://livebook.manning.com/book/time-series-forecasting-in-python-book/>. s.l.:s.n.
- Sakubu, D. S. K. J. G. & N. D., 2023. Predicting malaria dynamics in Burundi using deep Learning Models. .. *arXiv preprint arXiv:2306.02685*.
- Saunders, M., Lewis, P., & Thornhill, A., 2019. *Research methods for business students*. Pearson Education Limited.
- Shumway, R. H. & S. D. S., 2017. *Time Series Analysis and Its Applications: With R Examples (4th ed.)*. Springer. s.l.:s.n.
- Somyanothanaul, R. W. K. A. W., 2022. Forecasting COVID-19 cases using time series modelling and association rule mining. *BMC Med Res Methodology* 22, 281.
- Taiwo A. I, O. T. O. A. A. F. a. A. K. K., 2019. Modeling and Forecasting Periodic Time Series data with Fourier Autoregressive Model.. *Iraqi Journal of Science*.
- Tatem, A. J. G. P. W. S. D. L. & H. S. I., 2013. *Urbanisation and the global malaria recession*. *Nature*,. s.l.:s.n.
- Thomas Schincariol, E. R. S. J. T. C. F. G., 2021. Forecasting cross border malaria cases number : towards an early warning system to support malaria elimination plans. *7th international conference on Timeseries forecasting*.
- Trirat, P. S. Y. K. J. N. Y. N. J. B. M. .. & L. J. G., 2024. Universal Time-Series Representation Learning: A Survey..
- Wang M, W. H. W. J. L. H. L. R. D. T. e. a., 2019. A novel model for malaria prediction based on ensemble algorithms.. *PLoS ONE* 14(12): e0226910.
- Wangdi, 2010. Development of temporal modeling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhuta. *Malaria Journal* .
- Wiebe, M. A. J. D. G. a., 2009. Sage Encyclopedia of case study Research.
- Zinszer, K., 2014. Predicting malaria in a highly endemic country using clinical and environmental data..
- Zuo, J., 2022). Representation learning and forecasting for inter-related time series.

APPENDIX 1

Jupyternote book python codes

```
from sklearn.metrics import mean_squared_error, mean_absolute_error
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.seasonal import seasonal_decompose, STL
from statsmodels.stats.diagnostic import acorr_ljungbox
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.tsa.arima_process import ArmaProcess
from statsmodels.graphics.gofplots import qqplot
from statsmodels.tsa.stattools import adfuller
from tqdm import tqdm_notebook
from itertools import product
```

```

from typing import Union
import matplotlib.pyplot as plt
import statsmodels.api as sm
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

import pandas as pd
df = pd.read_excel ("C:\\Users\\hp\\documents\\malaria2015.xlsx")
df.head ()
df.head()
df.info()

df['date']= pd.Timestamp('2013-01-01') + pd.to_timedelta(df['week'].astype(str) + 'W')
df.set_index('date', inplace= True)
print(df.index.dtype)
print(df.head())
print(df.index.dtype)
df.index.inferred_freq
df.index.freq= df.index.inferred_freq

# Print the frequency
print(df.index.freq)

import pandas as pd

# Assuming 'df' is your DataFrame with the malaria cases data
# Assuming 'malaria_cases' is the column representing the malaria cases

# Calculate descriptive statistics
statistics = df['weekly suspected malaria'].describe()

# Print the descriptive statistics
print(statistics)

import pandas as pd

# Assuming 'df' is your DataFrame with the malaria cases data
# Assuming 'malaria_cases' is the column representing the malaria cases

# Calculate descriptive statistics
statistics = df['weekly positive malaria cases'].describe()

# Print the descriptive statistics

```

```
print(statistics)

import matplotlib.pyplot as plt

df['weekly suspected malaria'].plot()

plt.xlabel('Year-Week')
plt.ylabel('Weekly Suspected Malaria')
plt.title('Weekly Suspected Malaria Cases')

plt.xticks(rotation=45)
plt.tight_layout()

plt.show()
```

```
import matplotlib.pyplot as plt

df['weekly suspected malaria'].plot()

plt.xlabel('Year-Week')
plt.ylabel('Weekly Suspected Malaria')
plt.title('Weekly Suspected Malaria Cases')

plt.xticks(rotation=45)
plt.tight_layout()

plt.show()
```

```
import matplotlib.pyplot as plt

df['weekly positive malaria cases'].plot(color='green')

plt.xlabel('Year-Week')
plt.ylabel('weekly positive malaria cases')
plt.title('weekly positive malaria cases')

plt.xticks(rotation=45)
plt.tight_layout()

plt.show()
```

```
import matplotlib.pyplot as plt

# Set the figure size
```

```

plt.figure(figsize=(10, 6))

# Calculate the rolling mean and rolling standard deviation
rolling_mean = df['weekly suspected malaria'].rolling(window=4).mean()
rolling_std = df['weekly suspected malaria'].rolling(window=4).std()

# Plot the original data, rolling mean, and rolling standard deviation
plt.plot(df.index, df['weekly suspected malaria'], label='Original')
plt.plot(df.index, rolling_mean, label='Rolling Mean')
plt.plot(df.index, rolling_std, label='Rolling Std')

plt.xlabel('Year-Week')
plt.ylabel('Weekly Suspected Malaria')
plt.title('Weekly Suspected Malaria with Rolling Statistics')
plt.legend()

# Reduce the number of visible x-axis tick labels
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)

plt.tight_layout()
plt.show()

import matplotlib.pyplot as plt

# Set the figure size
plt.figure(figsize=(10, 6))

# Calculate the rolling mean and rolling standard deviation
rolling_mean = df['weekly positive malaria cases'].rolling(window=4).mean()
rolling_std = df['weekly positive malaria cases'].rolling(window=4).std()

# Plot the original data, rolling mean, and rolling standard deviation
plt.plot(df.index, df['weekly positive malaria cases'], label='Original')
plt.plot(df.index, rolling_mean, label='Rolling Mean')
plt.plot(df.index, rolling_std, label='Rolling Std')

plt.xlabel('Year-Week')
plt.ylabel('weekly positive malaria cases')
plt.title('weekly positive malaria cases with Rolling Statistics')
plt.legend()

# Reduce the number of visible x-axis tick labels
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)

```

```

plt.tight_layout()
plt.show()

import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
# Decompose the time series into trend, seasonality, and residuals
decomp = sm.tsa.seasonal_decompose(df['weekly suspected malaria'], model='additive',
period=13)

# Extract the trend, seasonality, and residuals
trend = decomp.trend
seasonal = decomp.seasonal
residual = decomp.resid

# Plot the decomposition
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(df['weekly suspected malaria'], label='Original', color='blue')
plt.legend(loc='best')
plt.title('Decomposition Plot')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(412)
plt.plot(trend, label='Trend', color='blue')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(413)
plt.plot(seasonal, label='Seasonality', color='blue')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(414)
plt.plot(residual, label='Residuals', color='blue')
plt.legend(loc='best')
plt.tight_layout()
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)

plt.show()

# Decompose the time series into trend, seasonality, and residuals
decomp = sm.tsa.seasonal_decompose(df['weekly positive malaria cases'], model='additive',
period=13)

```

```

# Extract the trend, seasonality, and residuals
trend = decomp.trend
seasonal = decomp.seasonal
residual = decomp.resid

# Plot the decomposition
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(df['weekly positive malaria cases'], label='Original', color='green')
plt.legend(loc='best')
plt.title('Decomposition Plot')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(412)
plt.plot(trend, label='Trend', color='green')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(413)
plt.plot(seasonal, label='Seasonality', color='green')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)
plt.subplot(414)
plt.plot(residual, label='Residuals', color='green')
plt.legend(loc='best')
plt.tight_layout()
num_ticks = 8
plt.xticks(df.index[::len(df.index)//num_ticks], rotation=45)

plt.show()

from statsmodels.tsa.stattools import adfuller

# Perform ADF test on suspected malaria cases data
adf_result = adfuller(df['weekly suspected malaria'])

# Extract and print the test statistics and p-value
test_statistic = adf_result[0]
p_value = adf_result[1]
print(f'ADF Test Statistic: {test_statistic}')
print(f'p-value: {p_value}')

# Interpret the test results
if p_value < 0.05:
    print("The data is stationary (reject the null hypothesis)")

```

```

else:
    print("The data is non-stationary (fail to reject the null hypothesis)")

from statsmodels.tsa.stattools import adfuller

# Perform ADF test on suspected malaria cases data
adf_result = adfuller(df['weekly positive malaria cases'])

# Extract and print the test statistics and p-value
test_statistic = adf_result[0]
p_value = adf_result[1]
print(f'ADF Test Statistic: {test_statistic}')
print(f'p-value: {p_value}')

# Interpret the test results
if p_value < 0.05:
    print("The data is stationary (reject the null hypothesis)")
else:
    print("The data is non-stationary (fail to reject the null hypothesis)")

import pandas as pd
import numpy as np
from statsmodels.tsa.stattools import adfuller

# Assuming df is your DataFrame

# Perform differencing on the column
df['diff_weekly_suspected_malaria'] = df['weekly suspected malaria'].diff()

# Drop rows with missing values in 'diff_weekly_suspected_malaria'
df.dropna(subset=['diff_weekly_suspected_malaria'], inplace=True)

# Apply the ADF test to check for stationarity
result = adfuller(df['diff_weekly_suspected_malaria'].dropna())

# Print the ADF test results
print('ADF Statistic:', result[0])
print('p-value:', result[1])
print('Critical Values:')
for key, value in result[4].items():
    print(f'{key}: {value}')
    # Interpret the test results
if p_value < 0.05:
    print("The data is stationary (reject the null hypothesis)")
else:
    print("The data is non-stationary (fail to reject the null hypothesis)")

```

```

from statsmodels.tsa.stattools import adfuller

# Extract the exogenous variables from the DataFrame
df.columns= df.columns.str.lstrip()
exog_variables = df[['weekly rainfall Received', 'average weekly max temp', 'average weekly
min temp']]
# Remove leading spaces from column names
exog_variables.columns = exog_variables.columns.str.lstrip()
# Perform ADF test for each exogenous variable
for column in exog_variables:
    result = adfuller(exog_variables[column])
    print(f'ADF Test for {column}:')
    print(f'ADF Statistic: {result[0]}')
    print(f'p-value: {result[1]}')
    print(f'Critical Values:')
    for key, value in result[4].items():
        print(f' {key}: {value}')
    print("-----")
    # Interpret the test results
if p_value < 0.05:
    print("The data is stationary (reject the null hypothesis)")
else:
    print("The data is non-stationary (fail to reject the null hypothesis)")

import pandas as pd
df.columns = df.columns.str.strip()
# Assuming you have already read the data into a DataFrame df

# Perform differencing on the exogenous factors
df['diff_weekly_rainfall'] = df['weekly rainfall Received'].diff()
df['diff_average_max_temp'] = df['average weekly max temp'].diff()
df['diff_average_min_temp'] = df['average weekly min temp'].diff()

# Perform differencing on the suspected malaria cases
df['diff_suspected_malaria_cases'] = df['weekly suspected malaria'].diff()

# Remove the first row with NaN values
df = df.dropna()

# Print the head of the DataFrame to check the differenced data
print(df.head())

import pandas as pd
import statsmodels.api as sm

```



```

import matplotlib.pyplot as plt

# assuming 'df' is your pandas DataFrame
# and 'weekly_suspected_malaria' is the column with the weekly suspected malaria cases

# calculate the difference
df['diff_suspected_malaria_cases'] = df['weekly suspected malaria'].diff()
# Remove the first row with NaN values
df = df.dropna()
# convert the index to a Datetimework

df.index.freq = 'W-TUE'
# STL decomposition
decomposition = sm.tsa.seasonal_decompose(df['diff_suspected_malaria_cases'],
model='additive')

# plot the decomposition
fig, ax = plt.subplots(4, 1, sharex=True, figsize=(12, 8))
ax[0].plot(df['diff_suspected_malaria_cases'], label='Differenced')
ax[0].legend(loc='best')
ax[1].plot(decomposition.trend, label='Trend')
ax[1].legend(loc='best')
ax[2].plot(decomposition.seasonal, label='Seasonality')
ax[2].legend(loc='best')
ax[3].plot(decomposition.resid, label='Residuals')
ax[3].legend(loc='best')
plt.show()

# Decompose the time series into trend, seasonality, and residuals
decomp = sm.tsa.seasonal_decompose(df['weekly positive malaria cases'], model='additive',
period=13)

# Extract the trend, seasonality, and residuals
trend = decomp.trend
seasonal = decomp.seasonal
residual = decomp.resid

# Plot the decomposition
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(df['weekly positive malaria cases'], label='Original', color='green')
plt.legend(loc='best')
plt.title('decomposition for weekly positive malaria cas')
num_ticks = 8
plt.xticks(df.index[:len(df.index)//num_ticks], rotation=45)

```

```

plt.subplot(412)
plt.plot(trend, label='Trend', color='green')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[:,len(df.index)//num_ticks], rotation=45)
plt.subplot(413)
plt.plot(seasonal, label='Seasonality', color='green')
plt.legend(loc='best')
num_ticks = 8
plt.xticks(df.index[:,len(df.index)//num_ticks], rotation=45)
plt.subplot(414)
plt.plot(residual, label='Residuals', color='green')
plt.legend(loc='best')
plt.tight_layout()
num_ticks = 8
plt.xticks(df.index[:,len(df.index)//num_ticks], rotation=45)

```

```
plt.show()
```

```

import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from itertools import product
# Plot ACF and PACF
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 6))
plot_acf(df['diff_suspected_malaria_cases'], lags=20, ax=ax1)
plot_pacf(df['diff_suspected_malaria_cases'], lags=20, ax=ax2)
plt.show()

```

```

import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from itertools import product
# Plot ACF and PACF
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 6))
plot_acf(df['weekly positive malaria cases'], lags=20, ax=ax1)
plot_pacf(df['weekly positive malaria cases'], lags=20, ax=ax2)
plt.show()

```

```

import pandas as pd
from statsmodels.tsa.statespace.sarimax import SARIMAX
from tqdm import tqdm_notebook
from itertools import product

```

```

import numpy as np
#Load the dataset
df = pd.read_excel("C:\\Users\\hp\\documents\\malaria2015.xlsx")
df.columns = df.columns.str.strip()
# Handle infinite or missing values
df.dropna(inplace=True)
df.replace([np.inf, -np.inf], 0, inplace=True)

#Prepare the dataset
target = df['weekly suspected malaria']
exog = df[['weekly rainfall Received', 'average weekly max temp', 'average weekly min temp']]

#Define the range of possible values for the orders p, q, P, and Q*
p = range(0, 6, 3)
d = 1
q = range(0, 6, 3)
P = range(0, 6, 3)
D = 0
Q = range(0, 6, 3)
s = 13 # Since the data is collected weekly, s = 13

#Generate a list of unique combinations of parameters
parameters = product(p, q, P, Q)
parameters_list = list(parameters)

#Train the model using the first 400 instances
target_train = target[:565]
exog_train = exog[:565]

def optimize_SARIMAX(endog, exog, order_list, d, D, s):
    results = []
    for order in tqdm_notebook(order_list):
        try:
            model = SARIMAX(
                endog,
                exog,
                order=(order[0], d, order[1]),
                seasonal_order=(order[2], D, order[3], s),
                simple_differencing=False
            ).fit(dispatch=False)
        except:
            continue
        aic = model.aic
        results.append([order, aic])
    result_df = pd.DataFrame(results)
    result_df.columns = ['(p,q,P,Q)', 'AIC']

```

```

# Sort in ascending order, lower AIC is better
result_df = result_df.sort_values(by='AIC', ascending=True).reset_index(drop=True)
return result_df
#Run the optimize_SARIMAX function and select the model with the lowest AIC
result_df = optimize_SARIMAX(target_train, exog_train, parameters_list, d, D, s)
print(result_df)

import pandas as pd
from statsmodels.tsa.statespace.sarimax import SARIMAX
from tqdm import tqdm_notebook
from itertools import product
import numpy as np
model= SARIMAX(target_train, exog_train, order= (3, 1, 3), seasonal_order=(0, 0, 0, 13),
simple_differencing= False)
model_fit= model.fit(dispatch= False)
print(model_fit.summary())
model_fit.plot_diagnostics(figsize=(10,8))

residuals = model_fit.resid
lbvalue, pvalue = acorr_ljungbox(residuals, np.arange(1, 11, 1))
print(residuals)

def rolling_forecast(endog: Union[pd.Series, list], exog:
Union[pd.Series, list], train_len: int, horizon: int, window: int,
method: str) -> list:
    total_len = train_len + horizon
    if method == 'last':
        pred_last_value = []
        for i in range(train_len, total_len, window):
            last_value = endog[:i].iloc[-1]
            pred_last_value.extend(last_value for _ in range(window))
        return pred_last_value
    elif method == 'SARIMAX':
        pred_SARIMAX = []
        for i in range(train_len, total_len, window):
            model = SARIMAX(endog[:i], exog[:i], order=(3,1,3),
seasonal_order=(0,0,0,13), simple_differencing=False)
            res = model.fit(dispatch=False)
            predictions = res.get_prediction(exog=exog)
            oos_pred = predictions.predicted_mean.iloc[-window:]
            pred_SARIMAX.extend(oos_pred)
        return pred_SARIMAX
    print(pred_SARIMAX)

target_train = target[:561]

```

```

target_test = target[561:]
pred_df = pd.DataFrame({'actual': target_test})

TRAIN_LEN = len(target_train)
HORIZON = len(target_test)
WINDOW = 1

pred_last_value = rolling_forecast(target, exog, TRAIN_LEN, HORIZON, WINDOW, 'last')
pred_SARIMAX = rolling_forecast(target, exog, TRAIN_LEN, HORIZON, WINDOW,
'SARIMAX')

pred_df['pred_last_value'] = pred_last_value
pred_df['pred_SARIMAX'] = pred_SARIMAX

print(pred_df)

import numpy as np
import pandas as pd
from statsmodels.tsa.statespace.sarimaxSARIMAX import SARIMAX

# Assuming you have the following data:
data = pd.DataFrame({
    'actual': [68, 51, 67, 56, 61, 85, 58, 57, 79, 64, 49],
    'pred_last_value': [52, 68, 51, 67, 56, 61, 85, 58, 57, 79, 64],
    'pred_SARIMAX': [68.818557, 60.106570, 68.488238, 61.254996, 60.7944702, 64.238223,
53.541036, 85.416874, 63.703904, 53.326564, 81.366124]
})

# Fit the SARIMAX model
model = SARIMAX(data['actual'], order=(3, 1, 3), seasonal_order=(1, 0, 2, 13))
results = model.fit()

# Generate 30-step forecast
forecast = results.get_forecast(steps=30)
forecast_df = forecast.conf_int().join(forecast.predicted_mean)

# Create a new DataFrame to store the actual, last value, and SARIMAX predictions
future_data = pd.DataFrame({
    'actual': [np.nan] * 30,
    'pred_last_value': [data['actual'].iloc[-1]] * 30,
    'pred_SARIMAX': forecast_df['predicted_mean']
})

# Combine the original data and the future data
all_data = pd.concat([data, future_data])

```

```

print(all_data)

from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_absolute_percentage_error
import numpy as np

# Calculate MAPE
pred_df['mape'] = np.abs((pred_df['actual'] - pred_df['pred_last_value']) / pred_df['actual']) * 100
mape_last_value = pred_df['mape'].mean()

pred_df['mape'] = np.abs((pred_df['actual'] - pred_df['pred_SARIMAX']) / pred_df['actual']) *
100
mape_SARIMAX = pred_df['mape'].mean()

print(f'MAPE for last value prediction: {mape_last_value:.2f}%')
print(f'MAPE for SARIMAX prediction: {mape_SARIMAX:.2f}%')

# Calculate MSE
mse_last_value = np.mean((pred_df['actual'] - pred_df['pred_last_value'])**2)
mse_SARIMAX = np.mean((pred_df['actual'] - pred_df['pred_SARIMAX'])**2)

print(f'MSE for last value prediction: {mse_last_value:.2f}')
print(f'MSE for SARIMAX prediction: {mse_SARIMAX:.2f}')

# Calculate MAE
mae_last_value = np.mean(np.abs(pred_df['actual'] - pred_df['pred_last_value']))
mae_SARIMAX = np.mean(np.abs(pred_df['actual'] - pred_df['pred_SARIMAX']))

print(f'MAE for last value prediction: {mae_last_value:.2f}')
print(f'MAE for SARIMAX prediction: {mae_SARIMAX:.2f}')

import pandas as pd
from statsmodels.tsa.statespace.sarimax import SARIMAX
from tqdm import tqdm_notebook
from itertools import product
import numpy as np
#Load the dataset
df = pd.read_excel("C:\\Users\\hp\\documents\\malaria2015.xlsx")
df.columns = df.columns.str.strip()
# Handle infinite or missing values
df.dropna(inplace=True)
df.replace([np.inf, -np.inf], 0, inplace=True)

#Prepare the dataset

```

```

target = df['weekly positive malaria cases']
exog = df[['weekly rainfall Received', 'average weekly max temp', 'average weekly min temp']]

#Define the range of possible values for the orders p, q, P, and Q*
p = range(0, 10, 6)
d = 0
q = range(0, 10, 6)
P = range(0, 10, 6)
D = 0
Q = range(0, 10, 6)
s = 13 # Since the data is collected weekly, s = 13

#Generate a list of unique combinations of parameters
parameters = product(p, q, P, Q)
parameters_list = list(parameters)

#Train the model using the first 400 instances
target_train = target[:565]
exog_train = exog[:565]

def optimize_SARIMAX(endog, exog, order_list, d, D, s):
    results = []
    for order in tqdm_notebook(order_list):
        try:
            model = SARIMAX(
                endog,
                exog,
                order=(order[0], d, order[1]),
                seasonal_order=(order[2], D, order[3], s),
                simple_differencing=False
            ).fit(dispatch=False)
        except:
            continue
        aic = model.aic
        results.append([order, aic])
    result_df = pd.DataFrame(results)
    result_df.columns = ['(p,q,P,Q)', 'AIC']
    # Sort in ascending order, lower AIC is better
    result_df = result_df.sort_values(by='AIC', ascending=True).reset_index(drop=True)
    return result_df

#Run the optimize_SARIMAX function and select the model with the lowest AIC
result_df = optimize_SARIMAX(target_train, exog_train, parameters_list, d, D, s)
print(result_df)

import pandas as pd
from statsmodels.tsa.statespace.sarimax import SARIMAX

```

```

from tqdm import tqdm_notebook
from itertools import product
import numpy as np
model= SARIMAX(target_train, exog_train, order=(2, 0, 3), seasonal_order=(3, 0, 2, 13),
simple_differencing= False)
model_fit= model.fit(dis= False)
print(model_fit.summary())

model_fit.plot_diagnostics(figsize=(10,8))

residuals = model_fit.resid
lbvalue, pvalue = acorr_ljungbox(residuals, np.arange(1, 11, 1))
print(residuals)

def rolling_forecast(endog: Union[pd.Series, list], exog:
Union[pd.Series, list], train_len: int, horizon: int, window: int,
method: str) -> list:
    total_len = train_len + horizon
    if method == 'last':
        pred_last_value = []
        for i in range(train_len, total_len, window):
            last_value = endog[:i].iloc[-1]
            pred_last_value.extend(last_value for _ in range(window))
        return pred_last_value
    elif method == 'SARIMAX':
        pred_SARIMAX = []
        for i in range(train_len, total_len, window):
            model = SARIMAX(endog[:i], exog[:i], order=(2,0,3),
            seasonal_order=(1,0,1,13), simple_differencing=False)
            res = model.fit(dis=False)
            predictions = res.get_prediction(exog=exog)
            oos_pred = predictions.predicted_mean.iloc[-window:]
            pred_SARIMAX.extend(oos_pred)
        return pred_SARIMAX
    print(pred_SARIMAX)

target_train = target[:561]
target_test = target[561:]
pred_df = pd.DataFrame({'actual': target_test})

TRAIN_LEN = len(target_train)
HORIZON = len(target_test)
WINDOW = 1

pred_last_value = rolling_forecast(target, exog, TRAIN_LEN, HORIZON, WINDOW, 'last')

```



```

pred_SARIMAX = rolling_forecast(target, exog, TRAIN_LEN, HORIZON, WINDOW,
'SARIMAX')

pred_df['pred_last_value'] = pred_last_value
pred_df['pred_SARIMAX'] = pred_SARIMAX

print(pred_df)

from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_absolute_percentage_error
import numpy as np

# Calculate MAPE
pred_df['mape'] = np.abs((pred_df['actual'] - pred_df['pred_last_value']) / pred_df['actual']) * 100
mape_last_value = pred_df['mape'].mean()

pred_df['mape'] = np.abs((pred_df['actual'] - pred_df['pred_SARIMAX']) / pred_df['actual']) *
100
mape_SARIMAX = pred_df['mape'].mean()

print(f'MAPE for last value prediction: {mape_last_value:.2f}%')
print(f'MAPE for SARIMAX prediction: {mape_SARIMAX:.2f}%')

# Calculate MSE
mse_last_value = np.mean((pred_df['actual'] - pred_df['pred_last_value'])**2)
mse_SARIMAX = np.mean((pred_df['actual'] - pred_df['pred_SARIMAX'])**2)

print(f'MSE for last value prediction: {mse_last_value:.2f}')
print(f'MSE for SARIMAX prediction: {mse_SARIMAX:.2f}')

# Calculate MAE
mae_last_value = np.mean(np.abs(pred_df['actual'] - pred_df['pred_last_value']))
mae_SARIMAX = np.mean(np.abs(pred_df['actual'] - pred_df['pred_SARIMAX']))

print(f'MAE for last value prediction: {mae_last_value:.2f}')
print(f'MAE for SARIMAX prediction: {mae_SARIMAX:.2f}')

import numpy as np
import pandas as pd
from statsmodels.tsa.statespace.sarimaxSARIMAX import SARIMAX

# Assuming you have the following data:
data = pd.DataFrame({
    'actual': [2, 2, 0, 1, 1, 0, 2, 6, 5, 1, 1],
    'pred_last_value': [2, 2, 2, 0, 1, 1, 0, 2, 6, 1, 1],

```

```
    'pred_SARIMAX': [-0.359775, 5.141341, 0.809185, 3.094283, -0.408399, 1.487610,
0.263123, 0.629804, 3.257232, 6.751317, 4.032482]
})
```

```
# Fit the SARIMAX model
model = SARIMAX(data['actual'], order=(2, 0, 3), seasonal_order=(3, 0, 2, 13))
results = model.fit()
```

```
# Generate 30-step forecast
forecast = results.get_forecast(steps=30)
forecast_df = forecast.conf_int().join(forecast.predicted_mean)
```

```
# Create a new DataFrame to store the actual, last value, and SARIMAX predictions
future_data = pd.DataFrame({
    'actual': [np.nan] * 30,
    'pred_last_value': [data['actual'].iloc[-1]] * 30,
    'pred_SARIMAX': forecast_df['predicted_mean']
})
```

```
# Combine the original data and the future data
all_data = pd.concat([data, future_data])
```

```
print(all_data)
```

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM
from tensorflow.keras.layers import Dense
# Load the data
lf = pd.read_excel("C:\\Users\\hp\\documents\\malaria2015.xlsx")
lf['date'] = pd.Timestamp('2013-01-01') + pd.to_timedelta(lf['week'].astype(str) + 'W')
lf.set_index('date', inplace=True)
# Strip whitespace from column names
lf.columns = lf.columns.str.strip()
print(lf)
```

```
import numpy as np
import pandas as pd
from keras.models import Sequential
from keras.layers import LSTM, Dense, Input, Dropout
from keras.regularizers import l2
from sklearn.preprocessing import StandardScaler
```

```

from sklearn.model_selection import train_test_split

# Assuming you have the necessary data in a DataFrame 'lf'

# Preprocess the data
scaler = StandardScaler()
X = lf[['weekly rainfall Received', 'average weekly max temp', 'average weekly min temp']]
y = lf['weekly suspected malaria'] # weekly suspected malaria cases

# Normalize the input and output variables
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1))

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.01,
random_state=42)

# Reshape the input data to include the exogenous factors
X_train = X_train.reshape(X_train.shape[0], X_train.shape[1], 1)
X_test = X_test.reshape(X_test.shape[0], X_test.shape[1], 1)

# Train the model
model = Sequential()
model.add(Input(shape=(X_train.shape[1], 1)))
model.add(LSTM(50, return_sequences=True, kernel_regularizer=l2(0.001),
bias_regularizer=l2(0.001)))
model.add(Dropout(0.2)) # Add dropout layer with a 20% dropout rate
model.add(LSTM(50, kernel_regularizer=l2(0.001), bias_regularizer=l2(0.001)))
model.add(Dropout(0.2)) # Add another dropout layer
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
history = model.fit(X_train, y_train, epochs=100, batch_size=3, validation_data=(X_test,
y_test))
# Print the model summary
model.summary()
# Make predictions on the test data
y_pred = model.predict(X_test)

# Inverse scale the predicted values
y_pred = scaler.inverse_transform(y_pred)
y_test = scaler.inverse_transform(y_test)

# Evaluate the model performance
mse = np.mean((y_test - y_pred) ** 2)
print(f'Mean Squared Error: {mse:.2f}')

```

```

# Forecast future values
future_weeks = 30 # Number of future weeks to forecast
future_data = X_test[-1].reshape(1, X_test.shape[1], 1) # Use the last row of X_test as the initial
input
future_predictions = []

for _ in range(future_weeks):
    future_prediction = model.predict(future_data)
    future_prediction = np.clip(scaler.inverse_transform(future_prediction), 0, None) # Clip
negative values to 0
    future_predictions.append(future_prediction[0][0])

    # Update the future_data with the predicted value
    future_data = np.concatenate((future_data[:, 1:, :], future_prediction.reshape(1, 1, 1)), axis=1)

# Print the forecasted weekly suspected malaria cases
print("Forecasted Weekly Suspected Malaria Cases:")
for i, prediction in enumerate(future_predictions):
    print(f"Week {i+1}: {prediction:.2f}")

# Plot the actual vs. forecasted values
import matplotlib.pyplot as plt

plt.figure(figsize=(20, 6))
plt.plot(range(len(y_test)), y_test, label='Actual')
plt.plot(range(len(y_test)), y_pred, label='Forecasted')
plt.plot(range(len(y_test), len(y_test) + future_weeks), future_predictions, label='Future
Forecast')
plt.xlabel('Time (Weeks)')
plt.ylabel('Weekly Suspected Malaria Cases')
plt.title('Actual vs. Forecasted Malaria Cases')
plt.legend()
plt.show()

from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_absolute_percentage_error
import numpy as np
# Evaluate the model on the testing data
loss = model.evaluate(X_test, y_test)
print(f'Test loss: {loss}')

# Make predictions on the testing data
y_pred = model.predict(X_test)

```

```

# Evaluate the predictions using various metrics
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
mape = mean_absolute_percentage_error(y_test, y_pred)

print(f'MSE: {mse}')
print(f'MAE: {mae}')
print(f'MAPE: {mape}')

import matplotlib.pyplot as plt
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.legend()

import numpy as np
import pandas as pd
from keras.models import Sequential
from keras.layers import LSTM, Dense, Input
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_absolute_percentage_error

# Assuming you have the necessary data in a DataFrame 'lf'

# Preprocess the data
scaler = StandardScaler()
X = lf[['weekly rainfall Received', 'average weekly max temp', 'average weekly min temp']]
y = lf['weekly positive malaria cases'] # weekly positive malaria cases

# Normalize the input and output variables
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1))

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.01,
random_state=42)

# Reshape the input data to include the exogenous factors
X_train = X_train.reshape(X_train.shape[0], X_train.shape[1], 1)
X_test = X_test.reshape(X_test.shape[0], X_test.shape[1], 1)

# Train the model

```

```

model = Sequential()
model.add(Input(shape=(X_train.shape[1], 1)))
model.add(LSTM(10))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
history = model.fit(X_train, y_train, epochs=10, batch_size=5, validation_data=(X_test, y_test))
model.summary()

# Make predictions on the test data
y_pred = model.predict(X_test)

# Inverse scale the predicted values
y_pred = np.clip(scaler.inverse_transform(y_pred), 0, None) # Clip negative values to 0
y_test = scaler.inverse_transform(y_test)

# Evaluate the model performance
mse = np.mean((y_test - y_pred) ** 2)
print(f'Mean Squared Error: {mse:.2f}')

# Forecast future values
future_weeks = 30 # Number of future weeks to forecast
future_data = X_test[-1].reshape(1, X_test.shape[1], 1) # Use the last row of X_test as the initial
input
future_predictions = []

for _ in range(future_weeks):
    future_prediction = model.predict(future_data)
    future_prediction = np.clip(scaler.inverse_transform(future_prediction), 0, None) # Clip
negative values to 0
    future_predictions.append(future_prediction[0][0])
    print(f'Forecasted weekly positive malaria cases for week {_ + 1}:
{future_prediction[0][0]:.2f}')

    # Update the future_data with the predicted value
    future_data = np.concatenate((future_data[:, 1:, :], future_prediction.reshape(1, 1, 1)), axis=1)

# Plot the actual vs. forecasted values
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
plt.plot(range(len(y_test)), y_test, label='Actual')
plt.plot(range(len(y_test)), y_pred, label='Forecasted')
plt.plot(range(len(y_test), len(y_test) + future_weeks), future_predictions, label='Future
Forecast')
plt.xlabel('Time (Weeks)')
plt.ylabel('Weekly positive malaria cases')

```

```
plt.title('Actual vs. Forecasted Malaria Cases')
plt.legend()
plt.show()

# Evaluate the predictions using various metrics
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
mape = mean_absolute_percentage_error(y_test, y_pred)

print(f'MSE: {mse}')
print(f'MAE: {mae}')
print(f'MAPE: {mape}')

plt.figure(figsize=(12, 6))
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.title("Training and Validation Loss")
plt.legend()
plt.show()
```

APPENDIX 2

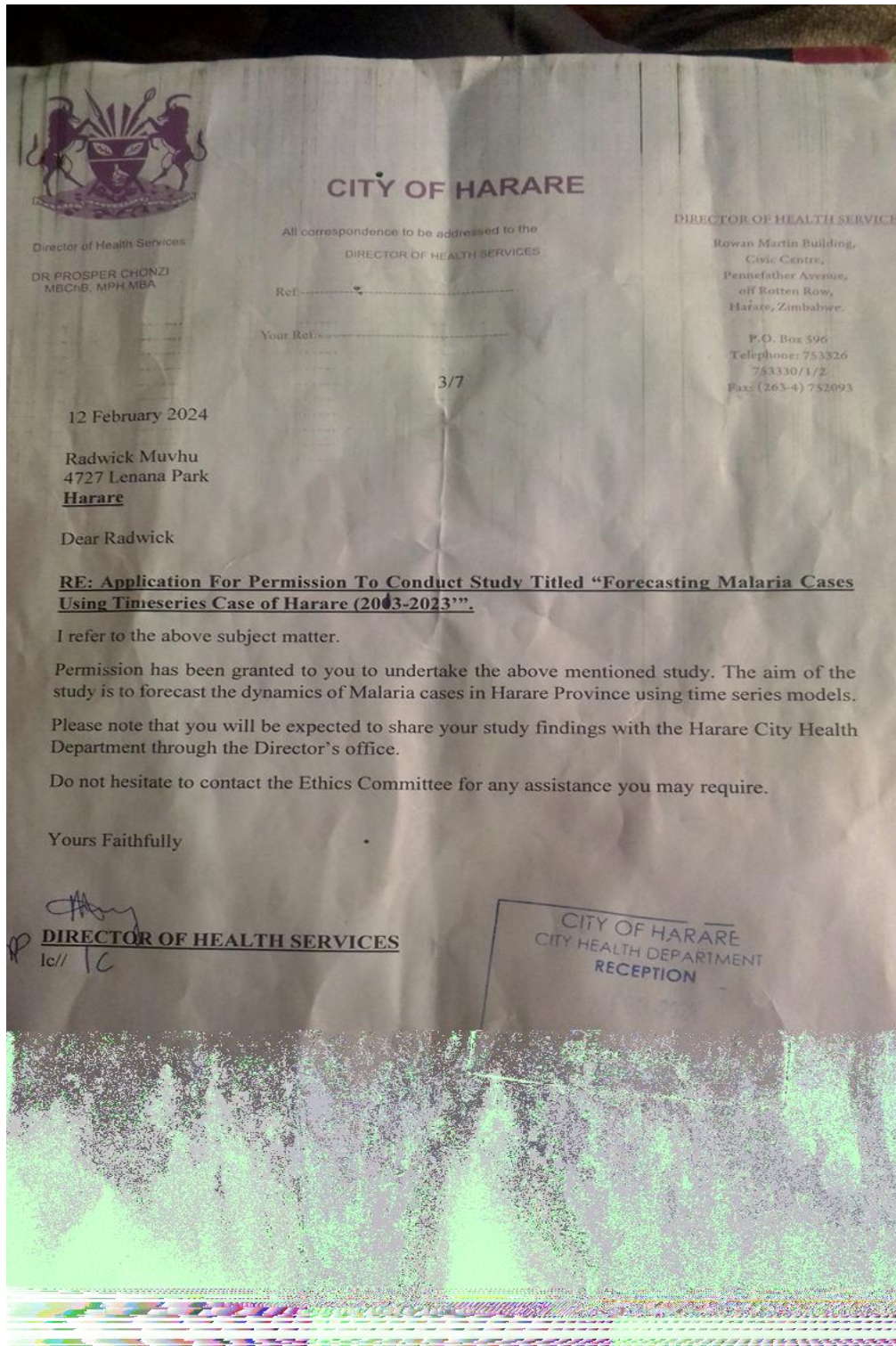


Figure 38 City of Harare approval form