FITTING AND FORECASTING MORTALITY TREND IN NIGERIA: AN APPLICATION OF THE LEE-CARTER MODEL

By

Ajijola, Lukman Abolaji, Olamide Patrick and Chukwuemeka, Prosper Toochukwu

Department of Actuarial Science and Insurance, Faculty of Management Science, University of Lagos. Email: lajijola@unilag.edu.ng

ABSTRACT

In a multi-ethnic and diverse country like Nigeria, which has inherent health system challenges, there is a need for timely and accurate mortality forecasts for effective resource allocation, and policy and social program design. This research, therefore, applies the Lee-Carter model to Nigerian mortality data which is a popular mortality prediction technique. The Model is modified due to Nigeria's structure comprising both rural and urban areas with different age geographic mortality rates. The research population consists of historical and projected mortality data for Nigeria from 12/31/1950 to 12/31/2100, obtained from the United Nations (UN) data bank. The final dataset was split into two subsets: the training set (from 1952 to 2023) and the validation set (from 2024 to 2100). This split allowed for accurate performance evaluation of the models on unseen data. The study predicts mortality trends in Nigeria using the Lee-Carter model and its variations, such as ARIMA, SARIMAX, and Gradient Boosting Regressor (GBR). The Lee-Carter Model with Gradient Boosting Regressor (GBR) model significantly outperformed the other two models with an exceptionally low MSE of 4.25e-06.

Keywords: Lee-Carter, Prediction, Fitting, Mortality, ARIMA, SARIMAX, GBR

1. Introduction

Mortality forecasting is an important component of demographic research and informs public health planning and socio-economic development. Accurate future mortality projections are necessary for governments, health organizations, and regulators to develop effective healthcare policies or optimize resource allocation (for example, pensions and insurance), and the costs of these programs must remain sustainable. Knowledge of trends in mortality enables decision-makers to anticipate future population structures and necessary public health priorities should a potentially massive global outbreak occur. Precision mortality forecasts are vital to a functioning public health system, leading to increased preparedness and more efficient interventions (Lee & Carter, 1992).

Nigeria not only boasts of the highest population in Africa but also has the highest number of deaths from traffic accidents on the continent (World Health Organization, 2023). The country also has distinctive public health and mortality challenges, coupled with a vastly heterogeneous population experiencing unequal healthcare access, marked by socio-economic inequalities, as well as high rates of infectious diseases (National Bureau of Statistics, 2022). These conditions

create the complexity of modeling and predicting mortality accurately in Nigeria. Understanding mortality trends in Nigeria is crucial for health, economic, and broader social planning. Mortality projections play a vital role in public health strategies to prevent avoidable deaths, corresponding with and generating changes that can affect overall life expectancy (Cairns, Blake, & Dowd, 2006).

For Nigeria, mortality rates are more than statistics, forecasting mortality accurately has become necessary, and using advanced techniques capable of covering the multiple dimensions and uncertainties surrounding mortality such as high burdens of disease, and disparities in access to healthcare across Nigeria is not negotiable. This served as a motivation to innovate new methodologies to enhance the fitting and forecasting of mortality rates in Nigeria.

In Nigeria, mortality rates are more than statistics; they can be seen as a matter of life and death, or worse, of a healthy future. By contrast, while it is exactly in critical such cases for forecasting accurate mortality that Nigeria would become necessary to use advanced techniques capable of covering the multiple dimensions and changing nature, the same as what happens in everyday life. Given the diversity of communities, high burdens of disease, and disparities in access to care across Nigeria, too many independent variables unique to individual settings drive mortality trends for more simple models (National Bureau of Statistics, 2022).

The Lee-Carter model is one of the most popular methods in modeling and forecasting mortality data. It has become popular because it can replicate long-term mortality trends from historical data by decomposing the observed age-specific patterns and levels of mortality over time. The Lee-Carter model is widely used due to its simplicity and accuracy, particularly in developed countries where reliable mortality data is available and mortality trends are relatively consistent (Lee & Carter, 1992). However, the utilization of the Lee -Carter model in forecasting mortality in a developing country such as Nigeria where data is limited has not been fully explored. This study aims to fill the existing research gap by utilizing the Lee-Carter model to fit and forecast mortality rates in Nigeria. Through the modification of the LCM, the study seeks to enhance the forecasting ability of the Lee-Carter model. Better forecasts help policymakers make better decisions, plan more effectively, and ultimately save lives (Lee & Carter, 1992; Cairns, Blake, & Dowd, 2006).

2. Literature Review

To provide a better understanding of forecasting mortality, this study adopts the Epidemiological Transition Theory, Compression of Morbidity Theory, and Stochastic Theory of Mortality. The epidemiological transition theory represents one of the most widely used theories in the forecast of mortality rates. This model, introduced by Abdel Omran in 1971, explains the evolution of the disease pattern and the mortality rate as societies develop and modernize.

The Compression of Morbidity Theory describes that while the life expectancy of a population continues to rise, the end-of-life period during which people may be expected to be unhealthy

gets shortened. More specifically, the stagnation of morbidity (or serious diseases) becomes prolonged so that a person doesn't become unwell far towards the last days of his or her life. Similarly, the Compression of Morbidity Theory includes provisions that improvements in the population, whether from healthcare, risk factors, or preventive measures, may increase the healthy life span of individuals and eventually reduce the burden of chronic diseases and disability in old age.

The Stochastic Theory of mortality perceives rates of deaths as being stochastic and affected by many non-statistical variables like climate change, outbreaks, or self-care practices of individuals. Rather than deterministic models that postulate a specific line of change in the mortality rates, stochastic models relax these assumptions and introduce uncertainties about future outcomes. Globally, studies on mortality forecasting have been devoted to using the Lee-Carter model within and across countries. In their 1992 paper, Lee and Carter figuratively describe explaining the mortality component of the demographic disease models regarding the US population during 1933–1987, thereby illustrating the capability of the model for effective forecasting of mortality for different cohorts over different ages. The major uniqueness of this study was that it fully delinked the age-specific mortality from the time-related trends, which was a safety turn for other uses of the model.

Booth et al. (2002) expanded the use of the Lee-Carter model to Australian, Canadian, and Swedish mortality. The results showed the robustness of the model for different demographic regimes but indicated simultaneously that the calendarization of the model is necessary in parallel to express the peculiar regional and cultural courses of mortality. The study confirmed that the Lee-Carter model was applicable in the long-term mortality forecasts for countries of different socio-economic levels of development and public health infrastructure. Further studies, among others, Girosi and King (2008) discussed the application of the Lee-Carter model in projecting mortality in low- and middle-income countries. They modified the model to take into consideration more volatile mortality patterns witnessed in countries beset by frequent epidemics, conflicts, or economic instability. With modification, the Lee-Carter model could be applied in countries with rather less reliable or inconsistent mortality data.

Liu and Yu (2011) set up a backtesting methodology to evaluate the prediction performance of the Lee-Carter model. They propose to use the Kolmogorov-Smirnov test to assess how close the percentile histogram resembles uniform distribution, which can complement the assessment of probabilistic prediction. They also address two issues with implementing the Lee-Carter model: robustness and drift uncertainty. Quantile regression (QR) were proposed for robust parameter estimation of the model for time-varying index k_t. They use the bootstrap method to incorporate the drift uncertainty. Finally, they illustrate the proposed methods through examining the model performance on our simulated data as well as actual mortality data from different countries. The findings of the study suggest that the QR method improves the prediction performance of the Lee-Carter model and there exists evidence for trend changes in male mortality in the last century.

Melnikov and Romaniuk, (2006) compares the performance of three mortality models in the context of optimal pricing and hedging of unit-linked life insurance contracts. Two of the

models are the classical parametric results of Gompertz and Makeham, the third is the recently developed method of Lee and Carter (1992) for fitting mortality and forecasting it as a stochastic process. First, quantile hedging techniques of Föllmer and Leukert (1999) are applied to price a unit-linked contract with payoff conditioned on the client's survival to the contract's maturity. Next, the paper analyzes the implications of the three mortality models on risk management possibilities for the insurance firm based on numerical illustrations with the Toronto Stock Exchange/Standard and Poor financial index and mortality data for the USA, Sweden and Japan. The strongest differences between the models are observed in Japan, where the lowest mortality for the next two decades is expected. The general mortality decline patterns, rectangularization of the survival curve and deceleration of mortality at older ages, are well pronounced in the results for all three countries.

Akinkugbe et al. (2017) used the Lee-Carter model to project mortality rates into the future for arguably Nigeria's two largest and most diverse cities, Lagos and Kano. Using historical mortality data from the National Population Commission and the World Health Organization, this chapter uses the Lee-Carter model to project mortality rates for these urban centers in the near and distant future. The results reflect that, in Nigeria, the urban mortality rates were falling, basically due to an improvement in access to health care and the reduction of infant and maternal mortality.

Shelleng, Sule, Kajuru and Kabiru, (2022) used Nigeria mortality data from 2009 to 2020 to compares and contrasts how well the Lee-Carter and ARCH models performed. Singular value decomposition (SVD) method, Langrage multiplier test, and autoregressive conditional heteroskedasticity (ARCH) effects were examined. Five (5) different ARIMA and ARCH models were fitted together with their criteria, i.e., AIC and BIC in order to determine the best model for Nigeria mortality data. ARIMA (0,1,0) had the lowest AIC and BIC values, and was determined to be the best ARIMA model. The mortality index is then modelled using ARIMA (0,1,0) and plugged back into the Lee-Carter model to predict the future mortality rate. Also ARCH (1) turned out to be the best ARCH model among other candidate models. The performance of Lee-Carter model and ARCH model was compared using error measures. Results obtained revealed that the ARCH model had the minimum RMSE and MAPE when compared with the Lee-carter model, therefore it was concluded that the ARCH model performs better than the Lee-Carter model on Nigeria mortality data.

3. Material and Methods

This section outlines the methodology employed in analyzing and predicting mortality trends in Nigeria using the Lee-Carter model and its variations, such as ARIMA, SARIMAX, and Gradient Boosting Regressor (GBR).

3.1. Population of the Study, Sample Size and Sampling Technique

The research population consists of historical and projected mortality data for Nigeria from 12/31/1950 to 12/31/2100, obtained from the United Nations (UN) data bank. This data includes demographic information across various age groups, gender categories, and regions

within Nigeria. The population provides comprehensive information to develop reliable mortality forecasts (United Nations, 2019). The extracted data was prepared by cleaning and organizing it into a format suitable for statistical analysis and model application. The final dataset was split into two subsets: the training set (from 1952 to 2023) and the validation set (from 2024 to 2100). This split allowed for accurate performance evaluation of the models on unseen data.

Given the extensive temporal scope of the data, there is no direct need for sampling in the traditional sense, as the entire dataset represents the population of interest. This is referred to as complete enumeration or a population study, where every available data point is utilized to ensure comprehensive analysis. By leveraging all available data, this study seeks to maximize the accuracy of the mortality forecasts.

However, within the data, stratification occurs based on age groups, gender, and possibly geographical or socioeconomic factors. Stratified sampling is thus applied to ensure that mortality rates are analyzed across different segments of the population, particularly to observe differences in mortality trends between different cohorts. The primary stratification variables include:

- *Age Groups*: Mortality data is stratified across various age cohorts to capture the differing mortality rates for each group, as mortality patterns tend to vary significantly by age.
- *Gende*r: Data is segmented by gender to analyze mortality differences between males and females, which is essential in identifying gender-specific trends (Booth et al., 2002).
- *Time Periods*: The data is also stratified into smaller time intervals, such as decades, to understand changes in mortality trends over time.

The decision to use the entire dataset ensures that the analysis captures both short-term and long-term mortality trends. Given the diverse range of years and demographic variables included, using the entire data pool avoids potential biases that might arise from excluding certain time periods or population groups. Additionally, this approach allows for accurate calibration of the predictive models, especially for the Lee-Carter model and its variations (ARIMA, SARIMAX, GBR), as larger datasets typically yield more reliable predictions.

3.2. Model Specification

Lee Carter Model

It has now been well accepted that mortality needs to be projected to allow future mortality improvement to be taken into account in the evaluation of mortality contingent products. It is also important to acknowledge that mortality trends have shown a great deal of uncertainty in the past (Pitacco, 2004). To incorporate randomness into mortality dynamics, Lee and Carter introduced a simple yet powerful statistical model for fitting and projecting mortality. Under the Lee-Carter framework, the log of age-specific central mortality rates are described as

$$logm_{xt} = a_x + b_x k_t + \epsilon_{xt}, \quad (1)$$

where $x = 1, 2, 3, \ldots, n$ represent ages and $t = 1, 2, \ldots, t_0$ represent years. Hence, through the Lee-Carter decomposition, the mortality improvement over time can be summarized with two age factors a_x and b_x, and one time-varying index k_t. Here k_t represents the time series of the general level of mortality, while a_x describes the age profile averaged over time, and b_x determines how much, at each age, the mortality rate responds to the changes in k_t.

After the Lee-Carter model is fit to a selected data set, a_x's and b_x's are treated as constants and the values ofk_t are modeled by a time series. In the original paper of Lee and Carter (1992), it is suggested that an autoregressive integrated moving average, specifically, ARIMA (0,1,0), is the most appropriate model for k_t, even though in some cases other ARIMA models might be preferable. The ARIMA (0,1,0) is equivalent to a random walk with drift and can be written as follows:

$$k_t = k_{t-1} + c_1 + \xi_t, \quad (2)$$

where ξ_t is $N(0, \sigma^2)$, independently and identically distributed. The drift term of this random walk, c_1, and its standard deviation, σ, are estimated from $k_1, k_2, \ldots k_{t_0}$. Forecasts of future values of k_t $(k_{t_0+1}, k_{t_0+2}, \ldots)$ can then be recursively generated using formula (2). More specifically, to forecast the time-varying index at time $t_0 + n$ given the data available up tot_0, the following equation is used:

$$k_{t_0+n} = k_{t_0} + n \cdot c_1 + \sum_{j=1}^{n} \xi_j \quad (3)$$

The following algorithm were follow in building the model

1. **Log Mortality Transformation**: The death rates were converted into their logarithmic form to stabilize variance.
2. **SVD Decomposition**: Using singular value decomposition (SVD), the Lee-Carter model components a_x, b_x, and k_t were estimated, where a_x is the mean log mortality, b_x represents the age-related component, and k_t captures the time-dependent mortality trend.
3. **Time Series Forecasting**: The time-dependent component was forecasted using an ARIMA (1,1,0) model. This configuration was chosen based on the underlying structure of the data.

Auto Regressive (AR) Models.

An AR(p) (Auto Regressive of order p) model is a discrete time linear equations with noise, of the form

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t. \quad (4)$$

Here p is the order,$\alpha_1, \alpha_2, \ldots \alpha_p$are the parameters or coefficients (real numbers), ε_t is an error term, usually a white noise with intensity σ^2. The model is considered either on integers $t \in \mathbb{Z}$, thus without initial conditions, or on the non-negative integers $t \in \mathbb{N}$. In this case, the relation above starts from $t = p$ and some initial condition $X_0, \ldots, X_{p-1}$ must be specified. The simplest case of an AR(1) model is

$$X_t = \alpha_1 X_{t-1} + \varepsilon_t \tag{5}$$

Time Lag Operator.

Let S be the space of all sequences $(x_t)_{t\epsilon\mathbb{Z}}$ of real numbers. Let us define an operator $L: S \to S$, a map which transform sequences in sequences. It is defined as

$$Lx_t = x_t - 1, \text{ for all } t \in \mathbb{Z}. \tag{6}$$

We should write $(Lx)_t = x_t - 1$, with the meaning that, given a sequence $x = (x_t)_{t\epsilon\mathbb{Z}} \in S =$, we introduce a new sequence $Lx \in S$, that at time t is equal to the original sequence at time $t - 1$, hence the notation $(Lx)_t = x_t - 1$. For shortness, we drop the bracket and write $Lx_t = x_t - 1$, but it is clear that L operates on the full sequence x, not on the single value x_t.

The time lag operator is a linear operator. The powers, positive and negative, of the lag operator are denoted by L^k:

$$L^k x_t = x_{t-k}; \text{ for } t \in \mathbb{Z} \tag{7}$$

With this notation, the AR model reads

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \varepsilon_t \tag{8}$$

Moving Average (MA) Models.

A MA (q) (Moving Average with orders p and q) model is an explicit formula for X_t in terms of noise of the form

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{9}$$

The process is given by a (weighted) average of the noise, but not an average from time zero to the present time t; instead, an average moving with t is taken, using only the last $q + 1$ times. Using time lags we can write

$$\left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t. \tag{10}$$

Auto Regressive Moving Average (ARMA) Models.

An ARMA (p, q) (AutoRegressive Moving Average with orders p and q) model is a discrete time linear equations with noise, of the form

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t \qquad (11)$$

or explicitly

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}. \qquad (12)$$

We may incorporate a non-zero average in this model. If we want that X_t has average μ, the natural procedure is to have a zero-average solution Z_t of

$$Z_t = \alpha_1 Z_{t-1} + \cdots + \alpha_p Z_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}. \qquad (13)$$

and take $X_t = Z_t + \mu$, hence solution of

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} + \bar{\mu} \qquad (14)$$

with

$$\bar{\mu} = \mu - \alpha_1 \mu - \cdots - \alpha_P \mu. \qquad (15)$$

Autoregressive Integrated Moving Average (ARIMA) Models

An ARIMA (p, d, q) (AutoRegressive Integrated Moving Average with orders p, d, q) model is a discrete time linear equations with noise, of the form

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)(1 - L)^d X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t. \qquad (16)$$

It is a particular case of ARMA models, but with a special structure. Set $Y_t \coloneqq (1 - L)^d X_t$. Then Y_t is an ARMA (p, q) model

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) Y_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t. \qquad (17)$$

and X_t is obtained from Y_t by d successive integrations. The number d is thus the order of integration. An example of this is the random walk is ARIMA(0,1,0).

We may incorporate a non-zero average in the auxiliary process Y_t and consider the equation

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d X_t = \left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t + \bar{\mu} \qquad (18)$$

$$\bar{\mu} = \mu - \alpha_1\mu - \cdots - \alpha_P\mu.$$

Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

SARIMA Model is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality. The SARIMA $(p,d,q)(P,D,Q)_S$ model is expressed as follows:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d\,(1-B^S)^D\,y_t = \theta_q(B)\,\Theta_Q(B^S)\varepsilon_t \qquad (19)$$

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)\left(1-\sum_{k=1}^{p}\Phi_k B^{kS}\right)z_t = \left(1-\sum_{j=1}^{q}\theta_j B^j\right)\left(1-\sum_{l=1}^{Q}\Theta_l B^{lS}\right)\varepsilon_t \qquad (20)$$

Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) Model

SARIMAX Models is a SARIMA model with Exogenous Variables (X), called SARIMAX $(p,d,q)(P,D,Q)_S$, where X is the vector of exogenous variables. The exogenous variables can be modeled by a multiple linear regression equation which is expressed as follows:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \omega_t, \qquad (21)$$

where β_0 is a constant parameter and $\beta_1, \beta_2, \ldots, \beta_k$ are regression coefficient parameters of exogenous variables, $X_{1,t}$, $X_{2,t}, \ldots, X_{k,t}$ are observations of k exogenous variables corresponding to the dependent variable y_t; ω_t is a stochastic residual; i.e., the residual series that is independent of input series.

$$\omega_t = \frac{\theta_q(B)\,\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d\,(1-B^S)^D}\,\varepsilon_t \qquad (22)$$

The general SARIMAX model equation can be obtained by substituting Equation (21) into Equation (22).

$$y_t = \beta_0 + \sum_{i=1}^{k}\beta_i X_{i,t} + \frac{\theta_q(B)\,\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d\,(1-B^S)^D}\,\varepsilon_t \qquad (23)$$

Below are the key steps involved in constructing this model:

I. **Log Mortality Transformation:** The death rates were transformed into their logarithmic form to stabilize the variance in the data, a common practice when modeling mortality rates.

II. **Singular Value Decomposition (SVD):** The SVD was applied to decompose the log-transformed mortality data into three primary components: a_x, b_x and k_{t}t. The component a_x represents the mean log mortality, which captures the average mortality across all ages. The component b_x, reflects the age-related deviations, indicating how mortality varies with age. Finally, k_t represents the time-varying index that captures the trends in mortality over the years. Together, these components provide a detailed decomposition of mortality, accounting for both age and time variations.

III. **Time Series Forecasting:** The time-dependent component k_t is the key focus for future forecasting, and for this, the SARIMAX (Seasonal ARIMA with exogenous variables) model was utilized. Specifically, a SARIMAX $(1,1,0) \times (0,1,1,12)$ model was chosen. This model specification includes a first-order autoregressive term, a differencing component to handle trends, and a seasonal component that addresses yearly cycles in mortality data. The SARIMAX model's ability to handle seasonality ensures that recurring patterns within the data, such as periodic changes in mortality rates, are captured, enhancing the overall predictive power of the model.

IV. **Prediction of Future Mortality:** Once k_t was forecasted using the SARIMAX model for the future years (2024-2100), the predicted values were combined with the previously estimated a_x and b_x to generate future mortality rates. This was done by first predicting the future log mortality as $a_x + b_x \cdot k_t$ and then transforming the log mortality values back to their original scale using the exponential function. This provided the predicted death rates for the future, allowing for a more comprehensive view of mortality trends that accounts for both long-term shifts and seasonal fluctuations.

3.3. Gradient Boosting Regression

The Gradient Boost Regression algorithm starts by making an initial guess (prediction) which represents all the samples of the training dataset. The initial guess is equal to the mean value of all the samples of the dependent variable in the training dataset provided the loss function used is one-half mean squared error. This mean value is taken as a leaf node. The next step is to create a tree based on the errors made by the previous tree. The errors made by the previous tree are the differences between the actual values and the predicted values. These difference are also known as Pseudo Residuals. After the calculation of the pseudo residuals, a tree is built using the independent attributes of the dataset with any greedy approach like Gini index to predict the pseudo residuals instead of the actual values of the dependent variable. If the newly built tree contains a leaf which possesses more than one value then the output value of that particular leaf is the mean of those values provided that one-half mean squared error is taken as the loss function. After the creation of a new tree, all the previously built trees are combined with the new tree to form a single prediction model. Then the above mentioned procedure is iterated again to form a new tree based on the errors made by the previous prediction model until a certain number of trees are built or the creation of more number of trees no longer

improves the fitness of the model. Also to counter the variance in the model and for better prediction with a testing dataset, Gradient Boosting uses a leaning rate generally between 0 and 1 to scale the contribution of any newly built tree after the first leaf node in the prediction model.

The procedure for building this model was as follows:

1. **Initial Forecast with ARIMA**: Similar to the first model, the death rates were first predicted using the ARIMA-forecasted values in conjunction with the Lee-Carter model.
2. **Residual Calculation**: The difference between the actual and predicted death rates, referred to as the residuals, was calculated.
3. **Residual Modeling with GBR**: A Gradient Boosting Regressor was then trained on the residuals, using the year and APC as features, to capture the non-linear patterns missed by the initial ARIMA model.

4. Results

Descriptive Statistics

Table 4.1 below provides the descriptive statistics of the dataset, summarizing the central tendencies and variability of the variables. The average death rate across all years is approximately 14.09, with a standard deviation of 6.59, indicating a moderate spread around the mean. The minimum observed death rate is 7.95, while the maximum is 30.70. Similarly, the annual percentage change (APC) shows a negative mean of -0.76, suggesting a general decrease in death rates over time. However, the variability of the APC is higher, with a standard deviation of 1.0 and a range from -2.8 to 0.78.

Table 4.1: Descriptive Statistics of the Dataset

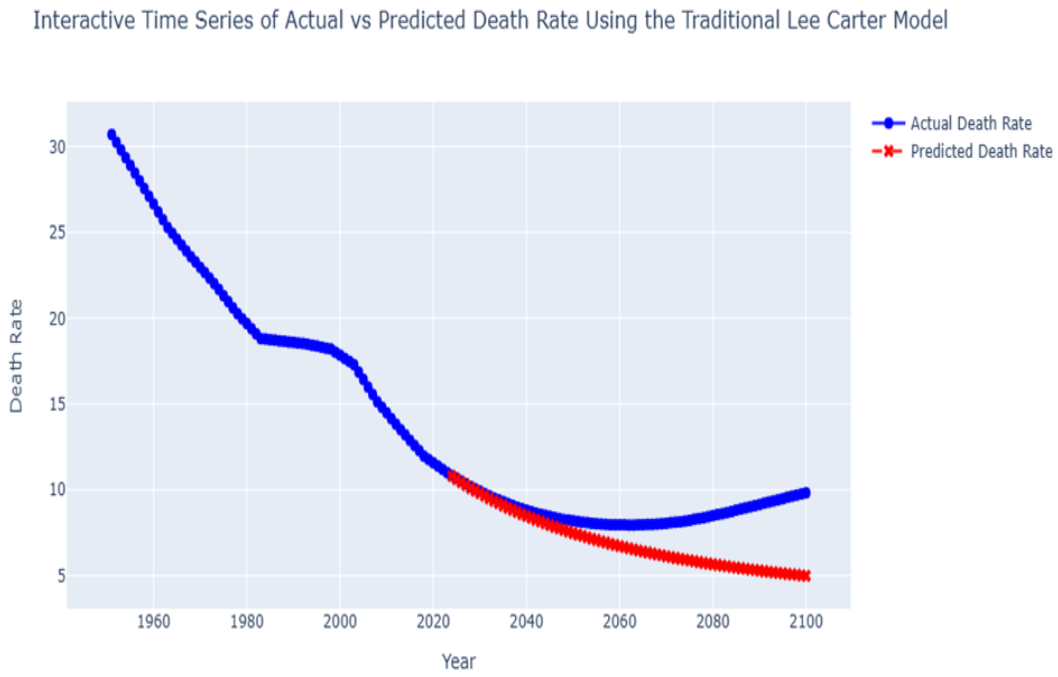
Statistic	Year	Death Rate	Annual Percentage Change (APC)
Count	150	150	150
Mean	2025.5	14.09	-0.76
Std. Dev	43.45	6.59	1
Min	1951	7.95	-2.8
25th Percentile	1988.25	8.56	-1.51
Median	2025.5	10.59	-0.95
75th Percentile	2062.75	18.65	-0.06
Max	2100	30.7	0.78

Author's Computation, 2025

Lee-Carter Model with ARIMA

The first predictive model applied was the traditional Lee-Carter model, which is frequently used in mortality forecasting. This model was enhanced with ARIMA for forecasting future mortality rates.

The future mortality rates were predicted by combining the forecasted k_t with the previously calculated a_x and b_x. Figure 4.1 below illustrates the performance of the Lee-Carter model with ARIMA. The blue line represents the actual death rates from the historical data (1952-2023), while the red line shows the predicted death rates from 2024 onwards.



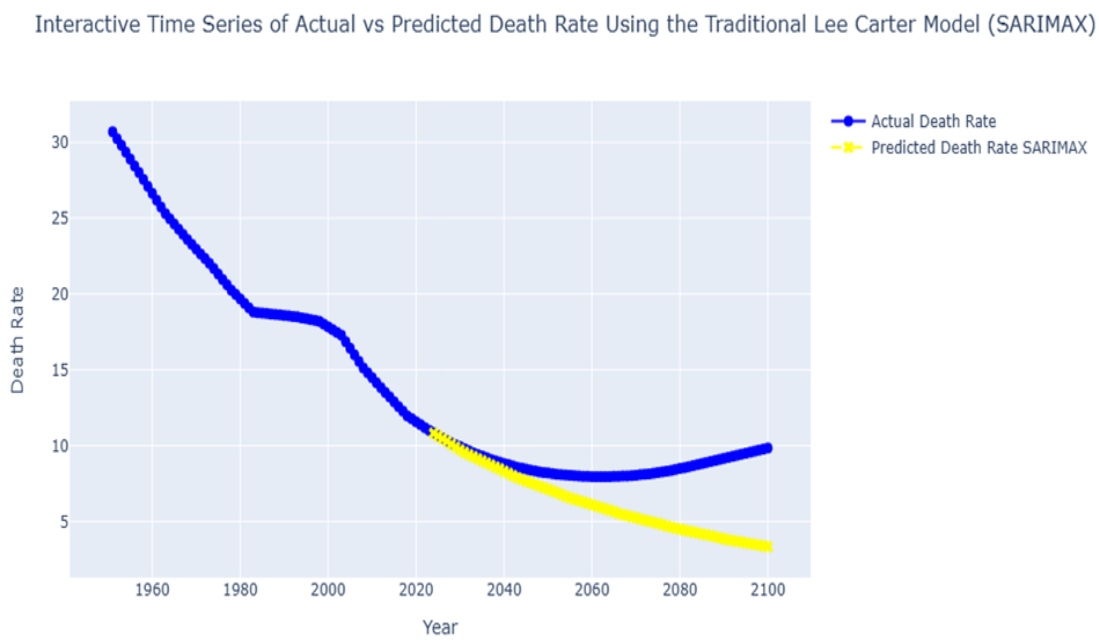
The model accurately captures the overall trend in mortality but demonstrates significant deviation after the year 2060. The observed discrepancy indicates that the ARIMA model, although effective in short-term predictions, struggled to generalize well over an extended period. The mean squared error (MSE) of the final predictions using the Lee-Carter model was calculated as 5.31, suggesting a considerable margin of error.

Lee-Carter Model with SARIMAX

The second predictive model applied was an enhanced version of the traditional Lee-Carter model, which incorporates seasonal effects using SARIMAX. This allows for capturing both the long-term trend and the seasonality in mortality rates.

Figure 4.2 illustrates the performance of the Lee-Carter model using the SARIMAX approach. The blue line represents the actual death rates derived from historical data (1952–2023), while the yellow line indicates the predicted death rates from 2024 onwards.

Figure 4.1: Actual vs Predicted Death Rate using Lee-Carter with SARIMAX



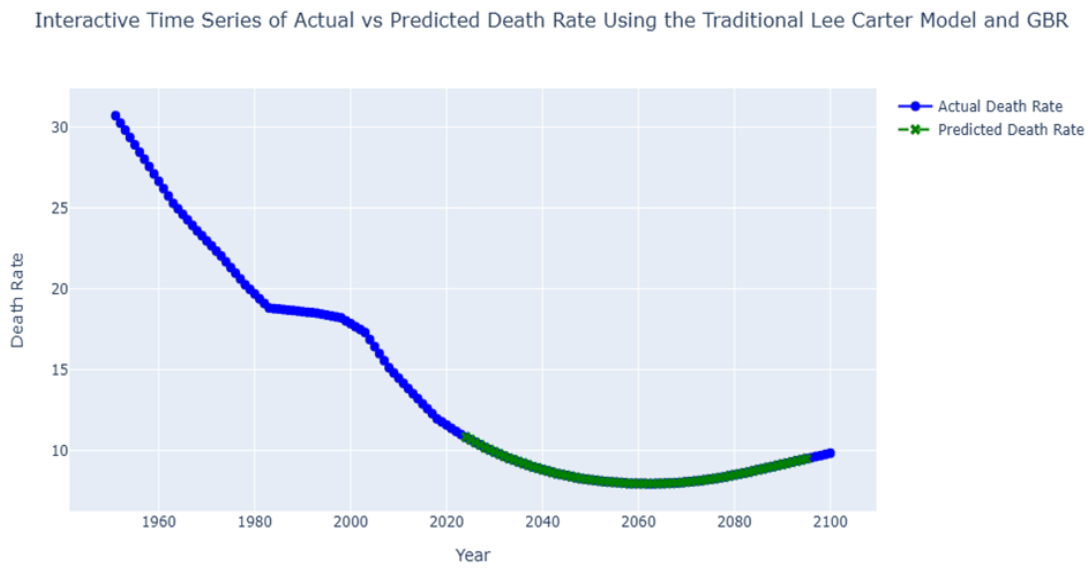
The SARIMAX model captures the overall declining trend in historical death rates, with a relatively good alignment between the actual and predicted rates up to around 2020. Beyond this point, the predicted death rates gradually begin to diverge from the actual data. Notably, after 2060, the divergence becomes more evident, as the model continues predicting a decline in death rates, while the actual data suggests fluctuations.

While SARIMAX demonstrates a more consistent and smoother long-term projection compared to the ARIMA model, the discrepancy after 2060 highlights the model's difficulty in accounting for long-term variations and cyclical patterns in mortality. This limitation is further emphasized by the Mean Squared Error (MSE) of **10.23**, indicating a higher margin of error compared to the ARIMA model, which had an MSE of 5.31. The larger MSE for SARIMAX suggests that despite offering better stability over time, it struggled with precise long-term forecasting.

Lee-Carter Model with Gradient Boosting Regressor

To improve the long-term forecasting accuracy, we combined the Lee-Carter model with Gradient Boosting Regressor (GBR) for the residuals that the ARIMA model and SARIMAX model could not predict effectively. This hybrid approach aimed to address the shortcomings of the ARIMA model and SARIMAX model in capturing non-linearities over time.

Figure 4.2: Actual vs Predicted Death Rate using Lee-Carter with GBR



The inclusion of the GBR model dramatically improved the model's ability to generalize over a long-time horizon, particularly beyond the year 2060. The final predictions closely align with the actual death rate trend, as seen in the gradual convergence of the lines. The MSE of the hybrid model dropped significantly to 4.25e-6, reflecting the enhanced performance. Figure 4.2 above shows the performance of the improved Lee-Carter model, where the predictions from the hybrid model (green line) are compared with the actual death rates (blue line).

4.1. Model Performance Comparison

The results of the two models reveal a clear distinction in their forecasting capabilities and the performance of these models was evaluated using the Mean Squared Error (MSE), which measures the average of the squared differences between the actual and predicted death rates. The table below summarizes the performance metrics of both models:

The Mean Square Errors of The Three Lee-Carter Models

Model	MSE
Lee-Carter with ARIMA	5.31
Lee-Carter with SARIMAX	10.23
Lee-Carter with GBR	4.25E-06

Author's Computation, 2025

A lower MSE value indicates better model performance, as it reflects a smaller error between the predicted and actual values proving the model's long-term predictive power by addressing the non-linear residuals.

The ARIMA-based model achieved an MSE of 5.31. This model captures the time series nature of the data, relying on the autoregressive and moving average components to forecast future death rates. However, its relatively high MSE indicates that it struggles to account for some of the non-linear patterns in the data.

Incorporating seasonal components into the model, the SARIMAX model yields an MSE of 10.23, which is higher than that of the ARIMA model. The inclusion of seasonal autoregressive terms likely introduced more complexity without improving the predictive accuracy, leading to an overfit to the seasonal noise present in the data.

The Lee-Carter Model with Gradient Boosting Regressor (GBR) model significantly outperformed the other two models with an exceptionally low MSE of 4.25e-06. The GBR leverages boosting techniques, which involve building multiple weak learners and combining them to form a strong learner, capturing the complex patterns in the data more effectively. The minimal error suggests that GBR excels at modeling non-linear relationships and produces highly accurate predictions for death rates.

5. Findings and Conclusion

In this study, the Lee-Carter model was employed to predict mortality rates from 1951 to 2100, with the period from 1952 to 2023 used for model training and the period from 2024 to 2100 designated for validation. To account for time-related mortality trends, the mortality data were transformed into a log format and decomposed using singular value decomposition (SVD), leading to the estimation of key components such as , and . These components represent the age-related mortality pattern and the time-varying trend, which were further forecasted using three different models: ARIMA, SARIMAX, and GBR.

The historical mortality data for Nigeria shows great variation in level of mortality amongst regions, ages and social classes. The general trend is that more people die in rural than urban areas as a result of poor health care accessibility, economic conditions and other aspects. The pace of mortality has been higher in the northern region of Nigeria as a result of poverty, war, and poor healthcare. Mortality rates rise and fall with age. Babies and the old aged societal members experience higher rates of death than the middle aged demographic. Factors such as social class also determine mortality. Those on the lower social class have more deaths than those on higher social classes due to less access to health care, education and even essential facilities.

The Lee-Carter model, which is particularly useful for forecasting mortality rates, while applied to the mortality data of Nigeria, required modifications to suit the peculiarities of the country's demographic data and socio-economic characteristics. However, the model was modified to account for Nigeria's rural-urban relations, age differentials and socioeconomic characteristics. In the calibration of the model, the parameters were adjusted to suit the particular mortality trends existing in Nigeria, where there are variations such as among different age groups and different geographical locations. These changes made it easier to

forecast the mortality patterns because the model was able to accommodate the diversity of the population in Nigeria.

The influencing factors that have been identified by the authors have an overall indication of the reduction of mortality rates over time especially in urban areas where healthcare as well as socioeconomic factors do improve. However, the pattern must change for low income groups as well lower regions where healthcare is inadequate the pattern must be unfavourable. Age-specific death rates among children below age five and preschoolers will continue to gradually improve as these public health trends will continue while within the aging population, age specific death rates will increase because of increased non-communicable disease burden and an increase in life span. The calibrated Lee-Carter model also suggests prospects for the equalization of loss of life in regions and social-economic stratifications but only on condition of political will.

The mortality projections have important consequences for the policy environment in Nigeria. In terms of health service delivery planning, the results, in this case, show the need for more resources to be directed towards improving health facilities in most rural regions and the provision of specific strategies to meet the needs of the high mortality regions. Health care expenditures should seek to narrow the rural-urban divide as well as resolve the socio-economic conditions that lead to increased death. Life insurers employ the mortality projections towards the re-calibration of life insurance rates and set reserves as the products offered today are postulated for future mortality trends. Social security programs including pensions will also need to be revised in the light of increasing life expectancy, especially amongst the elderly in such systems to be effective. The findings of this study underscore the effectiveness of the Lee-Carter model for mortality forecasting when integrated with different predictive techniques. While both ARIMA and SARIMAX provided reasonable predictions, their performances were overshadowed by the superior accuracy of the Gradient Boosting Regressor. This outcome reveals that traditional time-series models, though valuable, may not capture the complexities inherent in long-term mortality trends as effectively as machine learning approaches. The significantly lower MSE achieved by the GBR model emphasizes the potential of machine learning in improving mortality forecasting accuracy.

Moreover, this research highlights the importance of selecting the appropriate forecasting model for different types of data. The success of the GBR model suggests that machine learning models, which can manage non-linearity and incorporate feature interactions, can offer enhanced predictions compared to classical models. In public health and insurance industries, where accurate mortality predictions are crucial for policy planning and risk assessment, leveraging advanced models like GBR could lead to better decision-making and resource allocation.

References

Akinyemi, J. O., and Adebowale, A. S. (2017). Estimating the impact of mortality rate on life expectancy in Nigeria: A regional perspective. *African Population Studies, 31*(1), 3127–3145. https://doi.org/10.11564/31-1-1003

Booth, H., Maindonald, J., and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies, 56*(3), 325–336.

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance, 73*(4), 687–718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

Föllmer, Hans; Leukert, Peter (1999) : Efficient hedging: Cost versus shortfall risk, Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, No. 1999,18, urn:nbn:de:kobv:11-10056108 , http://hdl.handle.net/10419/61739

Girosi, F. and King, G. (2006). Demographic forecasting. *Cambridge University Press, Cambridge.*

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association, 87*(419), 659–671. https://doi.org/10.2307/2290201

Liu, X. and Yu, H. (2011). Assessing and extending the Lee-Carter model for long-term mortality prediction *Presented at the Living to 100 Symposium Orlando, Fla. January 5-7, 2011*. Retrieved from https://www.soa.org/globalassets/assets/files/resources/essays-monographs/2011-living-to-100/mono-li11-2b-liu.pdf

Makeham, W. M. (1867). On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries, 13*, 325–358. National Bureau of Statistics. (2022). *Nigeria: Statistical profile on health and mortality*. NBS.

Melnikov, A. and Romaniuk, Y. (2006). Evaluating the performance of Gompertz, Makeham and Lee-Carter mortality models for risk management with unit-linked contracts. December 2006 Insurance Mathematics and Economics 39(3):310-329. DOI: 10.1016/j.insmatheco.2006.02.012

Shelleng, A.U., Sule, Y.J., Kajuru, J.Y., and Kabiru, A. (2022). Comparative Study of Lee Carter and Arch Model in Modelling Female Mortality in Nigeria. UMYU Scientifica, 1(1), 241 – xxxx. https://doi.org/10.56919/usci.1222.012

United Nations. (2019). *World Population Prospects 2019: Highlights*. Department of Economic and Social Affairs, Population Division. https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

World Health Organization. (2023). *World mortality database*. Retrieved from https://www.who.int/data