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

# FORECASTING YOUTH MORTALITY IN ZIMBABWE USING HYBRID LEE-CARTER MODELS: A COMPARATIVE ANALYSIS OF ARIMA AND RANDOM FOREST APPROACHES

**BY**

**ASHLEIGH TAFADZWA VERENGAI B210275B**

**A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE HONORS BACHELOR OF SCIENCE DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS**

**2025**

SUPERVISOR: MR E. MUKONOWESHURO

## Authorship Declaration and Approval

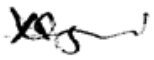
**Title of the Thesis:** Forecasting Youth Mortality in Zimbabwe Using Hybrid Lee-Carter Models. A Comparative Analysis of ARIMA and Random Forest Approaches

**Author:** Ashleigh T Verengai

**Program:** Bachelor of Science Honours degree in Statistic and Financial Mathematics

I, the undersigned author of the above-mentioned thesis, hereby declare that:

1. This thesis is my original work and has been prepared by me in accordance with the institution's requirements.
2. All sources, data, and references used in this thesis have been acknowledged and cited appropriately.
3. This thesis has not been submitted elsewhere for any degree or diploma.
4. I have obtained all necessary permissions for the inclusion of third-party content where applicable. I affirm that this declaration is made with full integrity and in compliance with the institution's policies and academic practice.

**Ashleigh Verengai**  **19/06/2025**

Student Signature Date

**B210275B**

Certified by:

**Mr.E. Mukonoweshuro** **19/06/2025**

Supervisor Signature Date

**Dr.M. Magodora**

Chairperson Signature Date

## DEDICATION

I dedicate this work to my beloved grandmother whose unwavering support, sacrifices and encouragement have been the foundation of my academic journey.

## ACKNOWLEDGMENTS

First and foremost, I want to sincerely thank God for giving me the strength and perseverance I needed to succeed academically.

My supervisor, Mr Mukonoweshuro, has provided me with direction, insightful input, and unwavering support throughout this project, for which I am incredibly grateful. Your guidance has greatly influenced both my academic development and this dissertation.

My sincere gratitude is extended to my family for their unwavering support, encouragement, and faith in my abilities. I am also appreciative of my friends who helped me along the process in many ways, exchanged ideas, and offered moral support.

## ABSTRACT

Youth mortality remains a critical public health issue in Zimbabwe, particularly for youth aged 0-24 years. Accurate mortality forecasting is vital for informing evidence-based health interventions and policies. In this research, there is a comparative evaluation between two hybrid models for mortality forecasting: the hybrid Lee-Carter model with Auto-Regressive Integrated Moving Average (LC-ARIMA) and the hybrid Lee-Carter model with Random Forest regression (LC-RF). Mortality rate data from 1990-2011 was obtained from UNICEF and Singular Value Decomposition was used to estimate parameters for the model. Estimates for the period 2012-2022 were generated by utilizing ARIMA and Random Forest techniques to estimate the time-varying mortality index ($k_t$). Quantitative evaluation of the models was performed using error metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Percentage Error (RMSPE). Results show that the LC-RF model performed better in projecting mortality for age brackets 0–4, 5–9, and 20–24 years, where it effectively captured non-linear trends. Conversely, LC-ARIMA fared better in the 10–14 and 15–19 age brackets. The results demonstrate the value of hybrid model approaches and affirm the necessity to apply suitable models to the underlying trends in the data. This study contributes to the body of literature by applying hybrid mortality models to a relatively understudied age cohort in Zimbabwe and offers results relevant for public health planning, actuarial purposes, and future mortality research. For areas of further studies, the researcher could use other predictors such as socioeconomic or health system predictors. Other machine learning algorithms such as LSTM and XGBoost can be used.

# TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

**ABBREVATIONS AND ACRONYMS**

LC Lee-Carter

ARIMA Autoregressive Integrated Moving Average

RF Random Forest

MSE Mean Squared Error

MAE Mean Absolute Error

RMSPE Root Mean Square Percentage Error

# CHAPTER 1: INTRODUCTION

## 1.0 Introduction

Youth mortality in Zimbabwe for the ages 0 to 24 is a serious public health challenge. This age group is affected by socioeconomic issues such as poverty, a weak health infrastructure system and limited access to health care (Levy and Sidel, 2009). The need to address youth mortality is not only important for individual and community well-being but also in shaping effective health policies and interventions.

Precise modeling and forecasting of mortality rates among youths is important in devising specific health policies and interventions. Precise forecasting will also provide indications on health policies that might be implemented to reduce mortalities, including the needs of the different age groups comprising the youth. This dissertation explores and compares two advanced techniques in the forecast of mortality rates among the youths: the Lee-Carter-ARIMA (LC-ARIMA) hybrid model and Lee-Carter-Random Forest (LC-RF). The study uses data-driven approaches to enhance predictive accuracy and inform evidence-based policy decisions.

The first chapter of the dissertation is an introduction, it details the background of the problem, presents a complete problem statement and states the objectives of the study. It states the research questions and describes assumptions and limitations of the study. Definitions of the main terms are given for better understanding.

In chapter 2, the literature on the Lee-Carter model and Random Forest approaches is thoroughly reviewed. The application of these models in mortality prediction and their relationships to youth mortality are also looked at. Chapter 3 discusses the datasets and analytical methods used to compare the two models as well as the research approach employed in this work. Chapter 4 discusses the analysis and results of the study. Chapter 5 summarizes the findings of this dissertation, discusses the implications and provides recommendations for further research. Using this systematic design, the present study is intended to help model and forecast mortality rates of Zimbabwean youths.

## 1.1 Background of the problem

*Global Perspective*

Youth mortality is considered a critical public health challenge worldwide, with the major burden experienced by low- and middle-income countries. According to UNICEF, in 2019 there were approximately 5.2 million deaths among children below the age of five, with the highest proportion from sub-Saharan Africa (UNICEF, 2020). Poverty and inadequate healthcare systems are some of the factors that contribute to preventable causes of death, such as malnutrition, infectious diseases and injuries (Liu et al, 2016). Sustainable Development Goal (SDG) 3 strives to minimize under-five mortality to at most 25 deaths per 1,000 live births by 2030 (United Nations, 2015).

*Continental Perspective*

Deaths are high among youths on the African continent compared to the rest of the world. Over 50% of the global annual deaths of children and youths are due to diseases such as HIV/AIDS, malaria and tuberculosis, and complications from maternal and neonatal health occur in sub-Saharan Africa alone (World Health Organization 2021). Death rates are decreasing slowly because health facilities do not have enough resources, and most people have limited access to preventive measures. However, new forecasting techniques and statistical models are being used to distribute resources and to guide interventions.

*National Perspective*

In Zimbabwe, youth mortality mirrors the continental trends, worsened by national challenges of economic instability, political tensions and strain on the healthcare system (Mavhandu-Mudzusi et al, 2018). For example, the infant mortality rate was recorded at 69 deaths per 1,000 born alive according to the Demographic and Health Survey that was done in 2015 in Zimbabwe (ZIMSTAT, 2016). There are many factors that are responsible for high youth mortality. Some of them are the non-availability of proper healthcare services, high prevalence of HIV/AIDS, and poor maternal healthcare (Chikanda & Matanda, 2020). Precise modelling of mortality trends is needed for public health policy and services impact.

*Direction Taken by Previous Studies*

There have been many studies done on mortality trends in Zimbabwe. For example, (Chikanda and Matanda, 2020) studied the effectiveness of health policies to avert juvenile mortality and emphasized the need for integrated healthcare services targeting adolescents. They also discussed the contribution of HIV/AIDS to youth mortality and demanded increased access to antiretroviral therapy and education. Since its creation in 1992, the Lee-Carter model has been well known for its ability to predict mortality. This is because it can accurately show the long-term trajectory of age specific death rates. In Zimbabwe, it was combined with ARCH models to predict death rates for individuals aged 0 to 85 (Taruvinga et al, 2017). Recent works have combined the Lee-Carter framework with machine learning methods such as Random Forests and Artificial Neural Networks forming hybrid models. This greatly improved predictive accuracy (Hong et al, 2021). Random Forests have been found quite promising in capturing nonlinear patterns and improving mortality forecasting accuracy across a wide range of settings (Czado et al, 2021).

Despite these developments, there are still some gaps in the modeling of mortality for youths aged 0 to 24 years in Zimbabwe. This dissertation compares the predictive performance of the Lee-Carter model and Random Forests in forecasting the rates of youth mortality. Through such advanced methods, the study will be able to give recommendations useful in the design of targeted interventions and health policies.

### 1.2 Statement of the problem

Youth mortality is a public health concern in Zimbabwe. Accurate death rate prediction is essential for making informed policy decisions and implementing effective health interventions. Conventional methods such as the Lee–Carter and ARIMA models may not adequately capture the non-linear relationships that occasionally show up in mortality statistics. Machine learning algorithms like Random Forest are more versatile but they lack the interpretability of traditional statistical models. This study aims to narrow the gap by comparing the accuracy of two hybrid models, Lee-Carter with ARIMA and Lee–Carter with Random Forest in modelling and forecasting mortality rates. Not only will this help improve health interventions for young people, but it would assist actuaries and insurers to develop life and health insurance for the youth.

### 1.3 Research Objectives

To better characterize and forecast mortality patterns in young populations of Zimbabwe, this study aims to accomplish the following:

1. To model overall mortality rates for ages 0 to 24 in Zimbabwe for the years 1990 to 2011 using LC-ARIMA and LC-Random Forest hybrid models.
2. To evaluate and compare the accuracy of LC-ARIMA and LC-Random Forest hybrid models in forecasting mortality for the target age groups using quantitative performance metrics for the years.
3. To examine age specific mortality forecast results and make inferences to guide targeted public health interventions in Zimbabwe.

### 1.4 Research questions

To fill the knowledge gap existing in mortality rate prediction in young populations, this study will attempt to answer the following questions:

1. How effectively can the LC-ARIMA and LC-Random Forest hybrid models be used to model overall mortality rates for Zimbabwean youths aged 0 to 24 during the period 1990 to 2011?
2. Which hybrid modeling approach, LC-ARIMA or LC-Random Forest, provides more accurate forecasts of mortality rates for youths aged 0 to 24 in Zimbabwe?
3. What are the public health implications of the age specific mortality projections generated by the LC-ARIMA and LC-Random Forest hybrid models for Zimbabwean youth?

### 1.5 Delimitations of the study

The study focuses solely on youth aged 0-24 years in Zimbabwe excluding other age categories. The study will be confined to Zimbabwe, which excludes regional and global trends in mortality with a view to providing specific insights for that country. Alternative machine learning methods, including LSTMs and support vector machines are not considered in this analysis. This

study looks only into overall mortality rates without showing subgroup analyses, for instance by cause of death, by gender, or by region. Forecasting accuracy is emphasized, instead of an in-depth exploration of the causal factors. Although the findings aim to inform health policy, the study does not address the implementation of interventions or evaluate existing public health programs.

### 1.6 Assumptions of the study

The mortality data for Zimbabwean youths aged 0 to 24 from 1990 to 2022 is accurate and reliable. It reflects true trends in mortality. It is assumed that the time-varying mortality index $k_t$ (kappa) of the Lee-Carter model is stationary enough to allow ARIMA models to provide valid forecasts after differencing. The Lee-Carter framework is assumed to be suitable for modeling and forecasting mortality rates in Zimbabwe, including its ability to capture age-specific mortality trends. It was assumed that the hybrid models of the LC-ARIMA and LC-Random Forest were comparably capable of handling the given data and generating meaningful forecasts about mortality rates.

### 1.7 Limitations of the study

The models assume that the future mortality will be predicted by the past mortality trend represented through $k_t$. This hypothesis may not consider the effects of unforeseen disruptions, like pandemics, economic crisis or major healthcare intervention. The model does not separate death by causes it only considers overall mortality. Age-specific parameters such as alpha ($\alpha$) and beta ($\beta$) remain constant during the period under study.

### 1.8 Definition of terms

*Mortality Rate*

The mortality rate refers to how often deaths occur in a specific population over a given period, typically expressed per 1000 or 100 000 people. It forms one of the critical indicators of public health and demographic features. (World Health Organization, 2021).

Lee-Carter Model

The Lee-Carter model captures how age specific mortality rates change over time by applying singular value decomposition to estimate parameters that reflect long term mortality trends (Pedroza, 2013).

*Random Forest*

Random forest is a machine learning algorithm that forms numerous decision trees while training and then joins together their outputs in efforts to improve predictive accuracy (Breiman, 2001).

*Forecasting*

Forecasting is a method of predicting future occurrences or evolutions of the situation which is founded on the statistical analysis of historical data. It is widely applied to the fields of economics, meteorology and public health for prediction of future values. (Hyndman Athanasopoulos, 2018).

## 1.9 Summary

This chapter has pointed out the significance of modeling and forecasting mortality rates of youths in Zimbabwe. This allows implementation of public health policies and interventions that are evidence-based. The research objectives and questions were formulated to guide the comparative analysis of the LC-ARIMA and LC-RF hybrid models in addressing this challenge. This chapter forms a basis for the next chapter that will undertake a broad review of the relevant existing literature and theoretical frameworks.

# CHAPTER 2: LITERATURE REVIEW

## 2.0 Introduction

This chapter aims to review both theoretical and empirical literature, which are relevant to modeling and forecasting mortality rates among the youth in Zimbabwe. The study will investigate various relevant theories and frameworks. Theoretical literature will focus on Lee-Carter model and Random Forest techniques which are of importance in this study. The empirical literature review will present findings from previous studies.

Gaps in relevant literature will be identified, which are the lack of comparative studies on mortality forecasting for youths aged 0 to 24 in Zimbabwe and integration of Lee- Carter with ARIMA and Random forests to forecast mortality rates in Zimbabwe. A conceptual framework that outlines the methodological approach for addressing these gaps will also be introduced.

The chapter is as follows: Section 2.1 presents the theoretical literature relevant to this study. Section 2.2 presents empirical studies that support and oppose mortality modeling and those with mixed findings. Section 2.3 identifies the research gaps this study tries to fill. Finally, section 2.4 shows the proposed conceptual framework and Section 2.5 concludes the chapter by summarizing the key points.

## 2.1 Theoretical Literature

This section deals with the theories and mathematical models that are used for this research. These are namely the Lee-Carter model and hybridization of the same with ARIMA and Random Forest. These methods form the theoretical framework of the modeling and forecasting of Zimbabwean youth mortality.

*The Lee-Carter model*

Lee and Carter first came up with the Lee-Carter model in 1992. It has widespread use in demography to model and predict mortality rates. The model decomposes the log of age-specific mortality as

$$ln\, m_{\{x,t\}} = \alpha_x + \beta_x k_t + \epsilon$$

where $\alpha_x$ represents the average age-specific log mortality, $\beta_x$ the sensitivity of mortality to changes in $k_t$ and $\epsilon$ the error term (Basellini et al, 2022).

Parameter estimation for the Lee-Carter model is done using the Singular Value Decomposition (SVD) technique (Taruvinga et al 2017). The parameters $\alpha_x$, $\beta_x$ and $k_t$ are identified under the constraints $\sum k_t$ =0 and $\sum \beta_x$=1 to ensure identifiability (Basellini et al, 2022).

let M be the matrix of log-transformed mortality rates, where:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2\ \ldots} & m_{1,t} \\ m_{2,1} & m_{2,2\ldots} & m_{2,t} \\ \vdots & \vdots & \vdots \\ m_{x,1} & m_{x,2} & m_{x,t} \end{bmatrix}$$

Using SVD, the matrix M can be decomposed as: $M = U\,\Sigma V^T$

$$\begin{bmatrix} m_{1,1} & m_{1,2\ \ldots} & m_{1,t} \\ m_{2,1} & m_{2,2\ldots} & m_{2,t} \\ \vdots & \vdots & \vdots \\ m_{x,1} & m_{x,2} & m_{x,t} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2\ \ldots} & u_{1,t} \\ u_{2,1} & u_{2,2\ldots} & u_{2,t} \\ \vdots & \vdots & \vdots \\ u_{x,1} & u & u_{x,t} \end{bmatrix} \begin{bmatrix} \delta_1 & 0\ \ldots & 0 \\ 0 & \delta_{2\ldots} & 0 \\ & \vdots & \\ 0 & 0 & \delta_x \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2\ \ldots} & v_{1,t} \\ v_{2,1} & v_{2,2\ldots} & v_{2,t} \\ \vdots & \vdots & \vdots \\ v_{x,1} & v_{x,2} & v_{x,t} \end{bmatrix}$$

Where:

U: The left singular matrix ($x \times x$), containing the left singular vectors.

Σ: A diagonal matrix ($x$ × t) of singular values, which represents the magnitude of the data's variance along the principal components.

$V^T$: The transpose of the right singular matrix (t × t), containing the right singular vectors.

The sensitivity of mortality at age $x$ ($\beta_x$) and the time varying index $t$ ($k_t$ ) are obtained by applying the following equations.

$$\hat{k} = \delta_1 \times (v_{1,1}\ \ v_{2,1}\ \ \ldots v_{t,1})$$

$$\widehat{b_x} = \frac{1}{\sum u_{x,1}} \times (u_{1,1}\ u_{2,1}\ \ldots u_{x,1})^T$$

We assume the residuals $\epsilon$ are independent and identically distributed with constant variance. Residuals could show age-dependent autocorrelation or heteroscedasticity. These restrictions may have an impact on model fit, particularly for age groups with erratic trends or sparse data (Cairns et al, 2009).

The Lee-Carter model has been widely used because it is somewhat simple. Mortality forecasts are generated using a single time index. This allows for probabilistic prediction for all mortality metrics, and it relies entirely on extrapolation avoiding subjective expert opinions or external input (Bassellini, 2022).

The model's dependence on extrapolation has advantages and disadvantages. It makes predictions reproducible, and data driven. However, it might not be as good at capturing sudden changes in the population or nonlinear transitions. This includes those brought on by wars and epidemics (Villegas et al, 2018).

Actuaries and insurers use this model for pricing pensions, life insurance and annuities. It is also suited for risk assessment and solvency estimation in financial applications due to its capacity to generate long range forecasts with uncertainty bands (Richards et al, 2007)

*ARIMA (Auto-Regressive Integrated Moving Average)*

ARIMA is a time series model that can be used for forecasting. It combines autoregression (AR), differencing (I), and moving averages (MA). It models temporal changes in mortality rates using past values and residuals. The $k_t$ index, after being estimated through a two-stage process, can be forecasted using ARIMA. $k_t$ follows a random walk model with a drift component (Haberman and Renshaw, 2011). The drift component captures the average annual change in mortality. A negative drift shows a decline in mortality and a positive drift shows an increase in mortality. This is mathematically expressed as:

$$k_t = k_{\{t-1\}} + d + \epsilon$$

Recent advancements have led to the hybridization of the Lee-Carter model with other machine learning models like Neural Networks and Random Forest. The ARIMA component is replaced with a more flexible non-linear machine learning model (Hong et al, 2021). The mortality index ($k_t$) which is usually forecasted using ARIMA is then forecasted using machine learning. This study will combine the Lee-Carter model and the Random Forest regressor to form a hybrid model.

*Random Forest*

Random Forest is a machine learning approach that develops several decision trees to improve a target variable's prediction accuracy (Breiman, 2001). The Law of Large Numbers prevents the model from overfitting and makes it robust to noise. For variance reduction and to prevent overfitting, the model aggregates predictions across several trees. By combining the multiple decision trees, Random Forest identifies the nonlinear associations and higher-order interactions within mortality data. Random Forest enhances and identifies the nonlinear trend of mortality rates within this hybrid approach by LC-RF. Random Forests are particularly helpful when considering nonlinear relationships between variables. They have the following advantages: 1) they are among the most accurate learning algorithms; 2) they can handle data that is unbalanced; and 3) they provide estimations of the predictive ability of variables utilized. A random subset of predictor variables is then evaluated at each node for splitting. There are predictions by the model based on composite predictions across several trees to reduce variance and prevent overfitting. Random Forest identifies nonlinear relationships and interactions of high order within mortality data. Random Forest refines and captures the nonlinear trend of mortality rates in this hybrid approach.

The Random Forest can be represented mathematically as,

$$\hat{h}_{RF}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{h}(x, \Theta_i)$$

where $\hat{h}(\Theta_i)$ are the random tree predictors and $\Theta_1 \ldots \Theta_n$ are random variables.

A bootstrap sample of the data is used to train each tree. A random subset of predictor variables is then taken into consideration for splitting at each node. By adding feature randomization and bagging (bootstrap aggregation), this improves the generalizability of the model. Random Forest performs well in high dimensional environments with varied interactions and non-linear correlations (Cutler et al, 2007; Biau and Scornet, 2016).

In mortality forecasting, Random Forest has been used to replace traditional time series models. $k_t$ values are forecasted and modeled by the Random Forest model (Hong et al, 2021). The method has been reported to produce encouraging results in countries that experience high complexity in mortality driven by environmental, health, and socioeconomic factors (Guelman et al, 2015 and Czado et al, 2021).

## 2.2 Empirical Literature Review

Empirical studies indicate evidence on the efficiency and shortcomings of both the Lee-Carter model and Random Forest in mortality forecasting. Various studies supported the model's use. For instance, a study by (Taruvinga et al, 2017) explored the efficiency of the Lee-Carter model combined with ARIMA in the forecasting of mortality trends in Zimbabwe. They also compared the Lee-Carter to the ARCH model in predicting mortality rates and discovered that it performed better than the ARCH model. They also discovered that even though the two models showed weakness in older age groups the Lee-Carter model was a better overall fit for Zimbabwean mortalities.

Similarly, (Hong et al, 2021) found that machine-learning methods including Random Forest made much better captures of non-linear mortality patterns than traditional demographic models. They also discovered that the Lee-Carter-ARIMA hybrid model is the most suitable model to use when predicting mortality rates for countries with good healthcare systems and longer life expectancy. Additionally, the Lee-Carter-Random Forest exhibited lower RMSPE values for Malysia's youth age groups (Hong et al, 2021). Forecasting accuracy increased significantly when Random Forest was applied to communities undergoing socioeconomic transformations (Czado et al, 2021). This supports the use of Random Forest to model mortality rates in Zimbabwe. The country experiences high rates of HIV/AIDS, malnutrition and access to healthcare is a challenge and this creates nonlinear mortality patterns.

However, some have pointed out certain misgivings with these methods. The Lee-Carter model might not be suitable for those populations that experience sudden socioeconomic discontinuities (Villegas et al, 2018). There was evidence revealed which showed that machine-learning approaches such as Random Forest are prone to overfitting when utilized on smaller data sets, limiting their generalizability (Kogure et al, 2020). Additionally, (Yu et al, 2017) emphasized the trade-off between accuracy and interpretability in Random Forest models, a concern for mortality studies aimed at informing public policy.
Several have reported inconclusive results. (Haberman and Renshaw, 2012) showed that the model works very well in a long-run projection but for the short-term, its results were unsatisfactory over smaller populations. (Hanewald et al, 2013) reported an improved accuracy by inclusion of machine-learning methods together with Lee-Carter, at the expense of higher computational complexity.

The Lee-Carter model has mainly been used in forecasting mortality rates for the purpose of pricing life insurance contracts (Taruvinga et al, 2017). When insurance companies are unable to predict mortality rates accurately this could result in overpriced premiums and other consumers might be unable to insure themselves. Accurate prediction allows them to provide new policies with reasonable prices (Hong et al, 2021)

*Research Gap*

Despite these improvements in mortality modeling, gaps exist. The 0 to 24 years age group has been rarely covered in studies due to varying mortality trends and determinants. Even though the Lee-Carter model is well-studied, its integration with machine-learning techniques such as Random Forest is unexplored in Zimbabwe (Hong et al, 2021). While many studies have been conducted on the subject, they were primarily focused on model accuracy and did not place much emphasis on explainability and usability in informing policy formulation. The study aims at filling such knowledge gaps through comparative evaluation of the LC-ARIMA and LC-Random Forest hybrid models' predictive power for youth mortality in Zimbabwe.

## 2.3 Proposed Conceptual Framework

This section builds on the theoretical and empirical findings from previous studies.

*Dependent Variable: Youth Mortality Rates (0–24 Years)*

This research compares two hybrid models that combine the Lee-Carter with both ARIMA and Random Forests. The study uses this dual-model approach because of the strengths and limitations stated in the literature regarding both linear and nonlinear methods of forecasting mortality rates.

Dependent variable in this study is the youth mortality rate. It is the mortality per unit of population in the 0–24-year age group in Zimbabwe. It considers mortality trends in several subgroups like children and young adults. Youth mortality is influenced by differing socio-economic, demographic, and health system determinants. This includes poverty, healthcare access levels, HIV/AIDS prevalence and education (Kembo and Van Ginneken, 2009). The study estimates and projects mortality rates between 1990–2022 with the aim of establishing historical patterns and informing future projections.

*Independent Variables/Predictors*

Independent variables are factors that influence youth mortality rates.

*1.Age-Specific Effects*

Mortality rates differ significantly for varying age groups, and the risks each group faces are unique to them. Lee-Carter model captures these differences employing two main parameters: $\alpha$ capturing baseline mortality levels for each age group and $\beta$ capturing the responsiveness of each age group to changes in mortality over time. Infants (0-4 years) die due to health-related causes, including neonatal and infectious diseases, while young adults (18-24 years) fall prey to accidents and drug misuse (Gakidou et al, 2010).

*2.Time Trends*

A thorough understanding of temporal dynamics is necessary to better understand mortality. The HIV/AIDS epidemic, health crises or economic challenges may prompt changes in trends. These

trends are represented by the mortality index $k_t$, summarizing overall changes in mortality levels across all age groups (Hyndman and Booth, 2008).

*Hybrid Modeling Approaches*

The LC-ARIMA and LC-Random Forest hybrid models share the same foundational structure. They both decompose a matrix of age specific mortality rates into age specific constants, a time varying constant and an error term. The difference between the two models is in forecasting the time varying mortality index $k_t$.

The first component of the conceptual framework is the mortality rate data input. The input data consists of the historical age-specific mortality rates from 1990 to 2011 for the Zimbabwean youths aged between 0 to 24 years. This data set is used for parameter estimation and model training. For the second stage, parameter estimation is carried out using the Singular Value Decomposition method.

Forecasting of the mortality index $k_t$ is done in the third step. The LC-ARIMA uses Partial Autocorrelation plots and ADF tests to determine the appropriate ARIMA model to use. The model assumes linearity and stationarity. The LC-Random Forest model uses Random Forest regression to find $k_t$. Conversely this model captures nonlinear relationships.

In the fourth step the mortality indexes obtained are then combined with the previously estimated $\alpha_x$ and $\beta_x$ parameters. This gives us the predicted age specific mortality rates for future periods.

The final step focuses on model evaluation. The ability of each model to forecast mortality rates is assessed using standard error metrics.

## 2.4 Expected Contributions

*Enhanced Predictive Accuracy*

This work combines the principles of the Lee-Carter framework with ARIMA and Random Forest, with the aim to improve prediction accuracy. ARIMA is the best tool for describing

temporal dependencies, while Random Forest captures non-linear relationships. These methods, when combined, offer a robust means of forecasting mortality rates.

*Model Comparison*

During the study, the two hybrid models are evaluated using metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) to check which of the two models is better suited to model and predict mortality rates.

*Policy Implications*

Policymakers can use accurate projections of youth mortality rates to design focused health interventions. This includes identifying high-risk age groups in programs, allocating scarce resources to areas where the average life expectancy is higher and developing programs that address various socio-economic factors related to health.

## 2.5 Summary

This chapter reviewed theoretical and empirical studies based on the Lee-Carter model, ARIMA, and Random Forest approaches. It outlined their advantages and disadvantages while concentrating on mortality modelling and predictions. It pointed out holes in earlier studies and offered a conceptual framework to fill them. In addition to the theoretical and empirical findings covered here, the methodology and data analysis strategies employed in this study will be covered in detail in the following chapter.

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.0 Introduction

This chapter attempts to show the methodology used in conducting the research on the comparative analysis of LC-ARIMA and LC-Random Forest Hybrid models. This methodology consists of research design, data sources, description of variables that include analytical model specifications, model diagnostic tests, and ethical considerations.

## 3.1 Research Design

The study follows a quantitative research design that uses statistical modeling and machine learning techniques to help analyze and predict mortality rates. This is an approach that enables the identification of patterns and trends in mortality rates in Zimbabwe, especially for youths. It also compares and evaluates the forecasting accuracy of two hybrid models. This comparison helps us to ascertain the most appropriate model for forecasting mortality rates to inform public health strategies, academics and even the insurance sector. The integration of the traditional Lee-Carter model and machine learning using Random Forest allows for a more comprehensive approach. The study is exploratory because it seeks to identify mortality patterns and trends among Zimbabwean youths.

This would allow the combination of two conventional yet powerful statistical techniques with the modern machine learning models, enhancing the robustness of the forecasting. In turn, this will enable the testing of the limitations and strengths of each model. The research design comprises of model performance evaluations. The accuracy of the two models is assessed using error metrics such as Mean Absolute Error (MAE).

The research design for this study is divided into the following sections, namely data collection and preparation, model development, forecast reconstruction and model evaluation.

## 3.2 Population and Sampling

This study uses mortality rate statistics for Zimbabwean youth between the ages of 0 and 24

years inclusive. Statistics from valid sources are used. The sample period is 1990 to 2022 in order to have reliable trend analysis as well as accurate forecasting

### 3.3 Data Sources

The data for this study is sourced from the UNICEF global database.

### 3.4 Description of Variables and Prior Expectations

This section outlines the independent and dependent variables that were used. It also includes the anticipated outcomes for the two hybrid models, considering their features and historical patterns.

The focus of the study is on the mortality rates of young individuals in Zimbabwe aged 0 to 24, which serve as the dependent variable. In this case the mortality rate is the central death rate for each age group across different time periods. These mortality rates are expressed in logarithmic form which is consistent with the Lee-Carter model.

The first independent variable is time which is represented by the year of observation. The range is from 1990 to 2022. The mortality index, $k_t$ is designed to track changes in mortality over time.

The second independent variable is age groups. The data is aggregated into five-year bands which are 0-4, 5-9, 10-14, 15-19 and 20-24. Each of these groups represents a distinct phase of youth development. Each group has different risk exposure and health vulnerabilities. This stratification allows for a more accurate modeling of age-specific mortality dynamics within the broader youth category.

*Prior Expectations*

According to historical death patterns and theoretical reasoning, mortality rates are expected to decrease slowly with time. This is because of the improvements in health and life standards in Zimbabwe. LC-ARIMA is expected to capture linear trends well. LC-ARIMA is the best model

to use for projecting mortality in cases where changes are slow and predictable. In contrast, the Lee-Carter (Random Forest) model is perhaps better suited to find non-linear trends. The model would be able to detect sudden shifts in mortality patterns due to epidemics of diseases or distortions in the healthcare system more effectively.

### 3.5 Analytical Model Specification and Justification

This section outlines the analytical frameworks applied in the study to calculate age-specific mortality rates of individuals between 0 and 24 years in Zimbabwe. Model specification is supposed to illustrate the mathematical form and contents of the forecasting models employed. It also gives variables employed and how they interact. The Lee-Carter model was regarded as the base model based on its precision in demographic forecasting. The model has been expanded to include a machine learning method, Random Forest, and a traditional time-series approach, ARIMA. The key elements and structural equation of both hybrid models are presented.

*The Base Model: Lee-Carter Specification*

The Lee-Carter model is the base model for this study. This model was combined with ARIMA and Random Forests to forecast mortality rates.

$$ln\, m_{\{x,t\}} = \alpha_x + \beta_x k_t + \epsilon$$

Where

$m_{\{x,t\}}$ : actual mortality rate for age $x$ in year $t$,

$\alpha_x$ : average mortality specific to age $x$ ,

$\beta_x$: reflects the sensitivity of mortality at age $x$ to changes over time ,

$k_t$: time varying index

$\epsilon$: the random variation for age $x$ at time $t$ that is not explained by the model

ln ($m_{\{x,t\}}$) is the natural logarithm of the mortality rate (logarithm to base e). The parameters are estimated using Singular Value Decomposition. The parameter $k_t$ is then forecasted into future periods using either a time series model or a machine learning model.

*LC-ARIMA Hybrid Model*

This is the hybridization of the Lee-Carter model and ARIMA to predict the mortality index $k_t$. The selection of the ARIMA parameters (p, d, q) is guided by the Augmented Dickey Fuller (ADF) tests, Partial Autocorrelation Function (PACF) plots and Autocorrelation Function (ACF) plots.

*LC-Random Forest Hybrid Model*

A Random Forest regression model is used in place of the ARIMA forecasting step. A Random Forest method is trained to forecast future values of $k_t$ ,once $\alpha_x$ and $\beta_x$ historical values have been estimated. To provide predictions that are more accurate, this machine learning method constructs decision trees and aggregates their results.

*Justification for Hybrid Approaches*

The idea of using hybrid models for this study comes from the need of enhancing and improving the existing Lee-Carter model first proposed in 1992. Traditional models may not adequately predict mortality rates with complex non- linear relationships. The original model has proven to be effective in high income countries and this study aims to investigate its performance in a more volatile environment and combined with machine learning methods. By integrating ARIMA and Random Forests this enables us to address these instances.

**Steps in Data Analysis**

1. *Data preprocessing*

- Firstly, mortality rate data is collected from the UNICEF database. The data is downloaded in Microsoft Excel format which is suitable for analysis.
- Load the dataset on Google Collab's Python using pandas library.
- Cleaning and handling missing values. It is important to identify missing values as they would greatly impact the accuracy of the results. Check for null entries.
- Normalization of data. Machine learning methods require the data to be in the same range. This is done by applying the MinMaxScaler from the sklearn.preprocessing module.

2. *Model estimation*

Parameter Estimation Using Singular Value Decomposition

The Lee-Carter parameters are obtained using SVD. The mortality rate data must be in matrix form.

let M be the matrix of log-transformed mortality rates, where:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2\ \ldots} & m_{1,t} \\ m_{2,1} & m_{2,2\ldots} & m_{2,t} \\ \vdots & \vdots & \vdots \\ m_{x,1} & m_{x,2} & m_{x,t} \end{bmatrix}$$

Using SVD, the matrix M can be decomposed as: $M = U\,\Sigma V^T$

$$\begin{bmatrix} m_{1,1} & m_{1,2\ \ldots} & m_{1,t} \\ m_{2,1} & m_{2,2\ldots} & m_{2,t} \\ \vdots & \vdots & \vdots \\ m_{x,1} & m_{x,2} & m_{x,t} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2\ \ldots} & u_{1,t} \\ u_{2,1} & u_{2,2\ldots} & u_{2,t} \\ \vdots & \vdots & \vdots \\ u_{x,1} & u & u_{x,t} \end{bmatrix} \begin{bmatrix} \delta_1 & 0_{\ \ldots} & 0 \\ 0 & \delta_{2\ldots} & 0 \\ & \vdots & \\ 0 & 0 & \delta_x \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2\ \ldots} & v_{1,t} \\ v_{2,1} & v_{2,2\ldots} & v_{2,t} \\ \vdots & \vdots & \vdots \\ v_{x,1} & v_{x,2} & v_{x,t} \end{bmatrix}$$

Where:

U: The left singular matrix ($x \times x$), containing the left singular vectors.

Σ: A diagonal matrix ($x \times t$) of singular values, which represents the magnitude of the data's variance along the principal components.

$V^T$: The transpose of the right singular matrix ($t \times t$), containing the right singular vectors.

With the use of python programming language, the parameters are obtained using the Singular Value Decomposition method. The mean mortality specific to age $x$ ($\alpha_x$) is obtained by dividing the logarithm of age specific death rates with time.

The sensitivity of mortality at age $x$ ($\beta_x$) and the time varying index $t$ ($k_t$) are obtained by applying the following equations using python.

$$\hat{k} = \delta_1 \times (v_{1,1} \ v_{2,1} \ \dots v_{t,1})$$

$$\widehat{b_x} = \frac{1}{\sum u_{x,1}} \times (u_{1,1} \ u_{2,1} \ \dots u_{x,1})^T$$

- Use python's numpy.linalg.svd() to decompose M.
- Numpy, pandas and matplotlib for the graphical representation were utilized in these following steps.
- $\alpha_x$ was estimated by averaging log mortality rate for all age groups for the years 1990 to 2011.
- $\beta_x$ was taken from the first column of matrix U, normalized so that the sum of $\beta_x$= 1.
- $k_t$ was calculated from the first row of V transpose times the largest singular value, adjusted so that the means are zero (by Lee-Carter constraints).
- Differencing (for ARIMA) for historical $k_t$ values' stationarity using the Augmented Dickey Fuller (ADF) test with the statsmodel package.
- Identify the suitable ARIMA model based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

Forecasting using ARIMA

- The ARIMA (1,2,1) model is used to predict values for $k_t$
- The statsmodels.tsa.arima.model.ARIMA class in python is used.

$$k_t = k_{\{t-1\}} + d + \epsilon$$

Where:

$d$ represents the drift parameter,

$\epsilon$ represents the uncorrelated error term

- Forecasted values for $k_t$ are plugged back into the Lee-Carter equation to generate mortality rate forecasts for each age group.

*Forecasting using Random Forest*

- Prepare $k_t$ as target variable.
- Create lag-based features from past values of $k_t$ as predictors.
- Split data into training and test datasets, though in this case it is 80:20.
- Use *sklearn.ensemble.RandomForestRegressor* to train a model with python.

3. *Model evaluation*

- The trained models of Lee Carter (ARIMA) and Lee Carter Random Forest are used for predicting mortality rates from 2012 to 2022.
- Predicted $k_t$ values are combined with $\alpha_x$ and $\beta_x$ to reconstruct predicted log mortality rates.
- Compare model performance using error metrics.
-

The error metrics used to compare the two models are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Percentage Error (RMSPE).

### 3.6 Model Diagnostic, Validation and Reliability Tests

*Pre-model diagnostic*

Before fitting the ARIMA model first test the historical $k_t$ values for stationarity. The Augmented Dickey Fuller (ADF) test was used for this.

Null hypothesis: The series is non-stationary.

Alternative hypothesis: The series is stationary.

*Reliability Tests*

For ensuring the reliability of the models, the subsequent tests were utilized. The Mean Absolute Error (MAE) statistic reports the average deviation of the predictions made by the model from the actual mortality rates as quantified by the average magnitude of prediction errors. The best predictive model is represented by a lower MAE.

$$MAE = \sum \frac{|m_{x,t,observed} - m_{x,t,fitted}|}{number\ of\ m_{x,t,observed}}$$

The average of the squared differences between observed and expected values is called the mean squared error, or MSE. It squares the variances and hence is outlier sensitive. This metric is useful for giving more weight to larger errors.

$$MSE = \sum \frac{(m_{x,t,observed} - m_{x,t,fitted})^2}{number\ of\ m_{x,t,observed}}$$

Expressing predictive accuracy as a percentage allows the extent to which the model correctly fits the actual data to be more intuitively obvious. RMSPE gives a normalized metric that is more understandable, particularly when working with public health measures.

$$RMSPE = \sqrt{\sum \frac{\left(m_{x,t,observed} - m_{x,t,fitted}\right)^2}{number\ of\ m_{x,t,observed}}} \times 100\%$$

### 3.7 Ethical Considerations

Findings were presented clearly in an objective manner to ensure that conclusions made from the findings are viable and plausible. This is to avoid misinformation through misrepresentation of findings. The researcher will report all results honestly.

There is a high requirement to maintain data confidentiality and privacy, especially in the case of sensitive information, such as that related to mortality rates. The UNICEF database anonymously collects data, so there is no chance of anyone's personally identifiable information appearing in the study.

The analytical methods ensured that the models were suitable in answering the research questions. Multiple metrics have been used to measure the predictive ability of the models. It, therefore, ensures that the study is free of any aspect of bias and manipulation.

The results from this study are meant to contribute to academic and public health knowledge. This means that results are presented in a way that does not lead to public misinterpretation or panic.

### 3.8 Summary

This chapter showed in detail the research methodology used in this study to compare the Lee-Carter (ARIMA) and Lee-Carter (Random Forest) hybrid models. The avoidance of bias and manipulation was emphasized throughout. The methodological choices aim to provide an accurate conclusion on which of the two hybrid models is most suitable to model and forecast youth mortality rates.

# CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION

## 4.0 Introduction

This chapter will forecast mortality rate data by using the LC-ARIMA and LC-Random Forest hybrid models. This study aims to compare the possibilities of these two hybrid models in forecasting. The model will be developed by using historical death rate data from 1990 to 2011. It will be validated by using data from 2012 to 2022.

## 4.1 Descriptive Statistics

The mortality rates in this study are divided into five age groups which are 0-4, 5-9, 10-14, 15-19 and 20-24. The most important descriptive statistics are clearly illustrated in the table below. These include the third quartile (Q3) and first quartile (Q1), mean, minimum and maximum, and standard deviation for every age group. As such, the central tendency and change of trend in youth mortality is better illustrated.

*Table 4. 1 Summary Statistics*

| Age Group | Mean | Median | Minimum | Maximum | Standard Deviation | Q1 | Q3 |
|---|---|---|---|---|---|---|---|
| 0 to 4 | 80.304 | 90.223 | 47.729 | 100.213 | 17.833 | 62.291 | 93.637 |
| 5 to 9 | 7.403 | 8.336 | 4.198 | 8.783 | 1.551 | 6.199 | 8.444 |
| 10 to 14 | 6.572 | 6.984 | 4.805 | 7.568 | 1.001 | 5.631 | 7.475 |
| 15 to 19 | 10.003 | 10.056 | 7.512 | 11.210 | 0.940 | 9.590 | 10.698 |
| 20 to 24 | 19.138 | 19.083 | 14.396 | 24.304 | 3.445 | 15.877 | 22.414 |

Table 4.1 is a summary of significant trends in the age-specific mortality rates among Zimbabwean youth. Mean rate for the age group 0-4 was the highest at around 80.3 deaths per 1000. This is evidence that youngsters below five years of age are the most vulnerable to death

compared to the rest of the age groups. It was the most spread-out group with a standard deviation of 17.8. The minimum and maximum of 47.7 and 100.2 respectively indicate the mortality rates varied widely throughout the study.

For the age groups 5-9 and 10-14, mean mortality rates decline with increasing age. The groups have means 6.57and 7.40 respectively. Both these groups have lower standard deviations, which implies that mortality trends among school aged children are more stable. This shows that mortality risks lessen during mid-childhood.

From the age group 15-19 and to 20-24 a gradual increase in mean mortality rates is observed. The mean mortality rates are 10.0 and 19.1 respectively. The age group 20-24 has a standard deviation of 3.45. This may be a result of economic hardship and other socioeconomic risks faced by young adults. It is marked by increased mortality in infancy and early childhood, followed by some stability during adolescence and then ultimately another peak during young adulthood.

### 4.2 Diagnostic Tests

*Augmented Dickey-Fuller (ADF) Test*

To ensure validity of the modeling process it is important to ensure that the mortality index series $k_t$ is stationary. The ADF test, a commonly used unit root test was performed to check the mortality index kappa for stationarity. A unit root test checks whether a time series shows non-stationarity or change in mean and variance over time.

Null hypothesis ($H_0$): The series is non-stationary (has a unit root).

Alternative hypothesis ($H_1$): The series is stationary.

Stationarity of the mortality indicators was obtained after second differencing. The ADF test statistic was -5.12 and p-value $1.27 \times 10^{-5}$. The critical value is higher than the test statistic and the p-value is far less than 0.05 significance level. This indicates that the ADF test decisively rejects the null hypothesis of a unit root (non-stationarity).

### 4.3 Analytical Model

*Lee-Carter Model Estimation*

The Lee-Carter model was adopted as the basis in projecting the mortality rates. To obtain the parameters the mortality rates are first logarithm transformed. The Singular Value Decomposition technique is then applied to the logarithm transformed mortality matrix. The technique decomposes the matrix into three matrices, namely, the U, V and S matrices.

From these three matrices, $\alpha_x$ was calculated as the average logarithm mortality rate for each age group across all years. $\beta_x$ was obtained from the first column of matrix U, normalized such that the sum of $\beta_x$= 1. $k_t$ was computed from the first row of V transpose multiplied by the largest singular value, rescaled to ensure that the means equal zero (according to Lee-Carter constraints).

These values for $\alpha_x$, $\beta_x$ and $k_t$ are then used to obtain future mortality values using ARIMA and Random Forests. The values are as shown below.

*Table 4. 2 : Age-specific Lee-Carter parameters*

| **Age Group** | **Alpha** $(\alpha_x)$ | **Beta** $(\beta_x)$ |
|---|---|---|
| 0-4 | 4.570371 | 0.115780 |
| 5-9 | 2.580941 | 0.026171 |
| 10-14 | 2.390467 | 0.186441 |
| 15-19 | 2.701620 | 0.265999 |
| 20-24 | 3.237091 | 0.405609 |

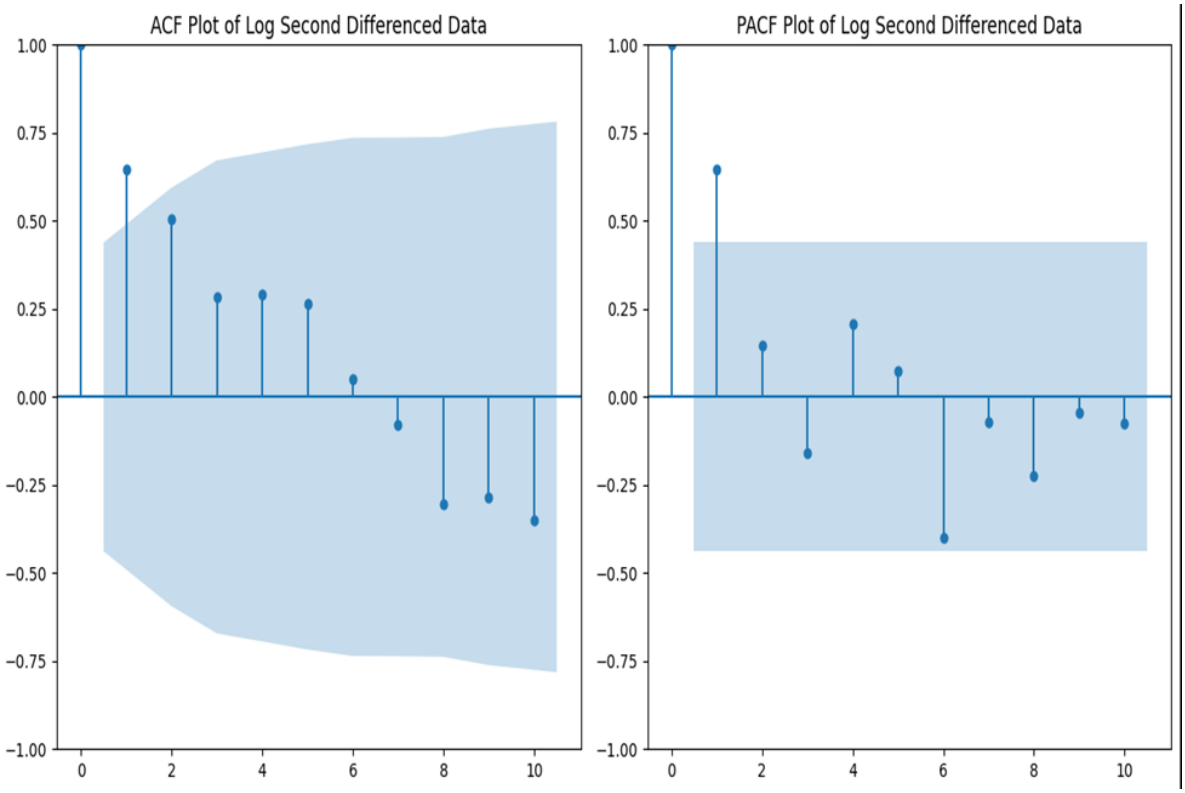
*Table 4. 3 Mortality Index Over Time (1990-2011)*

| Year | Kappa ($k_t$) | Year | Kappa ($k_t$) |
|---|---|---|---|
| 1990 | -0.380188 | 2001 | 0.16127313 |
| 1991 | –0.310297 | 2002 | 0.139966 |
| 1992 | -0.239386 | 2003 | 0.132223 |
| 1993 | -0.1696 | 2004 | 0.122638 |
| 1994 | -0.102782 | 2005 | 0.110784 |
| 1995 | -0.040617 | 2006 | 0.098632 |
| 1996 | 0.013064 | 2007 | 0.081493 |
| 1997 | 0.057386 | 2008 | 0.054769 |
| 1998 | 0.092236 | 2009 | 0.022196 |
| 1999 | 0.1167 | 2010 | 0.015353 |
| 2000 | 0.1769955 | 2011 | -0.05539 |

*ARIMA FORECASTING OF* $k_t$

To find an appropriate ARIMA model, plots of the Autocorrelation Function and Partial Autocorrelation Function have been obtained to propose Autoregressive and Moving Average terms to be included in the model. From these plots, there was a clear peak at lag 1 observed using ACF, and this indicates a first order MA. The PACF plot displayed a peak at lag 1 that was followed by the quick decline indicating AR of 1. An ARIMA (1,2,1) model is thus valid in the present research as the kappa series was stationary following second differencing.

*Figure 4. 1 ACF and PACF plots*




*Forecasted Kappa Using ARIMA*

The ARIMA (1,2,1) model was used to obtain forecasts for kappa from 2012 to 2022.

*Table 4. 4 : Forecasted Kappa (2012-2022)*

| **Year** | **Kappa** | **Year** | **Kappa** |
|---|---|---|---|
| 2012 | 0.010647 | 2018 | -0.195034 |
| 2013 | -0.023862 | 2019 | -0.229290 |
| 2014 | -0.057988 | 2020 | -0.263540 |
| 2015 | -0.092302 | 2021 | -0.297793 |
| 2016 | -0.126523 | 2022 | -0.332044 |
| 2017 | -0.160790 | | |

After obtaining the relevant mortality indexes for 2012 to 2022 the previously calculated alpha, beta and kappa were used to calculate mortality rates for 2012 to 2022 for each age group (0-4, 5-9, 10-14, 15-19 and 20-24).

*Figure 4. 2 ARIMA Forecasts (2012-2022)*

| | year | age_0_to_4 | age_5_to_9 | age_10_to_14 | age_15_to_19 | age_20_to_24 |
|---|---|---|---|---|---|---|
| 0 | 2012 | 85.801465 | 7.993260 | 7.234061 | 10.111486 | 19.832404 |
| 1 | 2013 | 83.759335 | 7.852050 | 7.213330 | 10.023280 | 19.300867 |
| 2 | 2014 | 82.382533 | 7.724988 | 7.213157 | 9.969164 | 18.779038 |
| 3 | 2015 | 81.431693 | 7.608238 | 7.185270 | 9.914479 | 18.269015 |
| 4 | 2016 | 80.300331 | 7.502967 | 7.149066 | 9.867751 | 17.763535 |
| 5 | 2017 | 79.448335 | 7.410735 | 7.098085 | 9.757079 | 17.260769 |
| 6 | 2018 | 78.370800 | 7.328986 | 7.016013 | 9.599106 | 16.791198 |
| 7 | 2019 | 77.708568 | 7.252287 | 6.906999 | 9.366610 | 16.319133 |
| 8 | 2020 | 77.052154 | 7.173160 | 6.767820 | 9.143794 | 15.864980 |
| 9 | 2021 | 75.967137 | 7.107707 | 6.640758 | 8.920840 | 15.420494 |
| 10 | 2022 | 75.169880 | 7.041880 | 6.494323 | 8.660791 | 14.972163 |

*Random Forest Regression Forecasting of $k_t$*

A machine learning method was used to forecast values for $k_t$ . Using historical kappa values the model generated lag-based features which were then used to train the Random Forest model. The data is first entered into a data frame with two columns years and kappa. Lagged features are created forming kappa lag 1, kappa lag 2 and kappa lag 3.

The Random Forest was trained with 100 estimators and a predetermined random state to guarantee reproducibility. The main objective was to produce forecasts for future periods therefore no specific set was necessary. The final three known $k_t$ values from the training data made up the starting input. The following $k_t$, which the trained model predicted for each anticipated year, was added to the lag sequence to create subsequent forecasts. This iterative process produced predicted mortality index values for each year from 2012 to 2022.

*Table 4. 5 Forecasted Kappa (2012-2022)*

| Year | Kappa Forecasts | Year | Kappa Forecasts |
|---|---|---|---|
| 2012 | 0.139273 | 2018 | 0.147070 |
| 2013 | 0.076462 | 2019 | 0.079082 |
| 2014 | 0.083448 | 2020 | 0.074081 |
| 2015 | 0.144287 | 2021 | 0.146535 |
| 2016 | 0.069857 | 2022 | 0.070392 |
| 2017 | 0.077001 | | |

The forecasted mortality indexes from 2012 to 2022 were used to obtain the mortality forecasts for these years in all the age groups in this study.

*Figure 4. 3 Random Forest Forecasts (2012-2022)*

| | year | age_0_to_4 | age_5_to_9 | age_10_to_14 | age_15_to_19 | age_20_to_24 |
|---|---|---|---|---|---|---|
| 0 | 2012 | 75.877627 | 7.313840 | 6.555398 | 10.364767 | 18.556818 |
| 1 | 2013 | 71.136511 | 6.957428 | 6.482432 | 10.241577 | 18.017698 |
| 2 | 2014 | 68.272751 | 6.641761 | 6.502834 | 10.258646 | 17.615108 |
| 3 | 2015 | 66.609146 | 6.355505 | 6.552306 | 10.329328 | 17.301824 |
| 4 | 2016 | 64.422376 | 6.098775 | 6.484581 | 10.230179 | 16.919122 |
| 5 | 2017 | 63.064935 | 5.879323 | 6.487372 | 10.225878 | 16.622674 |
| 6 | 2018 | 61.179827 | 5.690049 | 6.545422 | 10.303867 | 16.478456 |
| 7 | 2019 | 60.336062 | 5.512487 | 6.457904 | 10.156032 | 16.208955 |
| 8 | 2020 | 59.578615 | 5.330603 | 6.428932 | 10.109298 | 16.044274 |
| 9 | 2021 | 57.751267 | 5.187409 | 6.475742 | 10.171167 | 15.964296 |
| 10 | 2022 | 56.610749 | 5.041078 | 6.368105 | 10.000480 | 15.733970 |

### 4.4 Model Validation/ Model Fitness Tests

The forecasts from the two hybrid models were compared with the actual observed mortality rates from 2012 to 2022. The MAE, MSE and RMSPE were used.

*Table 4. 6 Forecasting Accuracy Comparison: LC-ARIMA vs LC-Random Forest*

| Age Group | Metric | ARIMA | Random Forest | Better Model |
|---|---|---|---|---|
| 0-4 | MAE | 22.729603 | 7.043024 | RF |
| | MSE | 531.945710 | 51.592178 | RF |
| | RMSPE% | 42.746166 | 13.390898 | RF |
| 5-9 | MAE | 2.056996 | 0.603541 | RF |
| | MSE | 4.580966 | 0.395840 | RF |
| | RMSPE% | 44.328594 | 13.052778 | RF |
| 10-14 | MAE | 0.415304 | 0.922381 | ARIMA |
| | MSE | 0.177305 | 0.874775 | ARIMA |
| | RMSPE% | 5.723657 | 12.538734 | ARIMA |
| 15-19 | MAE | 0.157537 | 0.4843994 | ARIMA |
| | MSE | 0.031664 | 0.336216 | ARIMA |
| | RMSPE% | 1.874688 | 6.275056 | ARIMA |
| 20-24 | MAE | 1.556289 | 1.091707 | RF |
| | MSE | 2.675532 | 1.237131 | RF |
| | RMSPE% | 10.104081 | 7.265891 | RF |

### 4.5 Findings and Discussion

This study compared the effectiveness of the LC-ARIMA and LC-Random Forest models in forecasting mortality rates. The LC-ARIMA hybrid model performed better than the Random

Forest hybrid model in the 10-14 and 15-19 age groups. The relatively smooth and linear trends in these age groups may have aligned more closely with the underlying assumptions of the ARIMA model.

For the 0-4, 5-9 and 20-24 age groups the LC-Random Forest model outperformed the ARIMA hybrid model. The Random Forest could detect intricate nonlinear patterns in mortality within these groups. The model found it difficult to predict mortality rates for the 10-14 and 15-19 age groups.

Following the prediction of mortality rates by the two hybrid models, the predicted mortality rates were verified against the actual mortality rates between the years 2012 and 2022. This is depicted graphically below for each group.

Plots of Actual versus Predicted Mortality

*Figure 4. 4 Forecast vs Actual Mortality rates for Age 0 to 4*

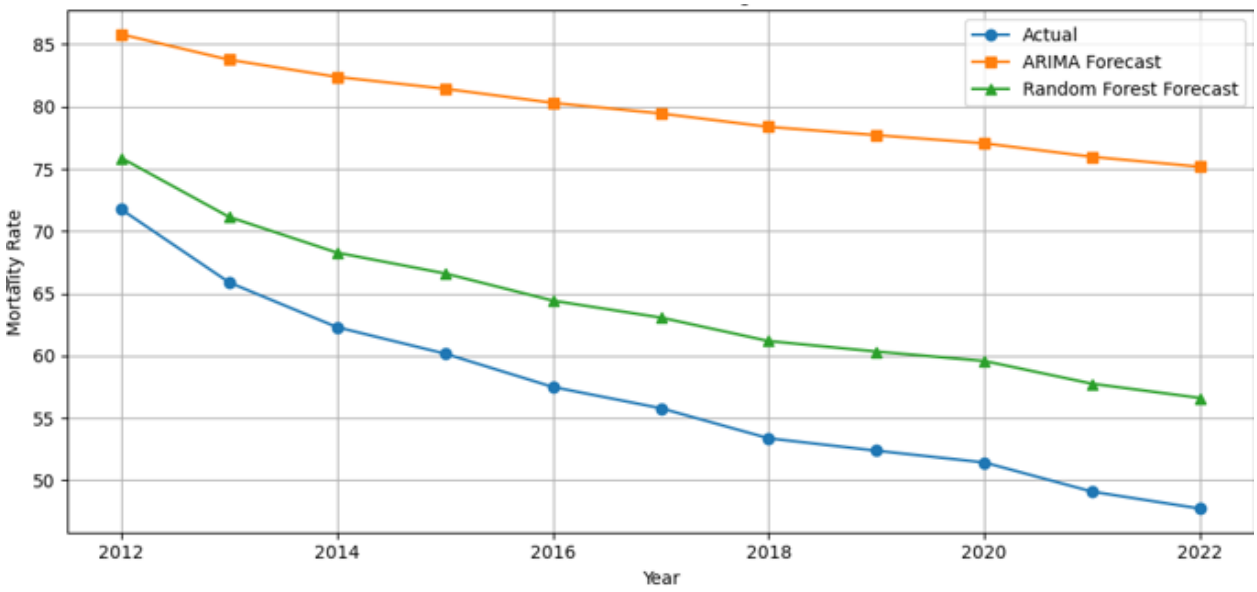


*Figure 4. 5 Forecast vs Actual Mortality rates for Age 5 to 9*

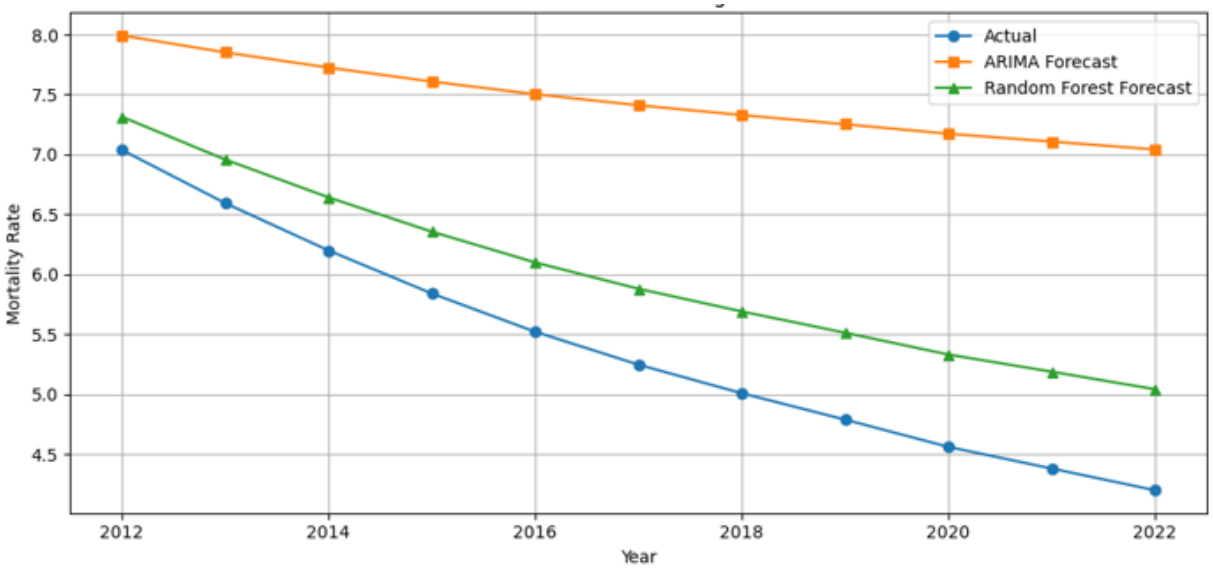



*Figure 4. 6 Forecast vs Actual Mortality rates for Age 10 to 14*

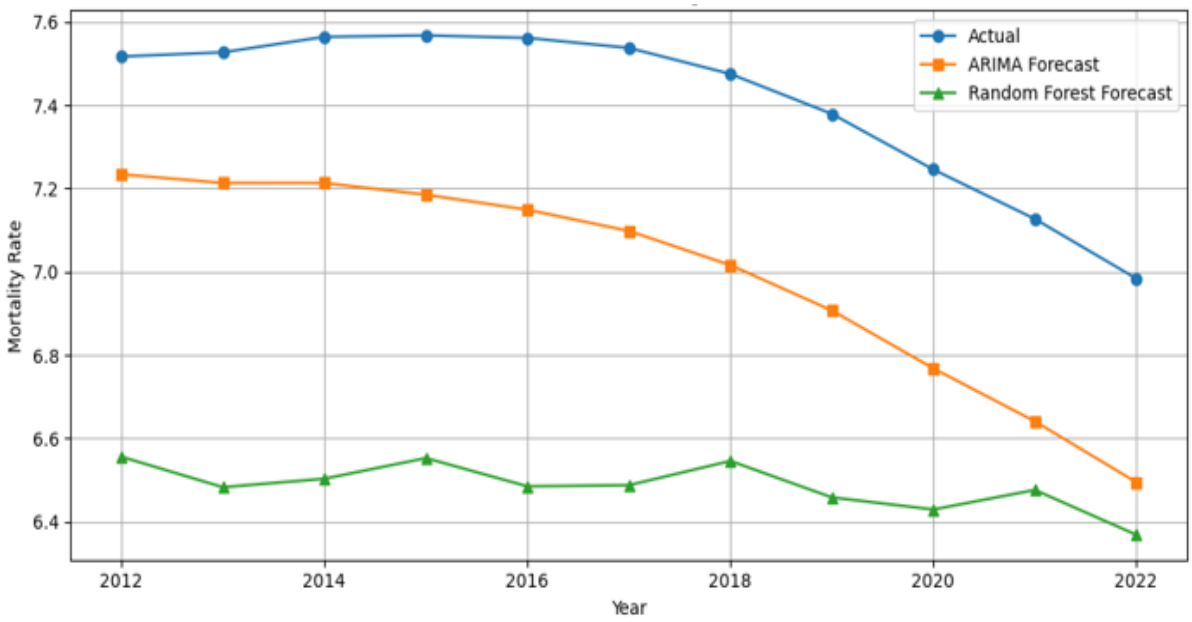


*Figure 4. 7 Forecast vs Actual Mortality rates for Age 15 to 19*

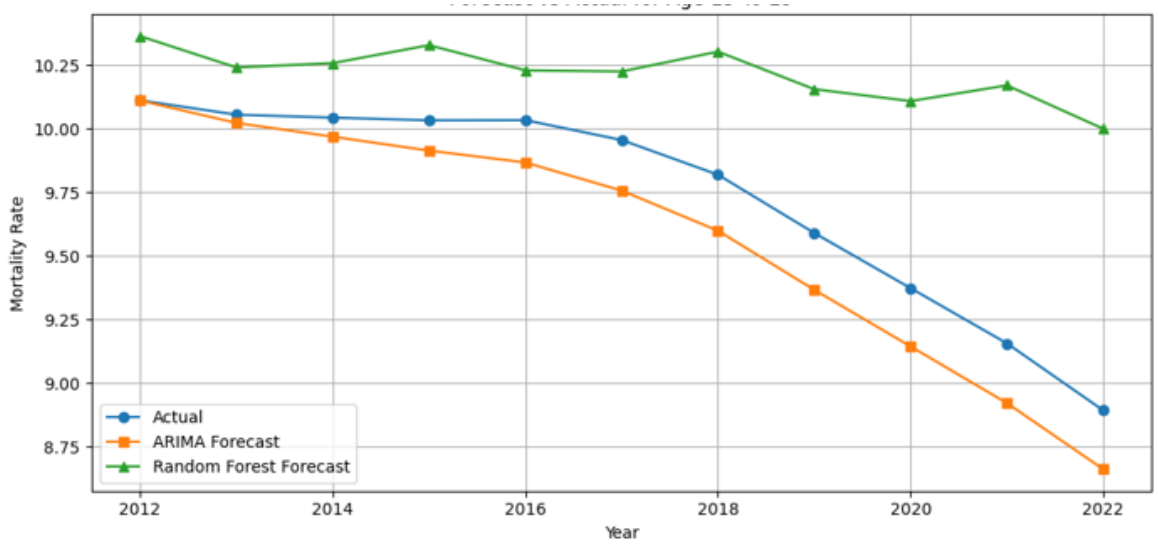
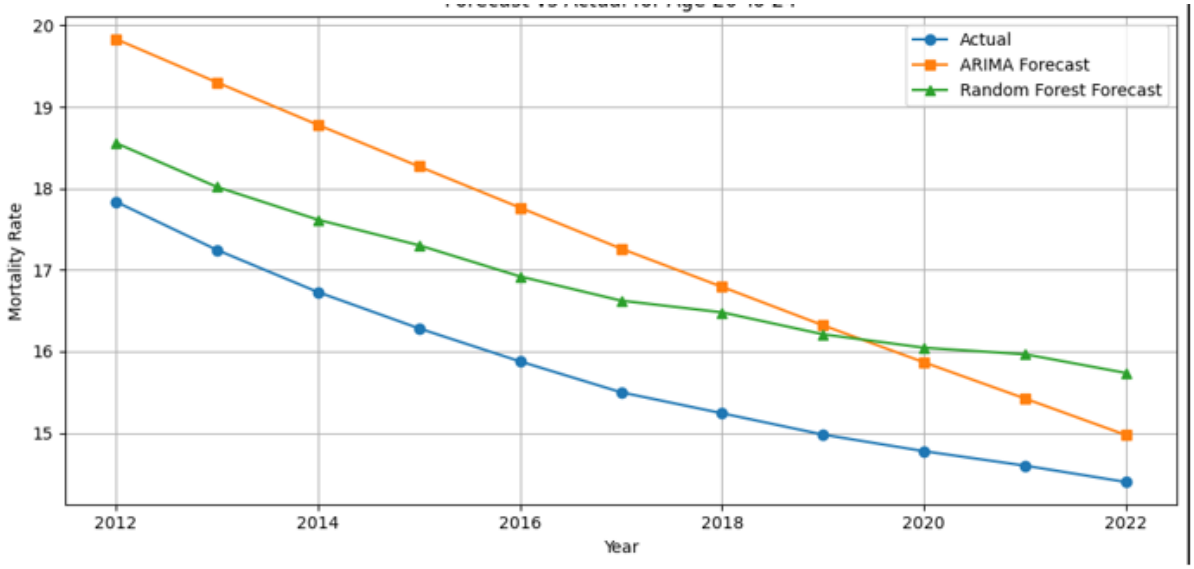



*Figure 4. 8 Forecast vs Actual Mortality rates for Age 20 to 24*




It was the aim of this study to not only compare model accuracy but to also describe how prediction results can be utilized to inform public health interventions. The LC-Random Forest model was most accurate in the prediction of mortality in the 0–4, 5–9, and 20–24 age groups. These results point to the need for increased investment in early childhood health interventions such as neonatal care, nutrition programs, and immunization campaigns. Similarly, the high risk

in the age group 20–24 years highlights the importance of attending to young adult health issues, including HIV prevention, mental health services, and injury control.

Conversely, the 10–14 and 15–19 age groups, where mortality trends were less volatile, were more accurately reflected by the LC-ARIMA model. This suggests that standard statistical models can still usefully inform policy for adolescent health, where trends are comparatively less volatile. The inferences from these projections can facilitate age-specific policy formulation and allow for more effective allocation of limited health resources.

Collectively, the interpretation of the model output provides evidence that hybrid forecasting tools can directly inform targeted public health responses in Zimbabwe.

### 4.6 Discussion of Results

The results emphasize that model performance varied by age bands in accordance with differences in patterns of mortality trends. LC-Random Forest was superior to LC-ARIMA in the 0–4, 5–9, and 20–24 age bands. These tend to be influenced by complex and non-linear mortality drivers such as malnutrition, neonatal conditions, and extrinsic causes such as HIV/AIDS and accidents. Random Forest, since the algorithm can capture non-linear interactions, was better suited for these patterns.

On the other hand, LC-ARIMA had a better fit for the 10–14 and 15–19 age groups that showed more stable and linear mortality trends. This is echoed by ARIMA's capacity to model stationary time series, suggesting its suitability for those environments characterized by smoother mortality profiles.

These findings validate the principle that model selection has to be data-based and age-dependent. For planning in public health purposes, this means interventions can be maximized through selecting the model that optimally describes the mortality profile of every age group. The hybrid approach thus provides a practical advantage through utilization of the virtues of both traditional and machine learning approaches.

### 4.7 Summary

This chapter showed a comparative analysis of two hybrid models with the Lee-Carter model as the base model. Historical data from 1990 to 2011 was used to obtain the parameters of the Lee-Carter model. Using this data the mortality index kappa was calculated using two different approaches Random Forests and ARIMA. The forecasted kappa was then used to forecast mortality rates from 2012 to 2022.

The LC-ARIMA hybrid model gave better results than the LC-Random Forest model for two of the age groups. The Random Forest model had the advantage of flexibility, and it successfully forecasted mortality rates for three of the age groups. This shows that the classical Lee-Carter model combined with ARIMA and Random Forest can be relied on when it comes to demography forecasting especially in Zimbabwe mortality profiles.

# CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS

## 5.0 Introduction

This chapter summarizes the findings from chapter 4. The chapter showed an analysis of the LC-ARIMA hybrid model and the LC-Random Forest hybrid model with a comparative basis. Chapter five provides research conclusions aligned to the objectives of the study. The researcher offers recommendations at the end of the chapter, addressing policymakers, those in the health sector and insurance sector.

## 5.1 Summary of the study

The introduction, background of the study, and the research questions of the study are provided at the beginning of the thesis chapter one. Chapter one consists of the study objectives to model and forecast mortality rates for ages 0 to 24 in Zimbabwe from 1990 to 2011 with LC-ARIMA and LC-RF hybrid models and to compare and assess the accuracy of LC-ARIMA and LC-RF hybrid models in predicting mortality for the specified age groups based on quantitative performance measures for the years 2012 to 2022.

Chapter two of the research addresses work by other researchers who have applied the Lee-Carter model. Pros and cons of the model when it comes to forecasting mortality rates are addressed in this chapter. The concept of the application of machine learning as a complement to the Lee-Carter model arose and was then followed by pointing out any research gaps regarding that field of research.

A detailed methodology for this study has been given in chapter three. In this way, chapter three showed that the Lee-Carter model, ARIMA and Random Forest model were to be used for the forecast of mortality rates. Motivation for the use of a hybrid modeling approach was also given in this chapter, showing how mortality modeling might benefit from the best of two different models. Finally, the chapter reflected on all those steps taken to ensure reliability and accuracy of the results.

In chapter four, the Lee-Carter model was used to obtain the parameters needed to model the mortality rates. ARIMA and Random Forest techniques were then used to obtain mortality

indexes and to forecast mortality rates. The two hybrid models were compared and assessed using error metrics.

## 5.2 Summary of the Findings

The two hybrid models successfully modeled the mortality rates. The parameters for the Lee-Carter model were successfully derived from the mortality rates. According to the study, the LC-Random Forest model performed better compared to the LC-ARIMA when it came to forecasting mortality rates since it successfully modeled three of the five age groups. The LC-ARIMA managed to provide better results than the machine learning method in the remaining age groups.

### 5.2.1 Objective 1

The first aim was to model Zimbabwean youths' mortality rates between the ages of 0 and 24 based on LC-ARIMA and LC-Random Forest models. This aim was accomplished by employing the Lee-Carter model to obtain the mortality index ($k_t$) for every year from 1990 to 2011. The mortality index was extracted by means of Singular Value Decomposition. The death index was recorded based on the time dynamics of Random Forest and ARIMA models. Age specific parameters ($\alpha_x$ and $\beta_x$) were attained from the model across the age groups. These parameters were used to reconstruct mortality.

### 5.2.2 Objective 2

The second objective was to evaluate and compare the accuracy of LC-ARIMA and LC-Random Forest hybrid models in forecasting mortality for the target age groups using quantitative performance metrics from 2012 to 2022. The two hybrid models were used to generate mortality forecasts from 2012 to 2022. Table 4.4.1 shows the error metric values for each of the hybrid models. Overall, the table shows that the LC-Random Forest model performed better than the LC-ARIMA model in three age groups (0-4, 5-9 and 20-24). The LC-ARIMA outperformed the LC-Random Forest in the 10-14 and 15-19 age groups.

### 5.2.3 Objective 3

The third objective sought to interpret age-specific mortality projections and derive implications for public health interventions. This was achieved through an analysis of which age categories each hybrid model forecasted more accurately and comparing these to possible real-world interventions. For instance, the LC-Random Forest model forecasted more accurately for mortality in age categories 0–4, 5–9, and 20–24, which are typically more subject to exogenous health and socioeconomic shocks. These results can inform public health actors to target intervention among these groups, for instance, through investment in neonatal care, school-age health interventions, and mental health in young adults. The trends of prediction thus provide practical advice for planning age-targeted health interventions in Zimbabwe.

## 5.3 Gaps and Limitations

Random Forest model could have been unable to provide more accurate predictions because not enough data was present between 1990 and 2022. The model could have overfitted in some age groups especially when the trend was linear leading to poor generalization. LC-ARIMA and LC-Random Forest models did not involve any socioeconomic or health-related factors

## 5.4 Project Constraints

The researcher could not obtain mortality rates for other years and age ranges from the national statistics bureau (ZIMSTATS) that would have been more accurate. Data for a wider age group and longer timespan would have caused the LC-Random Forest to perform better, which could not give more accurate mortality predictions despite being better than those of the LC-ARIMA predictions.

## 5.5 Recommendations

Several recommendations can be made for public health policy making and for future researchers. Public health authorities should invest in attaining age specific aggregated mortality data to enable

data driven research and policy making. Combining reliable mortality forecasting with national health planning can result in effective early warning systems and timely allocation of resources. Forecasts on the mortality of the younger generation can enable actuaries and insurance providers to develop life and health insurance for the youth. Based on these models, actuaries and insurers can further anticipate dependency ratios, healthcare expenses, and pension obligations.

# REFERENCES


1. Basellini, U., Camarda, C.G. & Booth, H., 2023. Thirty years on: A review of the Lee–Carter method for forecasting mortality. International Journal of Forecasting, 39, pp.1033–1049.
2. Biau, G. & Scornet, E., 2016. A random forest guided tour. Test, 25(2), pp.197–227.
3. Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5–32.
4. Chikanda, A. & Matanda, J., 2020. The impact of HIV/AIDS on the youth in Zimbabwe. African Journal of AIDS Research, 19(3), pp.217–225.
5. Cairns, A.J.G., Blake, D. & Dowd, K., 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal, 13(1), pp.1–35.
6. Cutler, D.R. et al., 2007. Random forests for classification in ecology. Ecology, 88(11), pp.2783–2792.
7. Czado, C., Nigri, A. & Rizzi, S., 2021. Machine learning and mortality forecasting: A comparative study. Demographic Research, 45, pp.1017–1040.
8. Gakidou, E., Cowling, K., Lozano, R. & Murray, C.J.L., 2010. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: A systematic analysis. The Lancet, 376(9745), pp.959–974.
9. Guelman, L., Guillén, M. & Pérez-Marín, A.M., 2015. A decision support system for mortality forecasting using machine learning techniques. Expert Systems with Applications, 42(3), pp.1356–1367.
10. Hong, W.H. et al., 2021. Forecasting mortality rates using hybrid Lee–Carter model, artificial neural network and random forest. Complex & Intelligent Systems, 7, pp.163–189.
11. Hyndman, R.J. & Athanasopoulos, G., 2018. Forecasting: Principles and Practice. 2nd ed. Melbourne: OTexts.
12. Haberman, S. & Renshaw, A.E., 2012. Parametric mortality forecasting models and the smoothing of mortality functions. Insurance: Mathematics and Economics, 50(3), pp.309–333.

13. Hanewald, K., Piggott, J. & Sherris, M., 2013. Population ageing and mortality improvement. Geneva Papers on Risk and Insurance - Issues and Practice, 38(3), pp.470–494.
14. Hyndman, R.J. & Booth, H., 2008. Stochastic mortality modeling. The Australian & New Zealand Journal of Statistics, 50(2), pp.165–183.
15. Kembo, J. & van Ginneken, J.K., 2009. Determinants of infant and child mortality in Zimbabwe: Results of multivariate hazard analysis. Demographic Research, 21, pp.367–384. Available at: http://dx.doi.org/10.4054/demres.2009.21.13 [Accessed 12 Mar. 2025].
16. Pedroza, C., 2013. A Bayesian forecasting model: Predicting US male mortality. *Biostatistics*, 14(1), pp.129–143.
17. Liu, L. et al., 2016. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet, 388(10063), pp.3027–3035.
18. Mavhandu-Mudzusi, A.H. et al., 2018. The impact of socio-economic factors on youth health in Zimbabwe. Journal of Public Health in Africa, 9(2), pp.123–130.
19. Ndlovu, R.S., 2018. Factors influencing infant and child mortality in Zimbabwe. University of the Western Cape.
20. Sauer, T. & Rau, R., 2018. Predicting death using random forests. Unpublished manuscript, 18 Sept.
21. Taruvinga, R. et al., 2017. Comparison of the Lee-Carter and ARCH in modelling and forecasting mortality in Zimbabwe. Asian Journal of Economic Modelling, 5(1), pp.11–22.
22. UN, 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Available at: https://sdgs.un.org/2030agenda [Accessed 12 Mar. 2025].
23. UNICEF, 2020. The State of the World's Children 2020. New York: UNICEF. Available at: https://www.unicef.org/reports/state-worlds-children-2020 [Accessed 12 Mar. 2025].
24. ZIMSTAT, 2016. Zimbabwe Demographic and Health Survey 2015. Available at: http://www.zimstat.co.zw [Accessed 12 Mar. 2025].
25. Levy, B.S. & Sidel, V.W., 2009. Social Injustice and Public Health. 2nd ed. New York: Oxford University Press.

26. Villegas, A.M., Millossovich, P. & Kaishev, V.K., 2018. Stochastic mortality modelling for population forecasts under regime shifts. Insurance: Mathematics and Economics, 78, pp.410–423.
27. Yu, W., Liu, T. & Valdez, E.A., 2017. Machine learning techniques in mortality modeling: A comparison. Risks, 5(4), p.57.

# APPENDIX 1

The following shows the code used to analyze mortality rate data using Google Colab's python.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

# Create a DataFrame
df = pd.DataFrame(data)
df.set_index('year', inplace=True)
print(df)

descriptive_stats = pd.DataFrame({
    'Mean': df.mean(),
    'Median': df.median(),
    'Min': df.min(),
    'Max': df.max(),
    'Std Dev': df.std(),
    'Q1': df.quantile(0.25),
    'Q3': df.quantile(0.75)
})


print(descriptive_stats.round(3))
```

```
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)
scaled_df = pd.DataFrame(scaled_data, columns=df.columns, index=df.index)
df = pd.DataFrame(data)
df.set_index("year", inplace=True)
min_value = df.min().min()
buffer = 0.001  # Small buffer
constant_to_add = abs(min_value) + buffer
df_scaled = df + constant_to_add

# Step 1: Convert to DataFrame
df = pd.DataFrame(data)

# Step 2: Set 'year' as index
df.set_index('year', inplace=True)

# Step 3: Log Transformation
log_mortality = np.log(df)

# Step 4: Calculate alpha (age-specific average log mortality)
alpha = log_mortality.mean(axis=0)

# Step 5: Center the data by subtracting alpha
centered_log_mortality = log_mortality - alpha
```

```
# Step 6: Apply SVD to the centered log mortality
U, s, Vt = np.linalg.svd(centered_log_mortality, full_matrices=False)

# Step 7: Extract first components for beta and kappa
beta = Vt[0, :]
kappa = U[:, 0] * s[0]

# Step 8: Normalize beta to sum to 1
beta = beta / np.sum(beta)

# Step 9: Adjust kappa accordingly and center to have mean zero
kappa = kappa * np.sum(beta)
kappa = kappa - np.mean(kappa)

# Step 10: Create DataFrames for parameters
alpha_df = pd.DataFrame(alpha, columns=['alpha'])
beta_df = pd.DataFrame(beta, index=df.columns, columns=['beta'])
kappa_df = pd.DataFrame(kappa.reshape(1, -1), columns=df.index, index=['kappa'])

# Step 11: Print Results
print("Alpha (Age-specific average log mortality):")
print(alpha_df)

print("\nBeta (Age-group sensitivity to mortality changes):")
```

```
print("\nKappa (Time-varying mortality index):")
print(kappa_df)

# Step 1: Log Transformation (if needed)
data['K_log'] = np.log(data['K'] - data['K'].min() + 1)  # Shift to avoid log(0)

# Step 2: First Differencing on Log Transformed Data
data['K_log_diff'] = data['K_log'].diff().dropna()

# Step 3: Second Differencing on Log Transformed Data
data['K_log_diff2'] = data['K_log_diff'].diff().dropna()

# Step 4: Check Stationarity of Log Differenced Data
adf_result_log_diff2, kpss_result_log_diff2 = check_stationarity(data['K_log_diff2'].dropna())
print('ADF Statistic (Log Second Differenced):', adf_result_log_diff2[0], 'p-value:', adf_result_log_diff2[1])

# Convert to DataFrame
data = pd.DataFrame(K, columns=['K'])

# Step 2: Log Transformation
data['K_log'] = np.log(data['K'] - data['K'].min() + 1)  # Shift to avoid log(0)

# Step 3: First Differencing
data['K_log_diff'] = data['K_log'].diff().dropna()
```

```
# Step 4: Second Differencing
data['K_log_diff2'] = data['K_log_diff'].diff().dropna()

# Step 5: Check Stationarity of Log Second Differenced Data
def check_stationarity(series):
    adf_result = adfuller(series)
    return adf_result

adf_result_log_diff2, kpss_result_log_diff2 = check_stationarity(data['K_log_diff2'].dropna())
print('ADF Statistic (Log Second Differenced):', adf_result_log_diff2[0], 'p-value:', adf_result_log_diff2[1])
# Step 6: ACF and PACF Plots
plt.figure(figsize=(12, 6))

# ACF Plot
plt.subplot(1, 2, 1)
sm.graphics.tsa.plot_acf(data['K_log_diff2'].dropna(), lags=10, ax=plt.gca())
plt.title('ACF Plot of Log Second Differenced Data')

# PACF Plot
plt.subplot(1, 2, 2)
sm.graphics.tsa.plot_pacf(data['K_log_diff2'].dropna(), lags=10, ax=plt.gca())
plt.title('PACF Plot of Log Second Differenced Data')

plt.tight_layout()
```

```
# Step 7: Fit ARIMA Model (Example: Replace p and q with identified orders)
p = 1  # Replace with identified AR order
d = 2  # Differencing order (second differencing)
q = 1  # Replace with identified MA order

model = ARIMA(data['K'], order=(p, d, q))
model_fit = model.fit()
print(model_fit.summary())

# Step 1: Fit the ARIMA(1, 2, 1) model
model = ARIMA(kappa_t, order=(1, 2, 1))
model_fit = model.fit()

# Step 2: Forecasting for 2012 to 2022
forecast_steps = 11  # 2012 to 2022
forecast = model_fit.get_forecast(steps=forecast_steps)
forecast_mean = forecast.predicted_mean
forecast_conf_int = forecast.conf_int()

# Step 3: Create a DataFrame for forecasted values
forecast_years = list(range(2012, 2023))
forecast_df = pd.DataFrame({
    'year': forecast_years,
    'kappa_forecast': forecast_mean
```

```
# Step 4: Print the forecasted kappa values
print(forecast_df)

# Forecasted Kappa values (from ARIMA model)
forecasted_kappa = np.array([
    0.006190, -0.034686, -0.078098, -0.123901, -0.171954,
    -0.222129, -0.274300, -0.328354, -0.384181, -0.441678, -0.500748
])

# Step 1: Calculate forecasted mortality rates using Lee-Carter model
mortality_rates_lee_carter = {}

for age_group in alpha_x.keys():
    mortality_rates_lee_carter[age_group] = [
        np.exp(alpha_x[age_group] + beta_x[age_group] * kappa) for kappa in forecasted_kappa
    ]

# Step 2: Create a DataFrame for the mortality rates
mortality_rates_lee_carter_df = pd.DataFrame(mortality_rates_lee_carter, index=range(2012, 2023))

# Step 3: Print the forecasted mortality rates
print(mortality_rates_lee_carter_df)

# Create a DataFrame
```

```
# Step 2: Create Lagged Features
def create_lagged_features(data, lags=3):
    for lag in range(1, lags + 1):
        data[f'kappa_lag_{lag}'] = data['kappa'].shift(lag)
    data.dropna(inplace=True)  # Remove rows with NaN values
    return data

# Add lagged features
data = create_lagged_features(data)

# Separate features (X) and target (y)
X = data.drop(columns=['year', 'kappa'])
y = data['kappa']

# Step 3: Train the Random Forest Model
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X, y)

# Step 4: Forecast for 2012 to 2022
forecast_years = np.arange(2012, 2023)
forecast_data = pd.DataFrame({'year': forecast_years})
# Initialize a DataFrame to store forecasted kappas
forecast_kappas = []
```

```
# Use the last 3 known kappas as the initial inputs
last_known_kappas = kappa_data[-3:]

for year in forecast_years:
    # Create lagged input from the last known kappas
    lagged_input = pd.DataFrame([last_known_kappas], columns=[f'kappa_lag_{i}' for i in range(1, 4)])

    # Forecast the next kappa
    predicted_kappa = rf.predict(lagged_input)[0]
    forecast_kappas.append(predicted_kappa)

    # Update the known kappas
    last_known_kappas = np.append(last_known_kappas[1:], predicted_kappa)

# Combine forecasted kappas into the DataFrame
forecast_data['kappa'] = forecast_kappas

# Print forecasted kappas
print("Forecasted Kappa (2012-2022):")
print(forecast_data)

# Calculate mortality rates for each age group and year
mortality_rates = pd.DataFrame({'year': forecasted_kappa['year']})
```

```
# Loop through each age group and calculate mortality rates
for age_group in alpha_x.keys():
    # Calculate ln(m_x,t) = alpha_x + beta_x * kappa_t
    ln_mortality_rate = alpha_x[age_group] + beta_x[age_group] * forecasted_kappa['kappa']

    # Convert ln(m_x,t) to m_x,t by exponentiating
    mortality_rates[age_group] = np.exp(ln_mortality_rate)

# Display the calculated mortality rates
print("Forecasted Mortality Rates (2012-2022):")
print(mortality_rates)

# Error metrics comparison
error_metrics = {}
age_groups = ["age_0_to_4", "age_5_to_9", "age_10_to_14", "age_15_to_19", "age_20_to_24"]

for age_group in age_groups:
    # ARIMA model metrics
    arima_mae = mean_absolute_error(actual_mortality[age_group], forecasted_arima[age_group])
    arima_mse = mean_squared_error(actual_mortality[age_group], forecasted_arima[age_group])
    arima_rmse = np.sqrt(arima_mse)

    # Random Forest model metrics
    rf_mae = mean_absolute_error(actual_mortality[age_group], forecasted_rf[age_group])
```

# APPENDIX 2

Mortality rate data obtained from the UNICEF database.

| age | 0 to 4 | 5 to 9 | 10 to 14 | 15 to 19 | 20 to 24 |
|---|---|---|---|---|---|
| 1990 | 80.3891 | 8.4445 | 4.8054 | 7.5121 | 14.4261 |
| 1991 | 84.1775 | 8.3933 | 4.8654 | 7.938 | 15.5039 |
| 1992 | 88.1863 | 8.3641 | 4.9449 | 8.3703 | 16.6547 |
| 1993 | 91.8857 | 8.3498 | 5.036 | 8.8096 | 17.865 |
| 1994 | 95.6638 | 8.3339 | 5.1395 | 9.2273 | 19.0828 |
| 1995 | 98.761 | 8.3257 | 5.2476 | 9.6331 | 20.2927 |
| 1996 | 100.142 | 8.3363 | 5.3683 | 10.0183 | 21.4207 |
| 1997 | 100.2128 | 8.3576 | 5.4953 | 10.3629 | 22.4141 |
| 1998 | 99.0299 | 8.3794 | 5.6313 | 10.656 | 23.2528 |
| 1999 | 96.8425 | 8.4326 | 5.7814 | 10.8952 | 23.8579 |
| 2000 | 94.4676 | 8.489987 | 5.953954 | 11.05685 | 24.20503 |
| 2001 | 92.5288 | 8.56886 | 6.133452 | 11.17585 | 24.30427 |
| 2002 | 91.2257 | 8.654164 | 6.337658 | 11.20967 | 24.12254 |
| 2003 | 90.2661 | 8.705526 | 6.545267 | 11.15477 | 23.72643 |
| 2004 | 90.4731 | 8.755755 | 6.753374 | 11.06051 | 23.21959 |
| 2005 | 91.3754 | 8.782592 | 6.970076 | 10.94887 | 22.60144 |
| 2006 | 93.6366 | 8.736558 | 7.155676 | 10.81633 | 21.95066 |
| 2007 | 94.4852 | 8.632428 | 7.308435 | 10.69757 | 21.28209 |
| 2008 | 93.3302 | 8.431092 | 7.411885 | 10.56303 | 20.53645 |
| 2009 | 90.2232 | 8.170743 | 7.46873 | 10.42944 | 19.81821 |
| 2010 | 85.3867 | 7.842523 | 7.510008 | 10.30601 | 19.11457 |
| 2011 | 79.9867 | 7.456319 | 7.519348 | 10.20118 | 18.4361 |
| 2012 | 71.7441 | 7.037364 | 7.517214 | 10.11209 | 17.83652 |
| 2013 | 65.8815 | 6.593729 | 7.527275 | 10.05584 | 17.24499 |
| 2014 | 62.2907 | 6.198847 | 7.564396 | 10.04416 | 16.72752 |
| 2015 | 60.1697 | 5.839784 | 7.567935 | 10.03357 | 16.28232 |
| 2016 | 57.4929 | 5.520516 | 7.561792 | 10.03434 | 15.87698 |
| 2017 | 55.7896 | 5.246018 | 7.537772 | 9.956116 | 15.49813 |
| 2018 | 53.3732 | 5.007682 | 7.475269 | 9.819549 | 15.24032 |
| 2019 | 52.3766 | 4.787394 | 7.379317 | 9.590238 | 14.9785 |
| 2020 | 51.4317 | 4.560103 | 7.2456 | 9.373231 | 14.77508 |
| 2021 | 49.0875 | 4.37943 | 7.126943 | 9.155967 | 14.59755 |
| 2022 | 47.7291 | 4.198438 | 6.983709 | 8.892184 | 14.39649 |