TIME SERIES ANALYSIS OF PROSTATE CANCER INCIDENCES IN MERU COUNTY

Original Research Article

Abstract

Cancer is a major health challenge. Globally, the estimated number of diagnosed cancer incidences is approximately 14.1 million people per year and a mortality rate of 8.2 million deaths per year. The primary objective was to develop robust predictive models to forecast prostate cancer incidences and identify significant trends and patterns that inform healthcare planning and interventions in Meru County Kenya using AutoRegressive Integrated Moving Averarge with exogeneous variable (ARIMAX) Models.The dataset used comprised historical records of prostate cancer incidences in Meru County. The data spaned from [Jan 2018] to [Nov 2023], providing a comprehensive overview of the trends over time. Additionally, exogenous variable age was included in the ARIMAX model to enhance the accuracy of the prostate cancer predictions. Data on the prevalence of prostate cancer was obtained from Meru Cancer Registry for 71 months.The ARIMAX model was fitted using the Box-Jenkins methodology which include four iterative steps that is model identification,parameter estimation, diagnostics and forecasting.The prostate cancer time series data was made stationary by differencing and log transformation. R programming (Version 4.3.3) software was used in the analysis. Further, given the highly sensitive nature of the forecast values, interpolated data from daily values to monthly values were used. The best models for the Prostate cancer incidences was ARIMAX (0,0,1). Majority of the Prostate cancer incidences were within the age group 70-79 years at 50.7%,ages 60-69 was 42.3% while 80-90 years was 7%. After log transformation and and differencing of the prostate cancer time series data the Augmented Dickey Fuller test was performed and the p-value was (.01) which was less than the significance level of (.05), the null hypothesis was rejected that the prostate cancer time series had a unit root. Therefore, there was sufficient evidence to conclude that the time series was stationary. Ljung-Box test checked for the presence of autocorrelation at multiple lags and a high p-value = .719 greater than 0.05 indicated that there is no significant autocorrelation remaining in the residuals, thus the ARIMAX model was adequate. The MA(1) coefficient was -0.9, which indicated strong short-term negative autocorrelation. A positive value of 0.587 suggested that as the external variable increases by one unit, the log-transformed and differenced prostate cancer monthly cases (lnPCa Monthlycases d1) were expected to increase by 0.5871 units, holding all else constant. Results show that the ARIMAX(0,0,1) model slightly outperformed the ARIMA (0,0,1) model.This study successfully modeled the trends of prostate cancer incidences in Meru County using ARIMAX models. The findings indicated a rising trend in incidences, with the ARIMAX model providing the most accurate forecasts by incorporating the external variable age.

Keywords: ARIMA,ARIMAX, Stationary, Autocorrelation, Forecasting, Prostate Cancer

1 Introduction

Cancer is a major health challenge. Globally, the estimated number of diagnosed cancer incidences is approximately 14.1 million people per year and a mortality rate of 8.2 million deaths per year. Cancer is a leading cause of premature death for persons between the ages 30–69 years in 134 of 183 countries, [16].

The three most common cancer types with high age-standardized cancer incidence and mortality rates in 2022 were colorectal, lung, and prostate cancer, according to data from the Global Cancer Observatory. Approximately 10 million deaths, which translates to one in six deaths, were due to cancer [2].

Concurrently, cancer has a significant economic and financial burden to society, [17]. The increasing burden for cancer prevention and reduction in mortality is a major public health concern. Primary cancer prevention includes the interventions made to reduce the incidence of cancer while secondary prevention is the efforts made to reduce second cancers among cancer survivors.

Mathematical models are used in the modeling of disease interactions within populations. Mathematical modelling can improve understanding and lead to better policies for the adoption of measures that would compound higher health and economic benefits in the context of resource restrictions. Furthermore, policymakers rely heavily on mathematical models to help them with resource allocation and management measures. However, the strand of literature on cancer research, particularly with a focus on Kenya, remains thin.

As a result, whereas the majority of similar research focuses on developed nations, the current research filled a gap in the literature and is therefore focused on Meru County, a county that has received relatively little attention. Divergent tendencies make this county a fascinating topic for cancer research.

According to [16], 8.2 million people die from cancer each year and 14.1 million people are anticipated to be diagnosed globally. Cancer surveys in low-income countries and high-prevalence settings are typically cross-sectional and independently implemented about once every five years, in contrast to high-income countries where longitudinal studies, such as the National Health and Nutrition Examination Survey, provide nationally representative trend estimates for health outcomes.

When modeling cancer data, inappropriate or restrictive assumptions can have a negative impact on the outcomes, which can therefore make it more difficult to use those outcomes.It is against this backdrop that ARIMAX model which includes an external variable age was used. Precise forecasts of prostate cancer for upcoming periods are crucial for primary and secondary prevention. They are also essential for organizing future services and resource distribution related to prostate cancer, as well as for creating and assessing programs for prostate cancer control.

1.1 Review of Related Literature

According to [6] in a study on forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using Autoregressive Integrated Moving Average (ARIMA) models, the results indicated that among the various models evaluated, the model with one autoregressive term (coefficient=0.45, p=0.028) as well as a differenced series provided the best fit, with a Mean Average Percentage Error(MAPE) of 5.2% and an external validation showed a MAPE of 5.8%. The study projected prostate cancer incident cases in 2022 to rise to 25,283 cases (95%, Confidence Interval: 23,233 to 27,333).

[19], the ARIMA model was applied to estimate the number of malaria cases in Bhutan's endemic areas, and the ARIMAX model was further employed to identify the predictors (meteorological factors). For four of the seven districts that were the subject of the study, their results showed that the mean maximum temperature lagged at one month was a strong positive predictor of an increase in malaria cases.

According to [12], combining point regression and time series models provide projection performance that is comparable: They showed that ARIMAX distinguished out in their comparison utilizing World Health Organization historical cancer data, achieving the lowest percentage error in five out of seven cancer cases. In five of the six instances, ARIMAX produced the lowest MSEs or percentage errors when compared to a single weighted average. Their results showed that ARIMAX outperformed the other two techniques, which were average annual percentage change (AAPC) approaches and joint point regression [12].

[11] conducted a study to forecast the incidence rates of top three cancers in Malaysia. The aim was to determine the best model between Box-Jenkins ARIMA and exponential smoothing in forecasting the incidence rates. The model with the least Mean Absolute Percentage Errors (MAPE) value, was determined as the best model and used to forecast the top three cancer incidences for the year 2017 to 2021. Results showed that the Exponential Smoothing model predominantly outperformed the ARIMA model.

Scarlet fever incidence in China was predicted by [13] using data from the People's Republic of China's National Health Commission between January 2011 and August 2022. The results indicated that the average monthly incidence of scarlet fever was 4462.17 (SD 3011.75) cases, and that the annual incidence showed an upward trend until 2019. With mean absolute errors indicating reduced values, the ARIMAX models beat the ARIMA models and had better prediction capabilities.

Men over 50 years of age have a higher chance of developing prostate cancer, according to [5]. Prostate cancer risk increases with age. If caught early enough, prostate cancer is among the most curable types of cancer. Following the implementation of PSA screening about thirty years ago, there has been a notable decline in both the incidence and death associated with prostate cancer. [16] reports that the median incidence of prostate cancer in Africa was 19.47/100,000 people, whereas the overall pooled incidence was 21.95/ 100,000 people.Globally, the incidence rate of prostate cancer is increasing by 3% every year.

Prostate cancer, at 10%, was the most common type of cancer in men in Meru County, followed by esophageal cancer. These findings are consistent with a study by [14] that found prostate cancer to be the most common cancer in men over 65 in the United States. The study found that men over the age of 55 accounted for 84.21% of instances of prostate cancer. The biggest risk of prostate cancer is among older people.

Although the exact causes of prostate cancer are unknown [18], age and family history have been recognized as risk factors. In the United States, one in six men will eventually acquire prostate cancer, according to [18]. The incidence of prostate cancer in men over 50 is more than 30%, and in people over 80, it rises to almost 80%. according to the report, 96% of prostate cancer cases affect men who are 55 years of age or older.

2 Methodology

A descriptive cross-sectional design was used which utilized aggregated monthly prostate cancer cases. The secondary data was collected from the Meru Cancer Registry. Monthly data from January 2018 to November 2023 was obtained due to the need for uniformly and consistently measured data. The Meru Cancer Registry, situated within the Imenti North Constituency of Meru County, operates at coordinates approximately 00 02' 46" N latitude and 370 39' 21" E longitude.The sample included the prostate cancer incidences in Meru County within January 2018 to November 2023.

2.1 Model Development of Time Series Data

An ARIMAX model was built on the dependent variable in this case the incidence of prostate cancer. To determine the appropriateness of the models and to substantiate the validity of the proposed

modeling framework, the Akaike Information Criterion (AIC) and the Mean Absolute Percentage Error (MAPE) was used. An ARIMA(p,d,q) model is given as:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.1}$$

Where Y_t is a given time series and ε_t is a white noise process.

The ARIMAX model is expressed as;

$$Y_t = \beta X_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.2}$$

Where Y_t is the incidence of prostate cancer.
Where X_t is a co-variate at time t and β is its coefficient.
The X_t represents exogenous/external variable (X_1 = age)

Parameter estimation and model identification are the two stages of the ARIMAX model development process. The order of the autoregressive (AR) component, p, is the degree of differencing of the original prostate cancer time series data, d is the degree of differencing, and q is the order of the MA component. The first step involved estimating the non-seasonal (p,d,q) and seasonal (P,D,Q) parameters [1]. The autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to determine these. The unknown coefficients for the AR and MA components of the ARIMA model were computed during the parameter estimation phase.

The prostate cancer time series data is then split into a training and testing set using an 80:20 ratio.The ARIMAX (0,0,1) model was generated using the auto.Arima() function models where the best ARIMA model based on Akaike information criterion (AIC) is obtained [3] . The series was tested through the Augmented Dickey-Fuller test. Differencing of the ARIMAX series was conducted. Based on the final selected model, the annual number of cases expected to be diagnosed in Kenya from 2024 to 2026 was forecasted. The 95% confidence intervals was calculated from the mean square errors of the ARIMAX model.

The ADF test is used to determine whether a time series is stationary by testing for the presence of a unit root. The ADF test is an extension of the Dickey-Fuller test, which accounts for higher-order autoregressive processes. The test involves estimating the following regression:
The model equation $\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \ldots + \delta_p \Delta y_{t-p} + \epsilon_t$ represents a time series model, where:

- Δy_t is the first difference of y_t ($y_t - y_{t-1}$).
- α is a constant.
- β is the coefficient on a time trend
- γ is the coefficient on y_{t-1}.
- δ_i are the coefficients on the lagged differences of y_t.
- ϵ_t is the error term.

Null Hypothesis H_0: The series has a unit root, alternative Hypothesis H_1: The series does not have a unit root.

2.2 Model Identification of Time Series Data

The model was specified and selected by plotting the ACF and PACF at different lags [8]. The model are initially identified by plotting the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) of the prostate cancer time series data.The Autocorrelation plot was used to obtain the order of the MA process, while the Partial Autocorrelation plot was used to obtain the order of the AR process [7].

The data has to satisfy the stationarity condition that is the mean, variance and autocorrelation have to be time invariant.

Table 1: Model Identification using the ACF and PACF

ACF/PACF	AR	MA	ARMA
ACF	Tails off	Cuts off at lag q	Tails off
PACF	Cuts off at lag p	Tails off	Tails off

2.3 Parameter Estimation of Time Series Data

In order to estimate the ARIMAX model, the Maximum Likelihood Method (MLE) was used. With the assumption of identically and independently distributed ε_t, the Log Likelihood (LL) function of y_t for t observations sample [4].

For an MA(1) model,

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}, \tag{2.3}$$

where ϵ_t are identically and independently distributed (i.i.d) normal errors with mean 0 and variance σ^2.

Given a time series $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the likelihood function, assuming $\epsilon_0 = 0$, is:

$$L(\mu, \theta, \sigma^2 \mid \mathbf{y}) = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \mu - \theta(y_{t-1} - \mu))^2}{2\sigma^2}\right). \tag{2.4}$$

Taking the natural logarithm, the log-likelihood function is:

$$\ell(\mu, \theta, \sigma^2 \mid \mathbf{y}) = -\frac{n-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \mu - \theta(y_{t-1} - \mu))^2. \tag{2.5}$$

The MLE estimates for μ, θ, and σ^2 are obtained by maximizing the log-likelihood function:

$$(\hat{\mu}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}}) = \arg\max_{\mu,\theta,\sigma^2} \ell(\mu, \theta, \sigma^2 \mid \mathbf{y}). \tag{2.6}$$

2.4 Model diagnostics of Time Series Data

In model diagnostics the models adequacy was determined. One of the assumptions is that the residuals/ model errors are white noise. Ljung-Box statistic was used to check if a given series is linearly independent. The test examines the null hypothesis of linear independence of the series and whether the residual series is a white noise series. When the Ljung-Box test P value is greater than .05, the residual series is white noise series, that is, the effective part of the original series is extracted sufficiently and the established model is valid. The diagnostic checking involves the analysis of the residuals by plot of the standardized residuals, the ACF of the residuals, and the p-values for Ljung-Box Q statistic. At this stage, the assumptions of the ARIMAX model are checked, such as the hypothesis of errors being independently and normally distributed.The ARIMAX(0,0,1) had a good fit and passed the residuals Ljung-Box test.

The Ljung-Box statistic Q is calculated as:

$$Q = n(n+2)\sum_{j=1}^{m}\frac{\hat{\rho}_j^2}{n-j}$$

where:

- n is the number of observations.
- m is the number of lags being tested.

- $\hat{\rho}_j$ is the sample autocorrelation at lag j.

The Q statistic follows a chi-squared (χ^2) distribution with m degrees of freedom under the null hypothesis.While fitting the time series models(ARIMAX) to the data we compute the sample autocorrelations of the residuals up to the specified lag m. Then the Computed Ljung-Box Q statistic is compared with the critical value from the chi-squared distribution with m degrees of freedom.

If the Q statistic is greater than the critical value, reject the null hypothesis, indicating that there is significant autocorrelation in the residuals. If the Q statistic is less than the critical value, fail to reject the null hypothesis, indicating that the residuals are independently distributed.If the test rejects the null hypothesis, it suggests that the residuals are not independently distributed and that there is significant autocorrelation. This indicates that the model may be inadequate and that additional lags or different model specifications might be needed. If the test fails to reject the null hypothesis, it suggests that the residuals are independently distributed and that the model is adequate with respect to capturing the time series dynamics. A high p-value (greater than .05) suggests that there is no significant autocorrelation remaining in the residuals, indicating a good model [15].After the ARIMAX model passing the tests, we proceeded to the prediction.

2.5 Forecasting of Time Series Data

Forecasting is making predictions of values whose real results have not yet been observed [10]. The model with the lowest MAE, MAPE, RMSE, or MSFE is considered to be the best. A perfect forecast would have MAE=MSE=RMSE, MSFE=0. As the value increases, the model's predictive power decreases [9]. The smaller the value, the better the forecast.

The ARIMAX (p, d, q) model equation for time series Y_t and exogeneous data X_t is;

$$\triangle Y_t = \varepsilon_t + \sum_{i=1}^{p} \triangle \psi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{m=1}^{M} \beta_m X_{t-m} \tag{2.7}$$

where, ψ_1,$\cdots$, ψ_p and θ_1,$\cdots$, θ_q are the parameters; ε_t ,$\varepsilon_{t-1}$are white noise error and β_1,$\cdots$, β_m are the parameters of independent variables input Y_t and time t.

The equation for RMSE, MAE, and MAPE are given by:

$$MAE = \frac{1}{2} \sum_{i=1}^{n} |P_i - O_i| \tag{2.8}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} |\frac{O_i - P_i}{O_i}| * 100 \tag{2.9}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2} \tag{2.10}$$

where O_i is the observed value, P_i is the predicted value and n is the number of observations [].

2.6 Back Transformation of Time Series Data

Back transformation of time series data after log transformation and differencing is a critical step in time series analysis, particularly when dealing with non-stationary data or when applying transformations to stabilize variance or achieve linearity. This process involves reverting the transformed data back to its original scale to interpret the results or make forecasts in the original units.To revert the log-transformed data back to its original scale, exponentiation was applied to each observation.To revert the differenced data back to its original scale, cumulative sum (integration) was applied to the differenced series [1].

3 Results and Discussion

3.1 Age distribution of the prostate cancer patients

According to the American cancer society about 6 in 10 cases are diagnosed in men aged 65 or older, and the average age at the time of diagnosis is around 66 years.

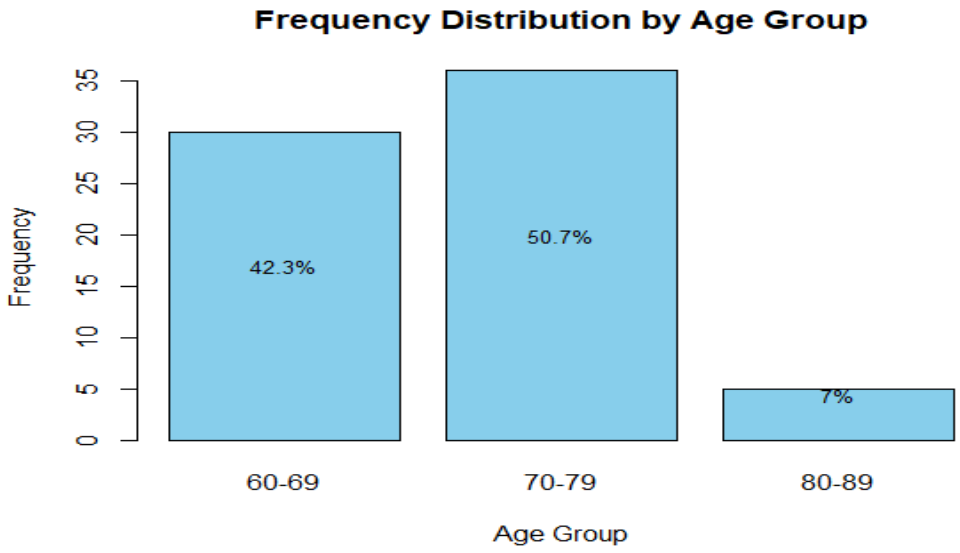


Figure 1: Age Groups

Majority of the cases were within the age group 70-79 years at 50.7% age 60-69 was 42.3% while 80-90 years was 7%. Age influences the screening and detection of prostate cancer. Screening tests such as the prostate-specific antigen (PSA) test are more commonly recommended for men aged 50 and older, or earlier for those at higher risk due to family history or other factors. Increased screening in older age groups led to more diagnosis of prostate cancer cases, [12].

3.2 Time series plot for Prostate Cancer Cases

Time series plots display observations over time.

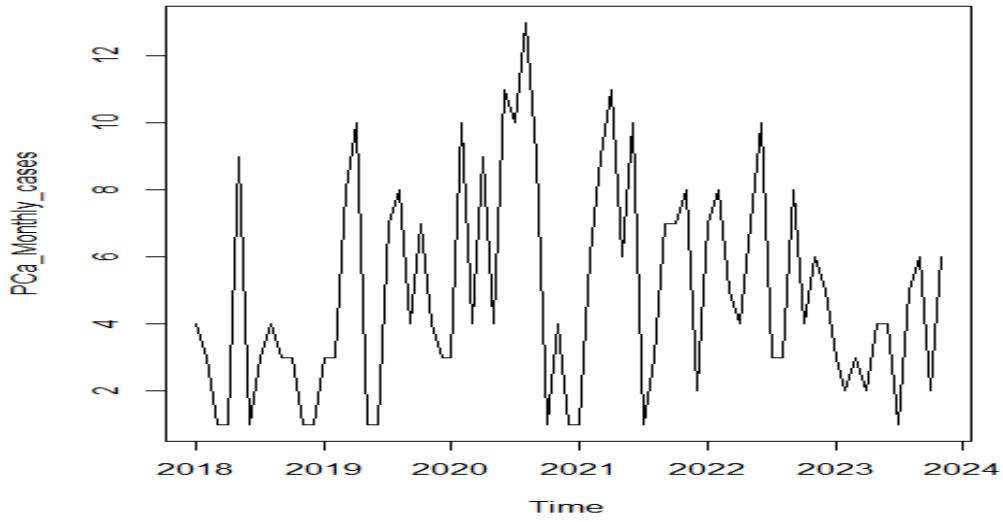


Figure 2: Time Series Plot of prostate cancer incidences

According to Figure 2, the prostate cancer cases ranged from 1 to 12 monthly cases in Meru County. The time series plot was fairly centered around the mean value of the number of prostate cancer incidence cases.Prostate cancer rates showed an increasing upward trend over the years.An essential aspect of selecting suitable modeling and forecasting techniques involves examining the patterns displayed in time series plots. These patterns typically stem from four main sources of variation within time series data: seasonal variations, trend variations, cyclic changes, and residual/ irregular fluctuations.

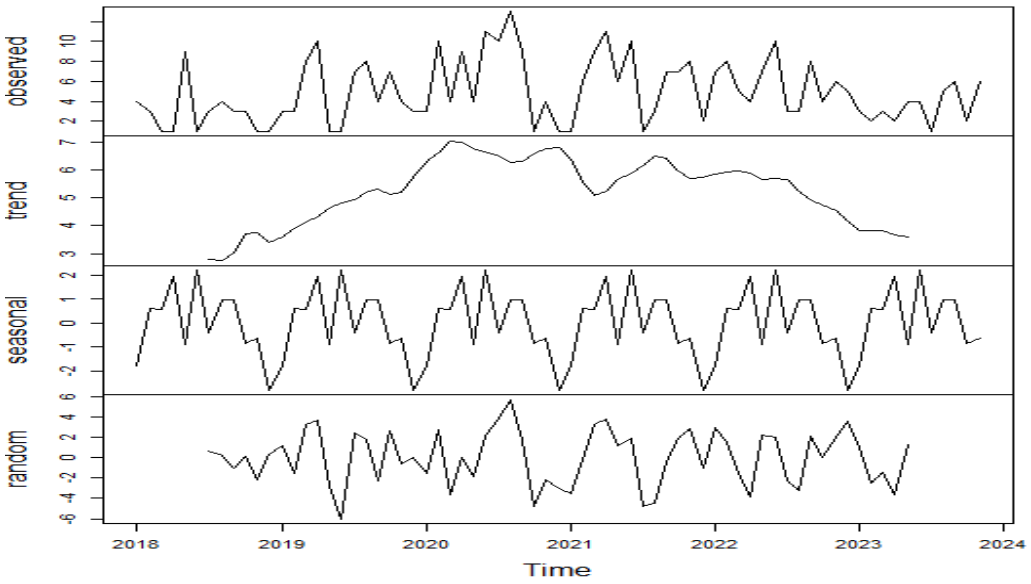


Figure 3: Model Decomposition prostate cancer incidence data

The time series of prostate cancer reported cases was plotted to observe long-term trends from January 2018 to November 2023. Data stationarity was tested using the augmented Dickey-Fuller (ADF). the non-stationary sequence was transformed into stationary sequences by difference and log transformation. Subsequent natural logarithm transformation and first differencing rended the modified time series stationary, as evidenced by the test results.

Table 2: Augmented Dickey-Fuller Test of Differenced log transformed prostate cancer incidences data

Dickey-Fuller value	Lag order	p-value
-6.4612	4	0.01

Since the p-value (0.01) is less than the significance level (typically 0.05), we reject the null hypothesis that the time series has a unit root. Therefore, we have sufficient evidence to conclude that the time series is stationary.It suggests that the Differenced log transformed prostate cancer incidences time series data in Meru County does not exhibit a unit root and is stationary, which is a prerequisite for fitting ARIMAX model effectively. This implies that trends and patterns observed in the data are likely to be reliable and are not due to non-stationarity, hence proceeding with proceed with ARIMAX modeling knowing that the basic assumption of stationarity is satisfied.

3.3 Model Identification for Prostate Cancer Incidence Data

The correlogram (Figure 4) shows the ACF of the differenced-log transformed data.

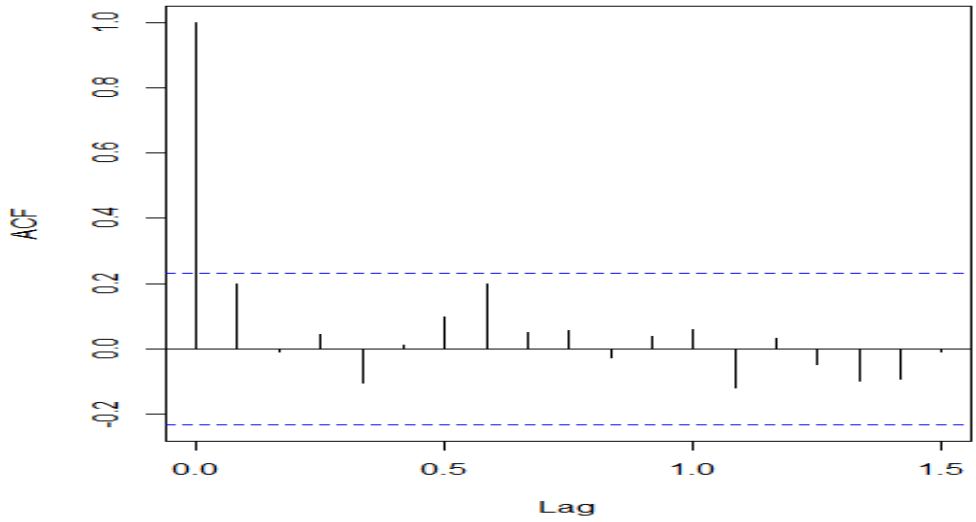


Figure 4: Autocorrelation of Differenced log transformed prostate cancer incidences data

According to the figure 4 a correlogram of the Autocorrelation function of the Difference log transformed prostate cancer incidences data the ACF had a significant spikes at lag 1, indicating a non seasonal MA(1) component. The correlogram (Figure 5) shows the PACF of the differenced-log transformed data.

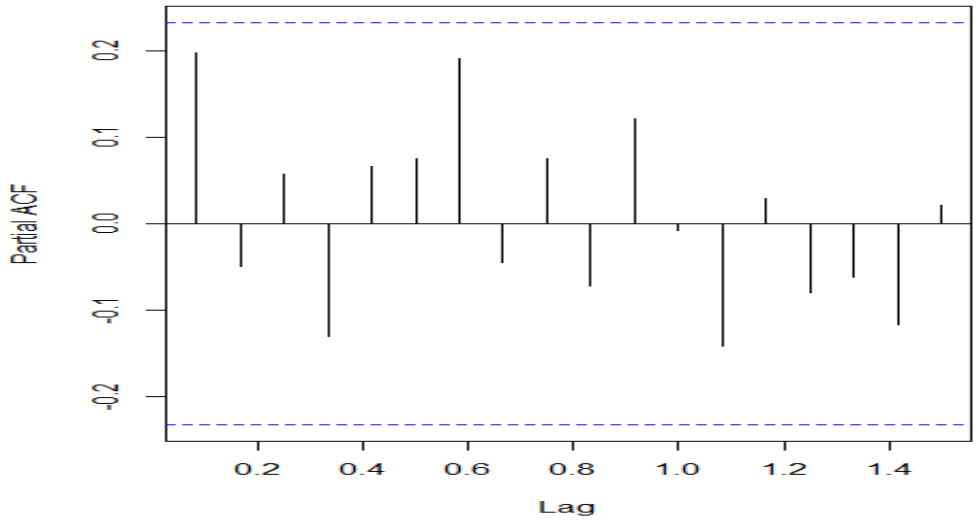


Figure 5: Partial Autocorrelation of Differenced log transformed prostate cancer incidences data

From the figure 5 a correlogram of the Partial Autocorrelation function of the Difference log transformed prostate cancer incidences data the PACF had exponential decay.

3.4 Parameter Estimation and Selection of the ARIMAX model for Prostate Cancer Incidence Data

This involves the estimation of parameters of different models using identification process and proceeds to the selection of the model using information criteria. The best ARIMAX model with the lowest Akaike Information Criteria (AIC) was selected.

Table 3: Comparison of ARIMAX models with corresponding AIC values

Model	AIC
ARIMAX(0,0,1)	169.8
ARIMAX(0,1,1)	201.35
ARIMAX(1,1,1)	194.03
ARIMAX(1,0,1)	175.21

From Table 3, ARIMAX (0,0,1) had an AIC value of 169.8, ARIMAX (0,1,1) had an AIC value of 201.35,ARIMAX (1,1,1) had an AIC value of 194.03 and ARIMAX (1,0,1) had an AIC value of 175.21. Thus ARIMAX (0,0,1) was chosen as the best . The summary results of the ARIMAX

Table 4: ARIMAX(0,0,1) Model Coefficients and Error Measures

Coefficients	**Estimate**	height**Standard Error (s.e.)**
ma1	-0.8975	0.0626
xreg	0.5871	1.2578
Model Statistics		height
σ^2	0.6123	
Log Likelihood	-81.96	
AIC	169.92	
AICc	170.28	
BIC	176.66	
Training Set Error Measures		height
ME	0.0467	
RMSE	0.7712	
MAE	0.6437	
MPE	NaN	
MAPE	Inf	
MASE	0.6646	
ACF1	0.1250	

The estimated coefficient was (ma1 = -0.8981) for the MA(1) term. A negative coefficient close to -1 suggested that the series has strong short-term negative autocorrelation. That is the last forecast error was positive, the model expects the next value to be lower, and vice versa. The Standard Error was (s.e.) = 0.0632 where this is precision of the estimated MA1 coefficient. A relatively small standard error indicated that the estimate was precise. This ARIMA(0,0,1) model with a negative MA(1) coefficient captures short-term fluctuations in the differenced log-transformed series of prostate cancer cases.

This model was an ARIMA(0,0,1) model with an additional regression term (indicated by xreg). It combined a linear regression component with an ARIMA model to capture the residual patterns in the data that the regression alone cannot account for. The coefficient for the external variable (age) suggests a positive relationship with the response variable. Specifically, for each unit increase in xreg, the log-transformed and differenced prostate cancer monthly cases are expected to increase by 0.5871 units, holding other factors constant.

3.5 Diagnostics and Evaluation of the ARIMAX model for Prostate Cancer Incidence Data

ARIMAX (0,0,1) was the best fit model, the model adequacy was further checked to draw empirical conclusion regarding the model as good fit hence thus used in estimation and forecasting. Ljung Box test coupled with the ACF, PACF, the normal Q-Q plot and histogram plots of the residuals were used in model diagnostics. The plots showed that the residuals from the model are similar to a white noise hence the model fits the data well. The p-value estimated by the Ljung Box test was greater than 0.05. The p-value was at 0.7096 which showed that the residuals were random, concluding that there is no significant autocorrelation, given the p-value of 0.7192, which exceeds the conventional significance level of 0.05.

Table 5: Ljung-Box test for Autocorrelation of Residuals in the ARIMAX(0,0,1) Model

X-squared	df	P-value
7.0664	10	0.7096

The p-value (0.7096), (greater than 0.05), we fail to reject the null hypothesis. Therefore there is no significant autocorrelation remaining in the residuals of the ARIMAX model. The residuals the ARIMAX model do not show significant autocorrelation, indicating that the model is a good fit in terms of capturing the temporal dependencies in the data. The model residuals are approximately white noise, which is a desirable outcome for time series models.

The normal Q-Q plot the residuals of the prostate cancer time series data indicates that residuals are located on the straight line except a few that are deviating from the normality. Hence the normality assumption is satisfied.

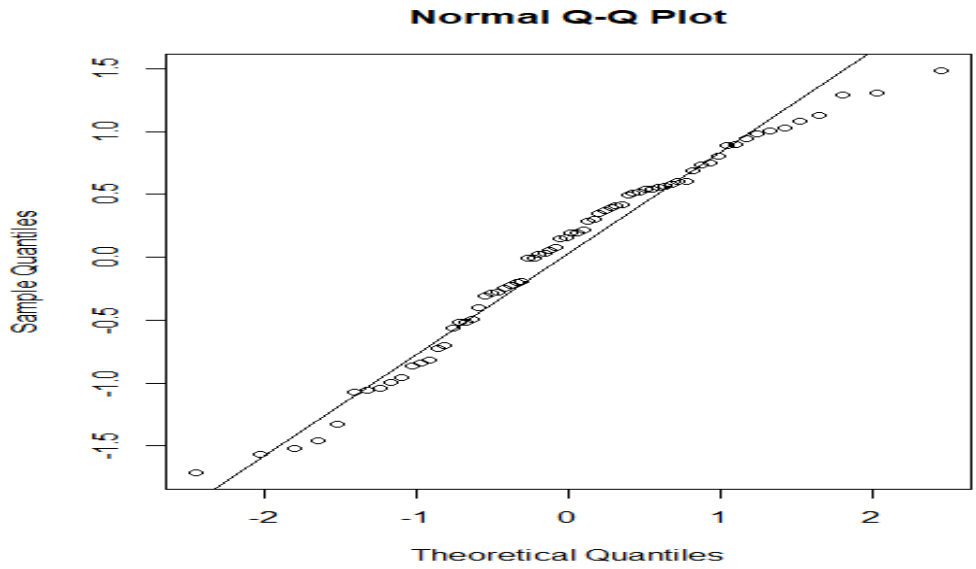


Figure 6: normal Q-Q plot of the residuals

The points on the Q-Q plot follow a straight line.

A bell-shaped histogram indicates that residuals are normally distributed, which supports the assumption of normality.

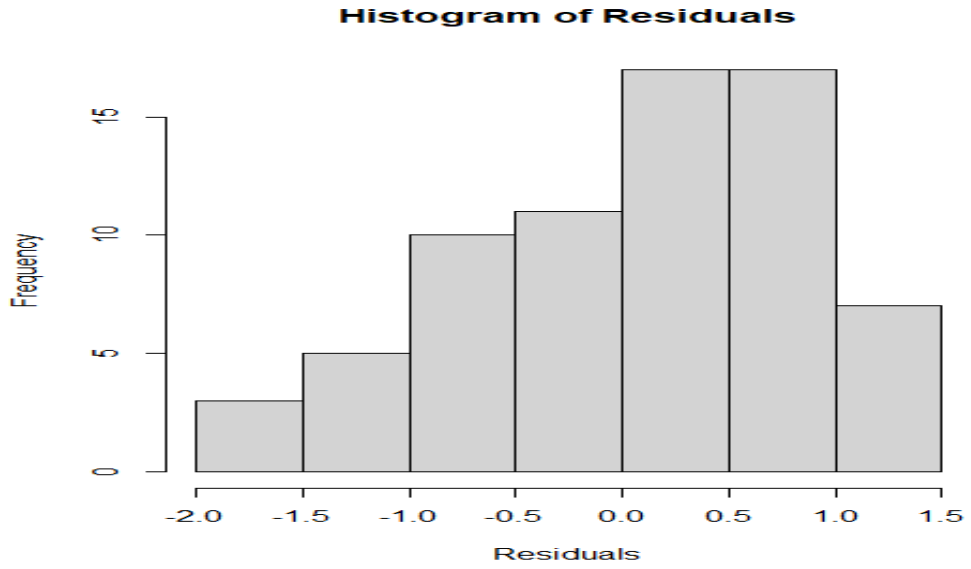


Figure 7: Histogram of Residuals

The residuals appear randomly scattered around zero with no clear patterns over time. Thus the model captured the data well and that there was homoscedasticity (constant variance of residuals over time).

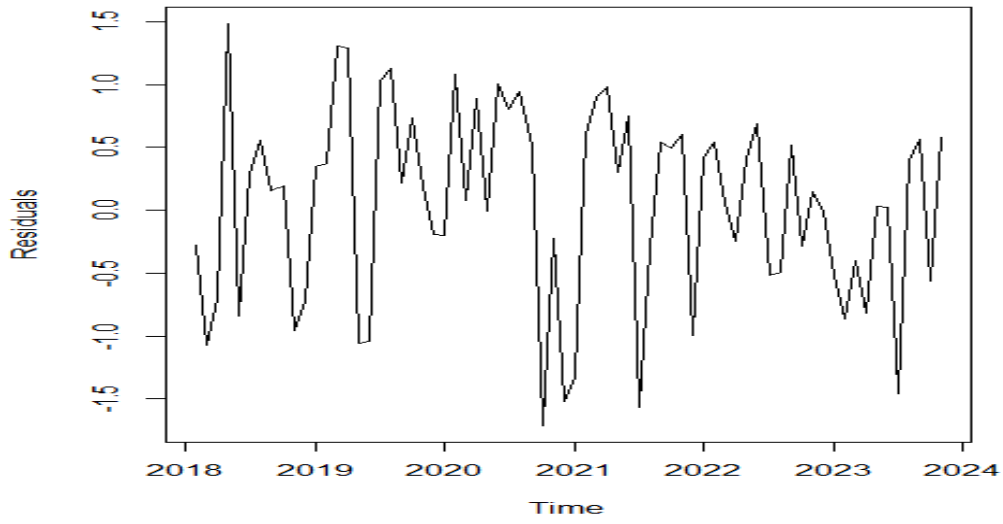


Figure 8: Residual Time Plot

3.6 Forecasts of the Prostate Cancer Incidence Data

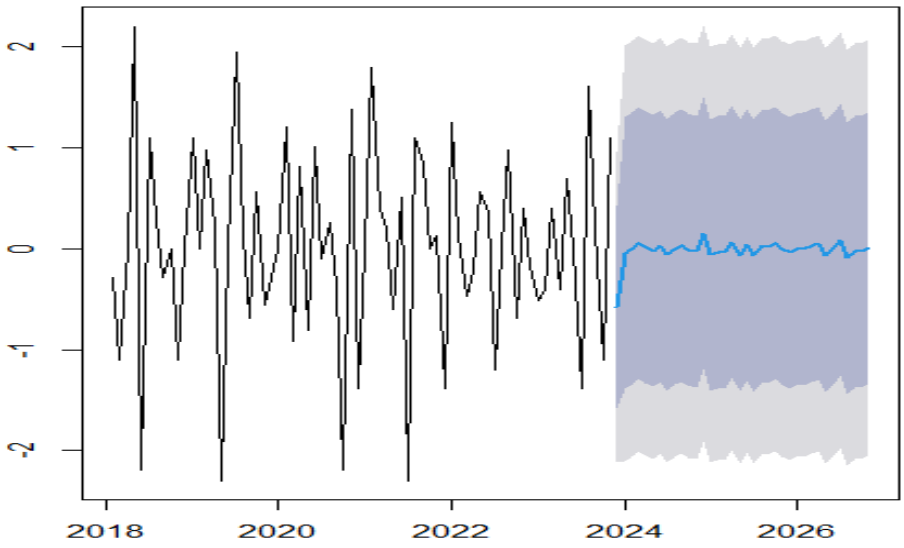
The forecast plot shows the predicted values continuing from the historical data, with reasonably narrow prediction intervals, suggesting confidence in the predictions.The model prediction effects was evaluated through MAE, RMSE, and MAPE.The Smaller values indicated better predictions.

Table 6: MAE,RMSE and MASE

MAE	RMSE	MASE
0.6437	0.7712	0.6646

MAE and RMSE values are relatively close, indicating that the model's prediction errors are fairly consistent in magnitude.

Figure 9: ARIMAX Forecasted values with confidence intervals



The predicted values represented the model's forecast, with 80% and 90% confidence intervals provide a range of uncertainty for these predictions. The actual values were within these intervals 80% and 90% confidence intervals.

4 Conclusion and Recommendations

This study modeled the trends of prostate cancer incidences in Meru County using ARIMAX models. The findings indicate a rising trend in incidences, with the ARIMAX model providing the most accurate forecasts by incorporating external variables such as age. These results underscore the importance of using advanced statistical models in epidemiological studies to inform public health policies.

Future research should explore the inclusion of additional exogenous variables in the ARIMAX model, such as lifestyle factors and genetic predispositions, to further enhance its predictive accuracy. Additionally, similar studies should be conducted in other counties to develop a comprehensive understanding of prostate cancer trends across Kenya.

References

[1] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[2] Machathoibi Takhellambam Chanu and Asem Surindro Singh. Cancer disease and its' understanding from the ancient knowledge to the modern concept. *World Journal of Advanced Research and Reviews*, 15(2):169–176, 2022.

[3] Avril Coghlan. A little book of r for time series. *Published under Creative Commons Attribution*, 3, 2015.

[4] Jonathan D Cryer and Natalie Kellet. *Time series analysis*. Springer, 1991.

[5] Don S Dizon and Arif H Kamal. Cancer statistics 2024: All hands on deck. *CA: a cancer journal for clinicians*, 74(1), 2024.

[6] Arul Earnest, Sue M Evans, Fanny Sampurno, and Jeremy Millar. Forecasting annual incidence and mortality rate for prostate cancer in australia until 2022 using autoregressive integrated moving average (arima) models. *BMJ open*, 9(8):e031331, 2019.

[7] Tartisio N Filder, Moses M Muraya, and Robert M Mutwiri. Application of seasonal autoregressive moving average models to analysis and forecasting of time series monthly rainfall patterns in embu county, kenya. 2019.

[8] Rob J Hyndman. A brief history of forecasting competitions. *International Journal of Forecasting*, 36(1):7–14, 2020.

[9] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[10] Amos Langat, George Orwa, and Joel Koima. Cancer cases in kenya; forecasting incidents using box & jenkins arima model. *Biomedical Statistics and Informatics*, 2(2):37–48, 2017.

[11] Norazliani Md Lazam, Syazreen Niza Shair, Nurin Haniah Asmuni, Aisyah Jamaludin, and Ardinei Ahmad Yusri. Forecasting the incidence rates of top three cancers in malaysia. In *AIP Conference Proceedings*, volume 2500. AIP Publishing, 2023.

[12] Jinhui Li, Nicholas B Chan, Jiashu Xue, and Kelvin KF Tsoi. Time series models show comparable projection performance with joinpoint regression: A comparison using historical cancer data from world health organization. *Frontiers in Public Health*, 10:1003162, 2022.

[13] Tingyan Luo, Jie Zhou, Jing Yang, Yulan Xie, Yiru Wei, Huanzhuo Mai, Dongjia Lu, Yuecong Yang, Ping Cui, Li Ye, et al. Early warning and prediction of scarlet fever in china using the baidu search index and autoregressive integrated moving average with explanatory variable (arimax) model: Time series analysis. *Journal of Medical Internet Research*, 25:e49400, 2023.

[14] Thanya Pathirana, Rehan Sequeira, Chris Del Mar, James A Dickinson, Bruce K Armstrong, Katy JL Bell, and Paul Glasziou. Trends in prostate specific antigen (psa) testing and prostate cancer incidence and mortality in australia: A critical analysis. *Cancer Epidemiology*, 77:102093, 2022.

[15] Robert H Shumway, David S Stoffer, Robert H Shumway, and David S Stoffer. Arima models. *Time series analysis and its applications: with R examples*, pages 75–163, 2017.

[16] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[17] Cristiana Tudor. A novel approach to modeling and forecasting cancer incidence and mortality rates through web queries and automated forecasting algorithms: Evidence from romania. *Biology*, 11(6):857, 2022.

[18] Andrew J Vickers, David Ulmert, Daniel D Sjoberg, Caroline J Bennette, Thomas Björk, Axel Gerdtsson, Jonas Manjer, Peter M Nilsson, Anders Dahlin, Anders Bjartell, et al. Strategy for detection of prostate cancer based on relation between prostate specific antigen at age 40-55 and long term risk of metastasis: case-control study. *Bmj*, 346, 2013.

[19] Kinley Wangdi, Pratap Singhasivanon, Tassanee Silawan, Saranath Lawpoolsri, Nicholas J White, and Jaranit Kaewkungwal. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and arimax analyses: a case study in endemic districts of bhutan. *Malaria Journal*, 9(1):1–9, 2010.