An Assessment Project on Time Series and Prediction
Department of Information Engineering and Computer Science and Mathematics

Time series Analysis and Forescasting of Cancer Mortality Rates Across the USA

Amos Tochukwu Ezeh
Matric No: 288543

1 Background of Study

The global mortality landscape is significantly impacted by cancer, claiming around 9.6 million lives and ranking as the second leading cause of death. Specifically, within the United States, cancer holds its position as the second leading cause of death. This study employs data from the research conducted by the Centers for Disease Control and Prevention [1], (`https://www.cdc.gov/pcd/issues/2020/19_0286.htm`) and data from `cancer.org`. The focus of this project is to conduct a time series analysis of the trends, and understanding the variations in yearly cancer mortality rates(Data is in multiples of hundred thousand i.e 100000) in the United States spanning from 1968 to 2023 (T=56). The ultimate goal is to employ this historical data to forecast potential mortality rates that could unfold in the future, providing valuable insights into the trajectory of cancer-related fatalities. It is worth mentioning that, the entire analysis will be done using GRETL

2 Preliminary Analysis

In the preliminary analysis phase, we closely examine the primary dataset sourced from [1] with the objective of comprehending the data's patterns, identifying trends, and assessing for any signs of seasonality or cycles. The aim is to discern whether the data represents a realization of a stationary process.

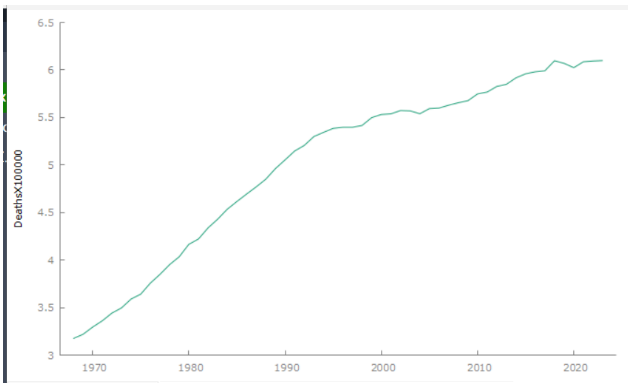


Figure 1: Primary dataset plot

Remark
The graphical examination of the preliminary plot, reveals a noticeable upward trend in the time

series data. The presence of trend in this time series strongly suggests that the series lacks stationarity.

2.1 Stationarity of The Time series

To be able to have a good forecast of the future, then the time series data needs to be a realisation of stationary process. "If the future is too different from the past the ARMA model will produce biased forecasts."

(1) First Difference
To achieve stationarity on this dataset; Firstly, i will adopt the first difference of the time series.

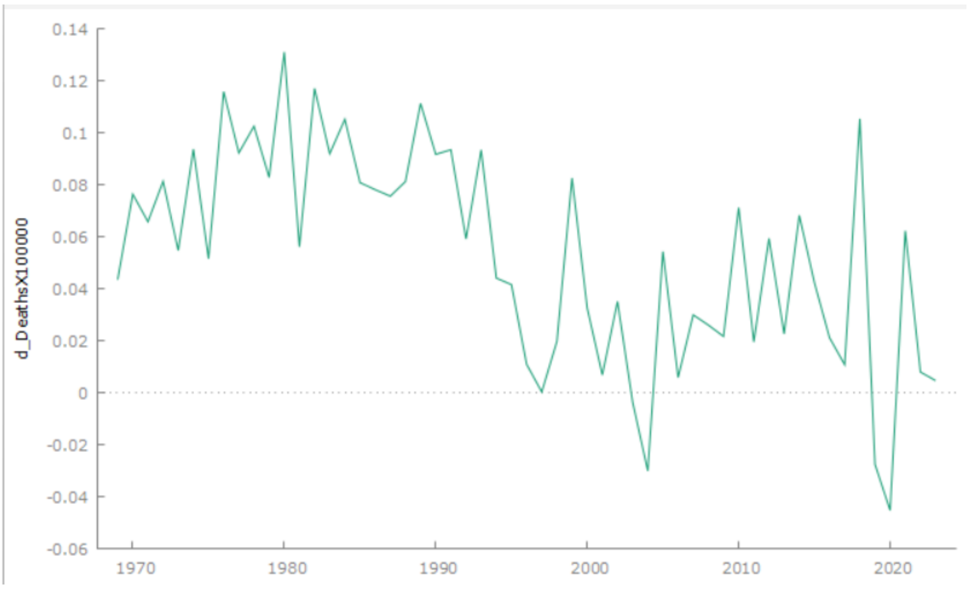
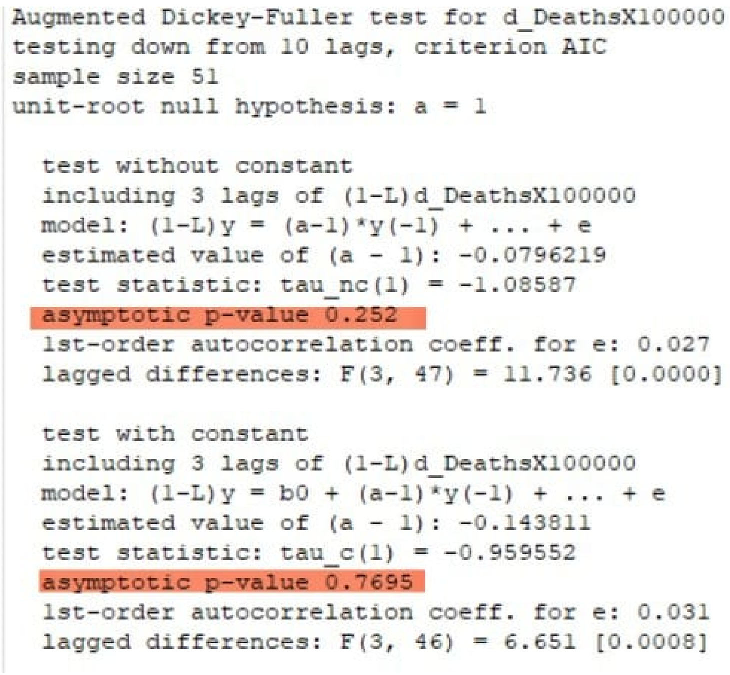


Figure 2: First difference plot

Remark
From the First difference plot, it could be observed that although the trend seen in the preliminary plot have almost disappeared. However, the first difference is not stationary. This is because the mean is not stationary as seen in the plot.

Moreover, the non-stationarity of the first difference is evident from the results of the Augmented Dickey-Fuller test (ADF Test) report provided below. The high p-value obtained from the ADF test strongly supports the null hypothesis of non-stationarity. Specifically, the p-value surpasses the 0.05 significance level, leading to the acceptance of the null hypothesis indicating non-stationarity.



```
Augmented Dickey-Fuller test for d_DeathsX100000
testing down from 10 lags, criterion AIC
sample size 51
unit-root null hypothesis: a = 1

  test without constant
  including 3 lags of (1-L)d_DeathsX100000
  model: (1-L)y = (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -0.0796219
  test statistic: tau_nc(1) = -1.08587
  asymptotic p-value 0.252
  1st-order autocorrelation coeff. for e: 0.027
  lagged differences: F(3, 47) = 11.736 [0.0000]

  test with constant
  including 3 lags of (1-L)d_DeathsX100000
  model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -0.143811
  test statistic: tau_c(1) = -0.959552
  asymptotic p-value 0.7695
  1st-order autocorrelation coeff. for e: 0.031
  lagged differences: F(3, 46) = 6.651 [0.0008]
```

Figure 3: Second difference plot

(2) Second Difference

Next, I consider the second difference of the time series dataset.

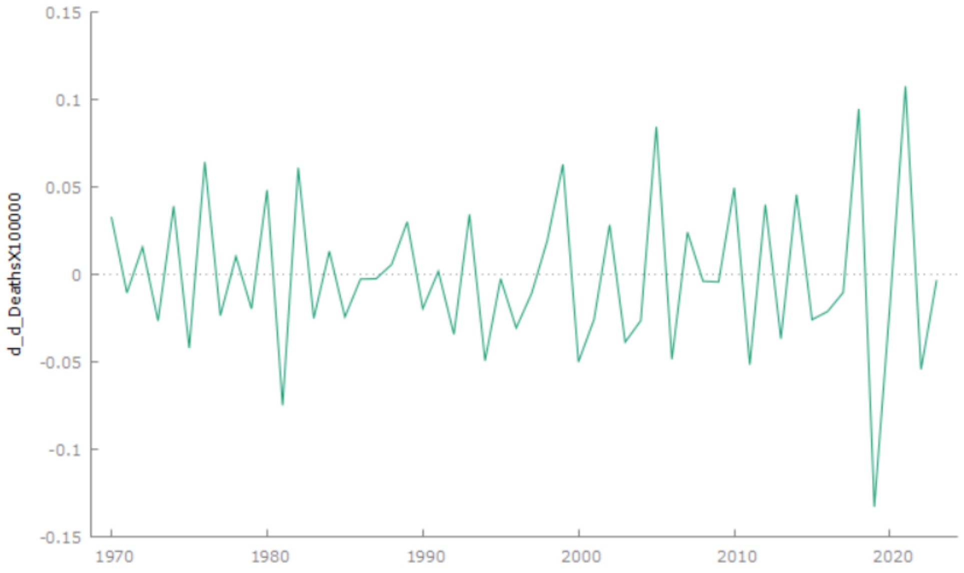
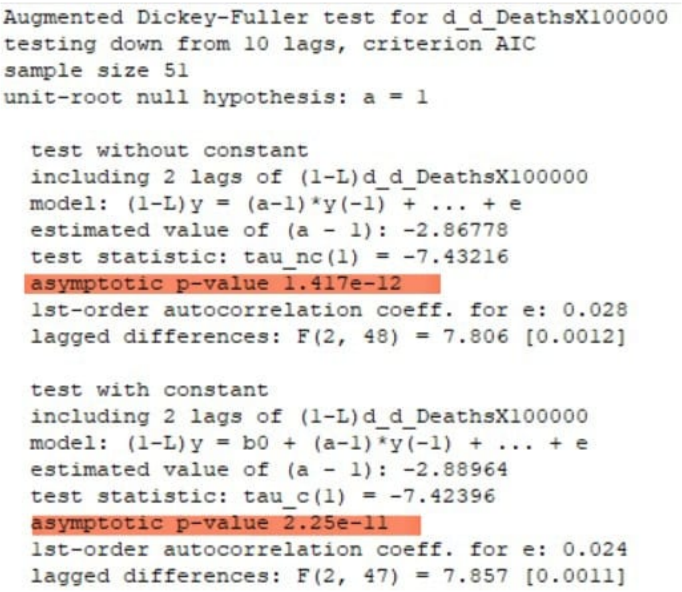


Figure 4: Second difference plot

Remark

The plot of the second difference of the time series suggest that the series is a realisation of a stationary process.

Furthermore, The stationarity of the second difference can also be seen from the result of the Augmented Dickey-Fuller test (ADF Test) of the second difference given below. The ADF test provides compelling evidence for stationarity as the p-value associated with the null hypothesis of non-stationarity is notably low, specifically below the 0.05 significance level. Consequently, we reject the null hypothesis of non-stationarity.



```
Augmented Dickey-Fuller test for d_d_DeathsX100000
testing down from 10 lags, criterion AIC
sample size 51
unit-root null hypothesis: a = 1

  test without constant
  including 2 lags of (1-L)d_d_DeathsX100000
  model: (1-L)y = (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -2.86778
  test statistic: tau_nc(1) = -7.43216
  asymptotic p-value 1.417e-12
  1st-order autocorrelation coeff. for e: 0.028
  lagged differences: F(2, 48) = 7.806 [0.0012]

  test with constant
  including 2 lags of (1-L)d_d_DeathsX100000
  model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -2.88964
  test statistic: tau_c(1) = -7.42396
  asymptotic p-value 2.25e-11
  1st-order autocorrelation coeff. for e: 0.024
  lagged differences: F(2, 47) = 7.857 [0.0011]
```

Figure 5: Second difference plot

3 Model Identification

To effectively model the ARMA(p,q) process, the crucial step is to ascertain the order of the process by determining appropriate values for p and q in the time series. This involves plotting the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) and leveraging Automatic criteria to establish the optimal order for the ARMA process.

3.1 ACF and PACF

To ascertain the order of the process, I plot the correlogram and investigate the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

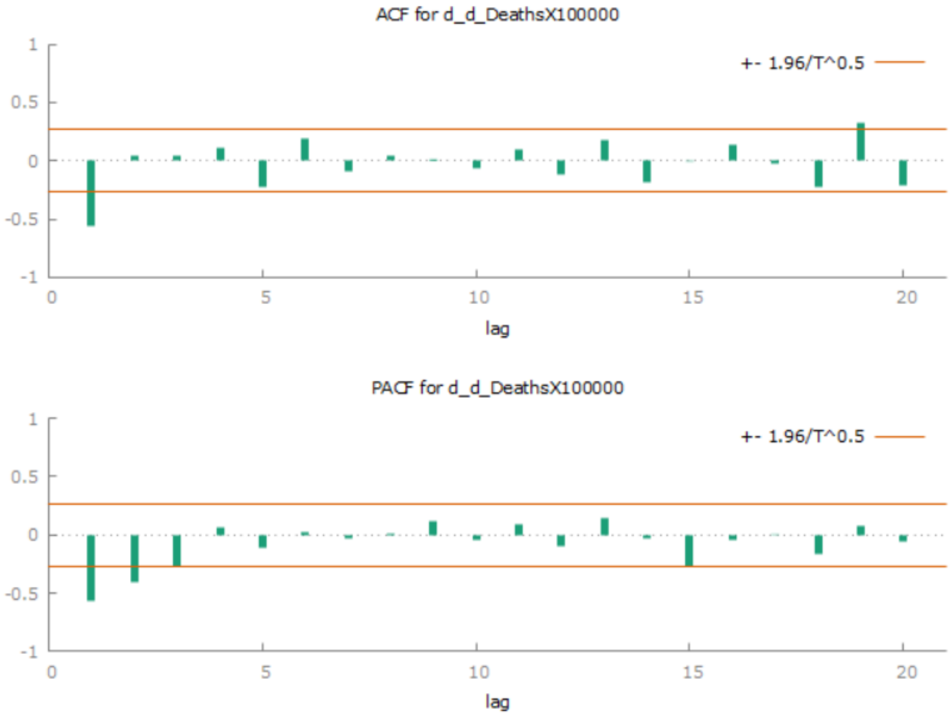
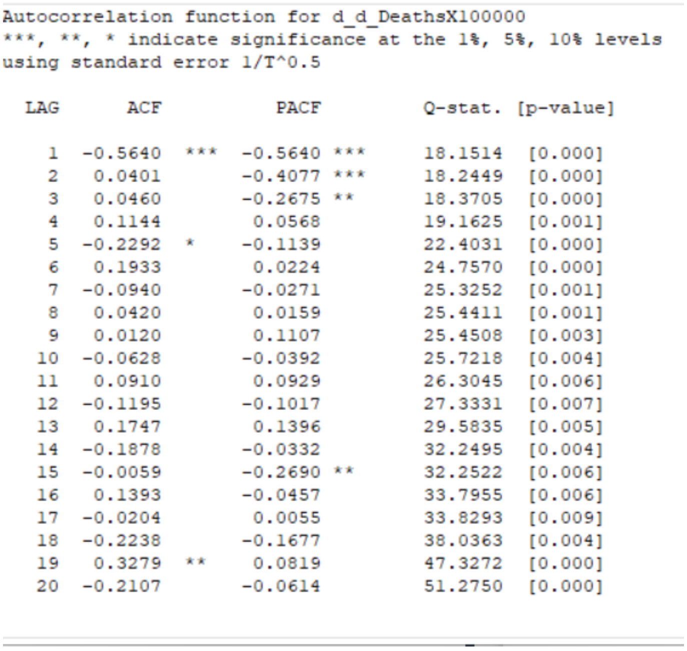


Figure 6: ACF and PACF plots



```
Autocorrelation function for d_d_DeathsX100000
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0.5

  LAG      ACF          PACF         Q-stat. [p-value]

    1  -0.5640  ***  -0.5640 ***     18.1514  [0.000]
    2   0.0401       -0.4077 ***     18.2449  [0.000]
    3   0.0460       -0.2675 **      18.3705  [0.000]
    4   0.1144        0.0568         19.1625  [0.001]
    5  -0.2292  *    -0.1139         22.4031  [0.000]
    6   0.1933        0.0224         24.7570  [0.000]
    7  -0.0940       -0.0271         25.3252  [0.001]
    8   0.0420        0.0159         25.4411  [0.001]
    9   0.0120        0.1107         25.4508  [0.003]
   10  -0.0628       -0.0392         25.7218  [0.004]
   11   0.0910        0.0929         26.3045  [0.006]
   12  -0.1195       -0.1017         27.3331  [0.007]
   13   0.1747        0.1396         29.5835  [0.005]
   14  -0.1878       -0.0332         32.2495  [0.004]
   15  -0.0059       -0.2690 **      32.2522  [0.006]
   16   0.1393       -0.0457         33.7955  [0.006]
   17  -0.0204        0.0055         33.8293  [0.009]
   18  -0.2238       -0.1677         38.0363  [0.004]
   19   0.3279  **    0.0819         47.3272  [0.000]
   20  -0.2107       -0.0614         51.2750  [0.000]
```

Figure 7: Table of Significance of ACF and PACF

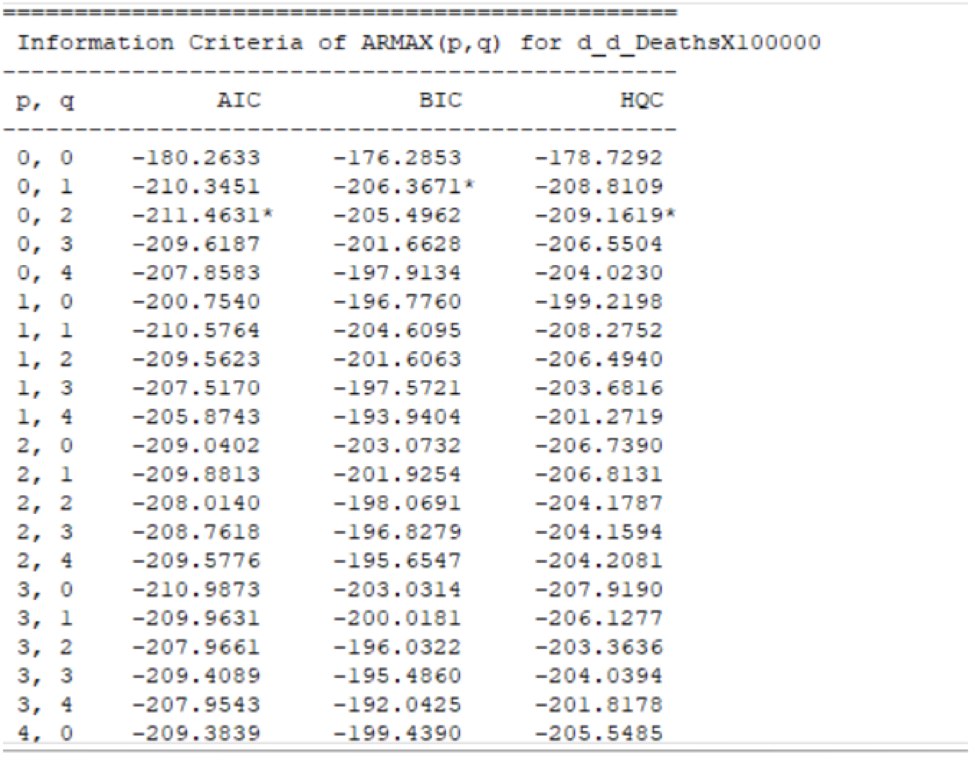
Remark
From the correlogram plot and the table of significance, the following was observed:

(1) **ACF:** Upon examining the Autocorrelation Function (ACF) plot for the second difference, a notable observation is that the plot is siginificantly different from zero only at lag one (1), 5 and 19. Since the first lag is always considered crucial in our decisions, This suggest that a pottential model is an **MA(1).**

(2) **PACF:** Similarly, Upon examining the Partial Autocorrelation Function (PACF) plot for the second difference, It is observed that the plot is significantly different from zero only on the first three lags. This suggest an **AR(3)** process.

3.2 Automatic Criteria

Here, Automatic criteria of AIC, BIC, HQC are investigated and their suggested optimal process is identified. Maximum AR lag order considered is $P = 4$ and maximum MA lag ordered is $Q = 4$. The information of the criteria is as follows:



```
================================================
 Information Criteria of ARMAX(p,q) for d_d_DeathsX100000
------------------------------------------------
 p, q          AIC            BIC            HQC
------------------------------------------------
 0, 0     -180.2633      -176.2853      -178.7292
 0, 1     -210.3451      -206.3671*     -208.8109
 0, 2     -211.4631*     -205.4962      -209.1619*
 0, 3     -209.6187      -201.6628      -206.5504
 0, 4     -207.8583      -197.9134      -204.0230
 1, 0     -200.7540      -196.7760      -199.2198
 1, 1     -210.5764      -204.6095      -208.2752
 1, 2     -209.5623      -201.6063      -206.4940
 1, 3     -207.5170      -197.5721      -203.6816
 1, 4     -205.8743      -193.9404      -201.2719
 2, 0     -209.0402      -203.0732      -206.7390
 2, 1     -209.8813      -201.9254      -206.8131
 2, 2     -208.0140      -198.0691      -204.1787
 2, 3     -208.7618      -196.8279      -204.1594
 2, 4     -209.5776      -195.6547      -204.2081
 3, 0     -210.9873      -203.0314      -207.9190
 3, 1     -209.9631      -200.0181      -206.1277
 3, 2     -207.9661      -196.0322      -203.3636
 3, 3     -209.4089      -195.4860      -204.0394
 3, 4     -207.9543      -192.0425      -201.8178
 4, 0     -209.3839      -199.4390      -205.5485
```

Figure 8: AIC, BIC and HQC information criteria

Remark
The Automatic criteria suggests the following:

(I) **AIC & HQC:** The AIC and HQC suggests MA(2)
(II) **BIC:** The BIC suggest an MA(1)

3.3 Potential Models

Kindly note that the models identified were based on the series of the second difference of the dataset(i.e $\Delta^2 = (1-L)^2$). Hence, to apply this models to the actual time series, the Autoregressive Integrated Moving Average (ARIMA) model with difference $d = 2$ will be used for the estimation of the model parameters. After the model identification process, the following are the potential models of the time series:

(1) ACF and BIC suggests:

$$\Delta^2 X_t \sim MA(1) \implies X_t \sim ARIMA(0,2,1)$$

(2) PACF suggests:

$$\Delta^2 X_t \sim AR(3) \implies X_t \sim ARIMA(3,2,0)$$

(3) AIC and HQC Suggest:

$$\Delta^2 X_t \sim MA(2) \implies X_t \sim ARIMA(0,2,2)$$

4 Model estimation and Checking

In the model estimation and checking stage, each potential model suggested above is used to estimate the values of the unknown model parameters. Subsequently, a critical test is conducted to verify the model's suitability for forecasting. This involves examining whether the obtained residuals represent a realization of a Gaussian white noise process.

To proceed, firstly, the dataset of the series will be partitioned as follows:

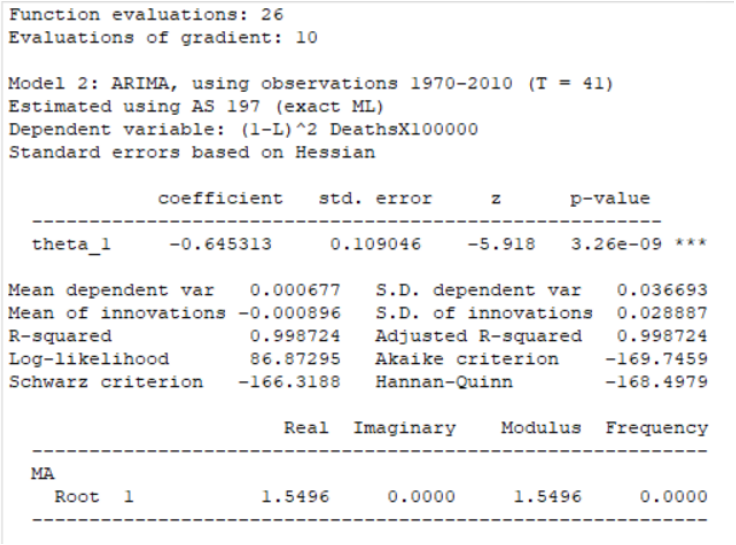
(A) **Training Data (1968-2010):** The Training dataset comprises of about 80% of the total dataset of the timeseries. From this range, the model parameter will be estimated.

(B) **Testing Data (2011-2023):** The Testing dataset comprises of about 20% of the total dataset of the timeseries. From this dataset, the performance and validation of the model will be conducted, by comparing the prediction of the model by the actual value. The Mean Squared Error(MSE), Root Mean squared Error(RMSE) and other will derived.

Lastly, I compared The Root Mean Squared Error (RMSE) of each suggested model in order to choose the model that performed best. The Optimal model is the model with the least Mean squared error.

4.1 ARIMA(0,2,1)

Considering the ARIMA(0,2,1), the estimates of the model parameters is obtained as follows:



```
Function evaluations: 26
Evaluations of gradient: 10

Model 2: ARIMA, using observations 1970-2010 (T = 41)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L)^2 DeathsX100000
Standard errors based on Hessian

             coefficient   std. error     z      p-value
  --------------------------------------------------------
  theta_1     -0.645313     0.109046    -5.918   3.26e-09 ***

Mean dependent var    0.000677   S.D. dependent var   0.036693
Mean of innovations -0.000896   S.D. of innovations  0.028887
R-squared             0.998724   Adjusted R-squared   0.998724
Log-likelihood        86.87295   Akaike criterion    -169.7459
Schwarz criterion   -166.3188   Hannan-Quinn        -168.4979

                        Real  Imaginary    Modulus  Frequency
  -----------------------------------------------------------
  MA
    Root  1           1.5496     0.0000     1.5496     0.0000
  -----------------------------------------------------------
```

Figure 9: Parameter Estimation of ARIMA(0,2,1)

In this case, with a highly acceptable P-value, I obtained the value of the model parameter as $\theta = -0.645313$.

Following the model estimation, the subsequent verification involves ensuring that the obtained model estimate is well-suited for forecasting future values. This entails a meticulous analysis of the residual error's correlogram to confirm that the residual error aligns with the characteristics of a white noise process.

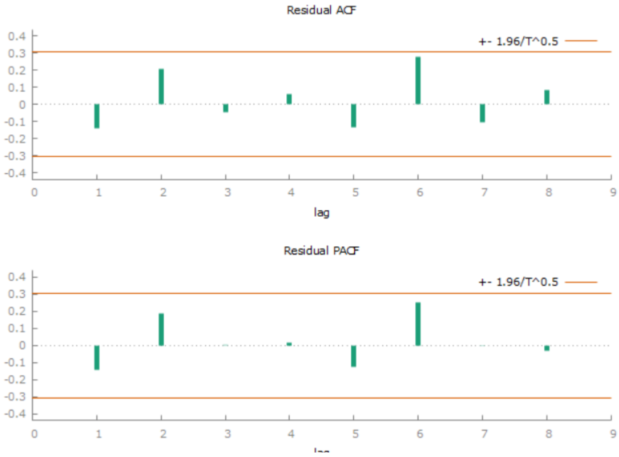


Figure 10: Residual Correlogram of ARIMA(0,2,1)

From this plot, it can be seen that all lags are not significantly different from zero. Thus, the residual is a white noise. Next, I check if the residual can be considered as a realisation of a Gaussian white noise. For this, I analyse the q-statistics (Normality test) of the residual and the qq-plot of the residual.

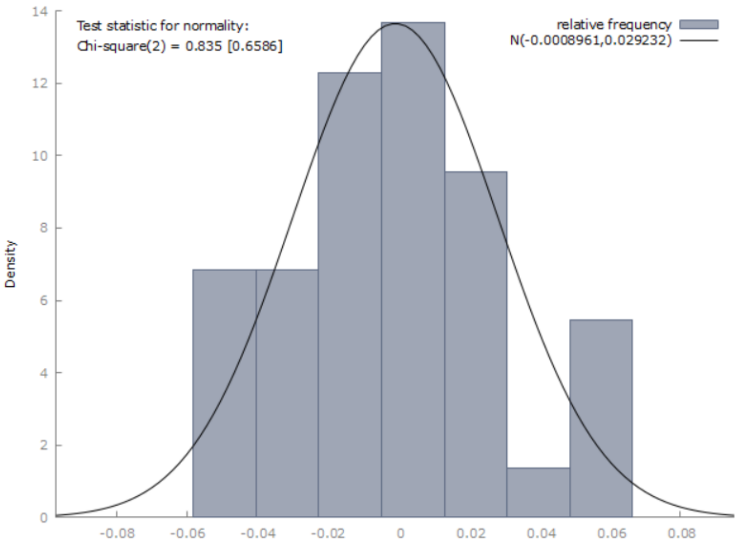


Figure 11: Residual Normality of ARIMA(0,2,1)

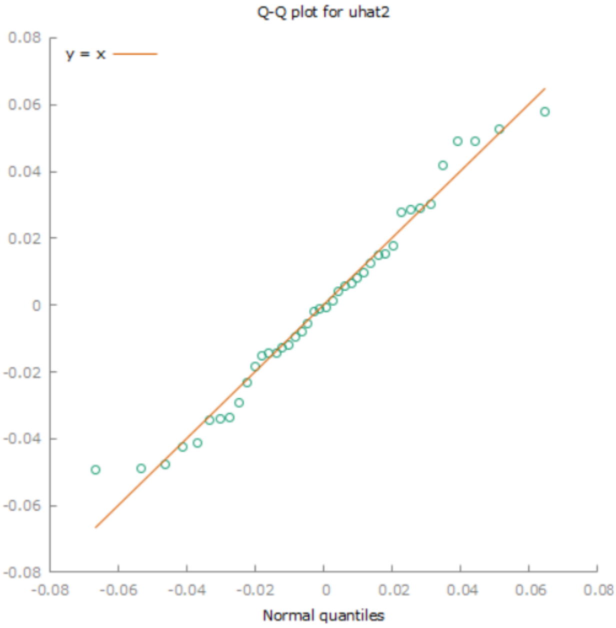


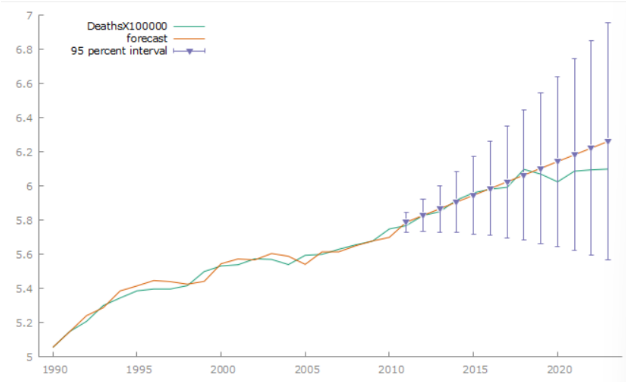
Figure 12: Residual QQ-Plot of ARIMA(0,2,1)

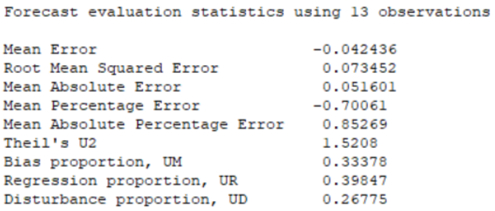
Remark
From the Normality plot, it is observed that the p-value is about **0.65** (which is highly greater than 0.05). Also from the qq-plot, it can be seen that the data sets are not so different from the red plot.

Therefore, I conclude that the residual is a realisation of a Gaussian white noise. Hence the model is appropriate for forecasting.

4.1.1 Forecasting of ARIMA(0,2,1)

Using the testing dataset, the following is obtained.





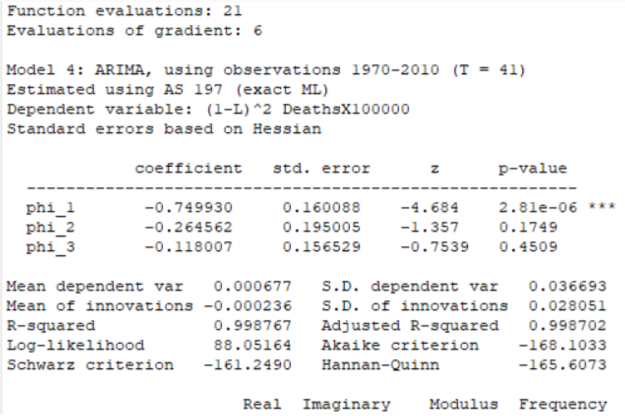
```
Forecast evaluation statistics using 13 observations

Mean Error                         -0.042436
Root Mean Squared Error             0.073452
Mean Absolute Error                 0.051601
Mean Percentage Error              -0.70061
Mean Absolute Percentage Error      0.85269
Theil's U2                          1.5208
Bias proportion, UM                 0.33378
Regression proportion, UR           0.39847
Disturbance proportion, UD          0.26775
```

Remark
In conclussion, the obtained Root Mean Squared error of the ARIMA (0,2,1) model is :

RMSE = 0.073452

4.2 ARIMA(3,2,0)

In this case, the model estimation is as given below. However, the ϕ_2 and ϕ_3 values can be consider an oversighted value because the p-value is higher than the acceptable threshold of 0.05.



```
Function evaluations: 21
Evaluations of gradient: 6

Model 4: ARIMA, using observations 1970-2010 (T = 41)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L)^2 DeathsX100000
Standard errors based on Hessian

             coefficient   std. error      z      p-value
  --------------------------------------------------------
  phi_1       -0.749930     0.160088    -4.684    2.81e-06 ***
  phi_2       -0.264562     0.195005    -1.357    0.1749
  phi_3       -0.118007     0.156529    -0.7539   0.4509

Mean dependent var   0.000677   S.D. dependent var   0.036693
Mean of innovations -0.000236   S.D. of innovations  0.028051
R-squared            0.998767   Adjusted R-squared   0.998702
Log-likelihood       88.05164   Akaike criterion    -168.1033
Schwarz criterion   -161.2490   Hannan-Quinn        -165.6073

                        Real   Imaginary    Modulus  Frequency
```

Figure 13: Parameter Estimation of ARIMA(3,2,0)

Furthermore, the residual correlogram below also suggest that the residual is a realisation of white noise process.

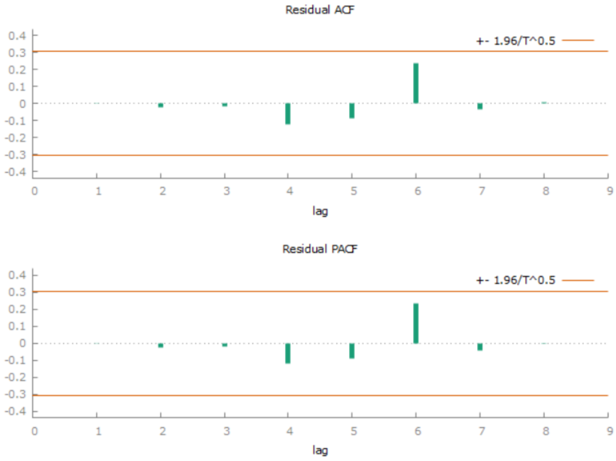


Figure 14: Residual Correlogram of ARIMA(3,2,0)

Lastly, from the Normality plot of the residual (with a p-value = 0.67) and the qq-plot, we can conclude that the residual is a realisation of a Gaussian White Noise process.

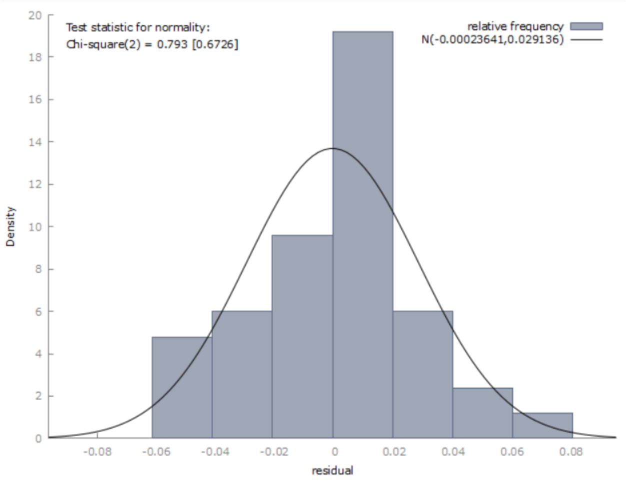


Figure 15: Residual Normality of ARIMA(3,2,0)

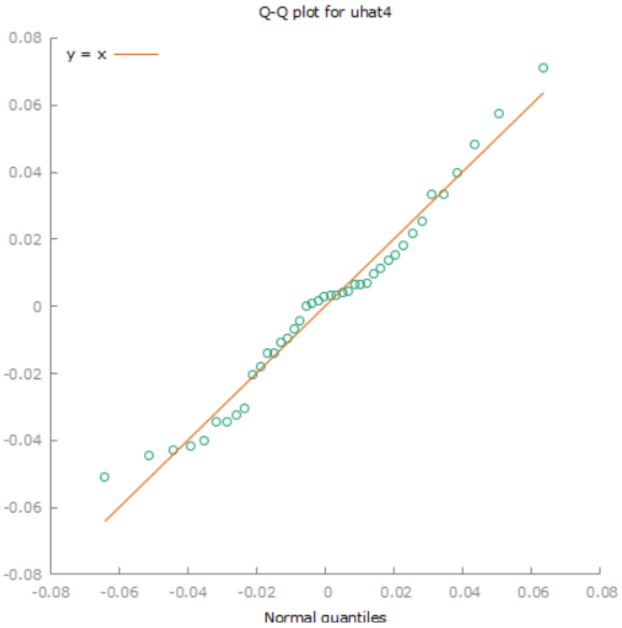
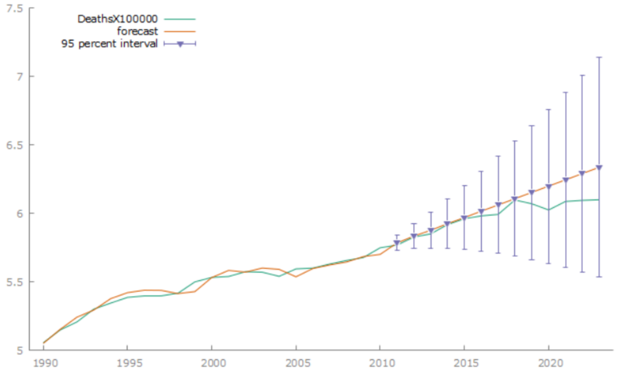


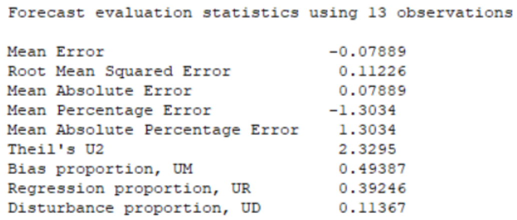
Figure 16: Residual QQ-Plot of ARIMA(3,2,0)

Thus the model is an appropriate model for forecasting the future value.

4.2.1 Forecasting of ARIMA(3,2,0)

The forecast of the testing dataset obtained by the ARIMA(3,2,0) is give below.





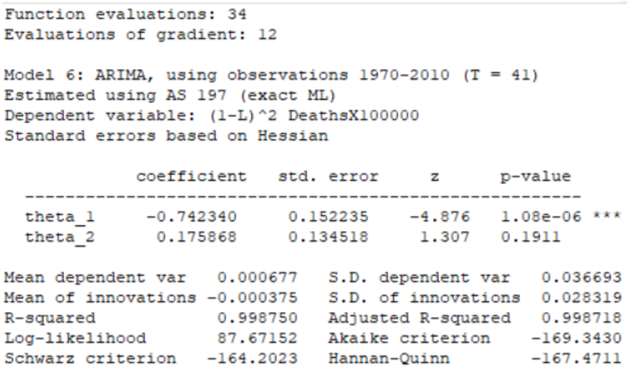
```
Forecast evaluation statistics using 13 observations

Mean Error                        -0.07889
Root Mean Squared Error            0.11226
Mean Absolute Error                0.07889
Mean Percentage Error             -1.3034
Mean Absolute Percentage Error     1.3034
Theil's U2                         2.3295
Bias proportion, UM                0.49387
Regression proportion, UR          0.39246
Disturbance proportion, UD         0.11367
```

Remark
In conclussion, the obtained Root Mean Squared error of the ARIMA (3,2,0) model is :

RMSE = 0.11226

4.3 ARIMA(0,2,2)

The model estimate for this suggested model is given below.Also, the θ_2 values can be consider an oversighted value because the p-value is higher than the acceptable threshold of 0.05.



```
Function evaluations: 34
Evaluations of gradient: 12

Model 6: ARIMA, using observations 1970-2010 (T = 41)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L)^2 DeathsX100000
Standard errors based on Hessian

             coefficient   std. error     z      p-value
  --------------------------------------------------------
  theta_1     -0.742340     0.152235    -4.876   1.08e-06 ***
  theta_2      0.175868     0.134518     1.307   0.1911

Mean dependent var    0.000677   S.D. dependent var    0.036693
Mean of innovations  -0.000375   S.D. of innovations   0.028319
R-squared             0.998750   Adjusted R-squared    0.998718
Log-likelihood        87.67152   Akaike criterion     -169.3430
Schwarz criterion    -164.2023   Hannan-Quinn         -167.4711
```

Figure 17: Parameter Estimation of ARIMA(0,2,2)

Similarly, the residual correlogram of the model also suggest that the model is a realisation of a White noise process since all lags of the ACF and PACF of the residual is not significantly different from zero as seen in the figure below:

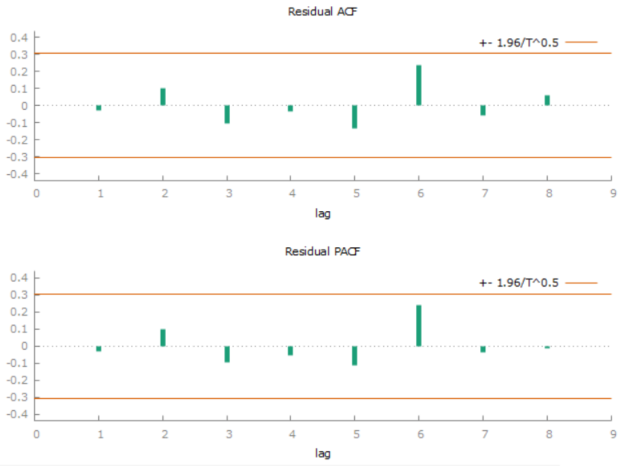


Figure 18: Residual Correlogram of ARIMA(0,2,2)

Also, the Normality test shows $p - values = 0.8168$ which is highly greater than the 0.05 threshold. Similarly, the qq-plot also shows a great correlation between the quantities. This implies that the residual of the process can be considered a realisation of a Gaussian white noise.

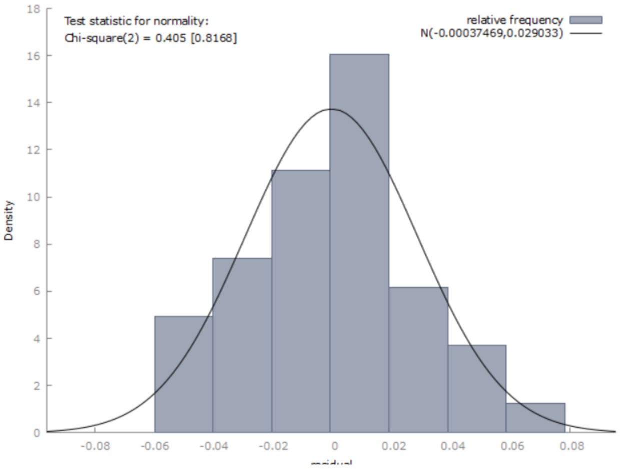


Figure 19: Residual Normality plot of ARIMA(0,2,2)

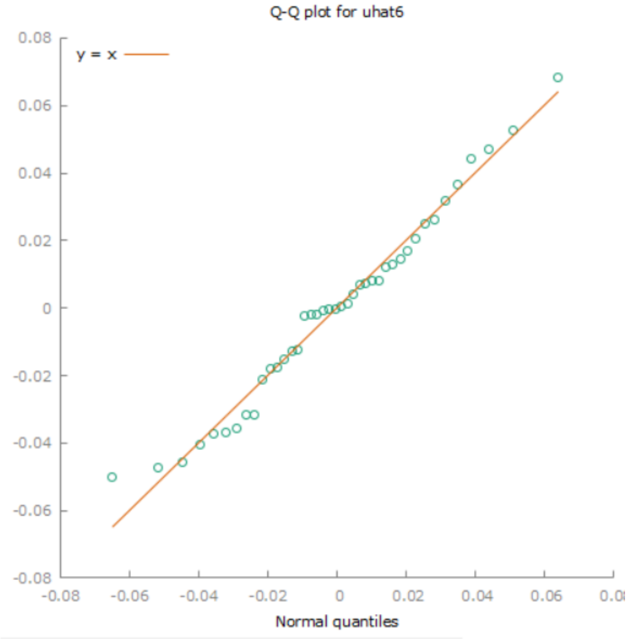
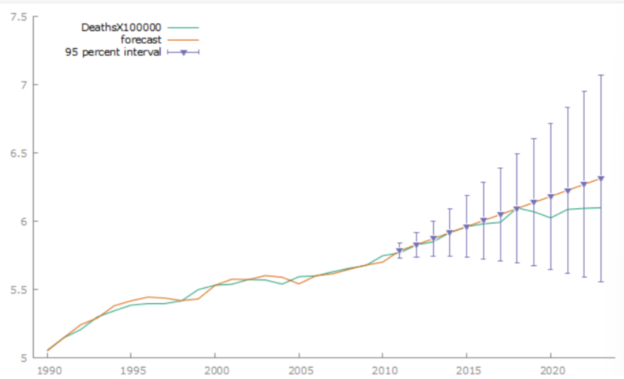


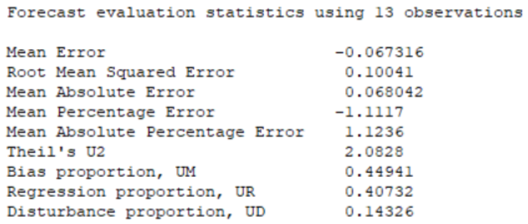
Figure 20: Residual QQ-plot of ARIMA(0,2,2)

Given that the residual of the model exhibits the characteristics of a Gaussian white noise process, it can be reasonably concluded that the model is indeed appropriate for forecasting future values.

4.3.1 Forecasting of ARIMA(0,2,2)

The ARIMA(0,2,2) model was applied to forecast the testing dataset, and the obtained forecast is provided below.





```
Forecast evaluation statistics using 13 observations

Mean Error                          -0.067316
Root Mean Squared Error              0.10041
Mean Absolute Error                  0.068042
Mean Percentage Error               -1.1117
Mean Absolute Percentage Error       1.1236
Theil's U2                           2.0828
Bias proportion, UM                  0.44941
Regression proportion, UR            0.40732
Disturbance proportion, UD           0.14326
```

Remark

In conclusion, the Root Mean Squared Error (RMSE) of the ARIMA(0,2,2) model is :
RMSE = 0.10041

General Remark

Observing the parameter estimates for all three models, it's notable that constants were excluded. This decision stems from the fact that when constants were added, the associated p-values exceeded the acceptable threshold. This suggests an oversight regarding the inclusion of constants in the model parameters.

Summary

in the model estimation and checking, the root mean squared error of each proposed model is given below.

Model	RSME
ARIMA(0,2,1)	0.073452
ARIMA(3,2,0)	0.11226
ARIMA(0,2,2)	0.10041

Optimal Model

The optimal model for the time series is the model with the least (smallest) root mean squared error (RSME). From the table above, we can see that the optimal model is the **ARIMA(0,2,1)**.

5 Future Forecast of the Time series

With the ARIMA(0,2,1) model identified as the optimal choice for the time series, the process for forecasting future values involved the following steps:

- Restoration of the complete dataset spanning from 1968 to 2023.
- Estimation of the unknown parameters of the ARIMA(0,2,1) model utilizing the entire data range.
- Subsequently, the forecast for future values was generated as shown below

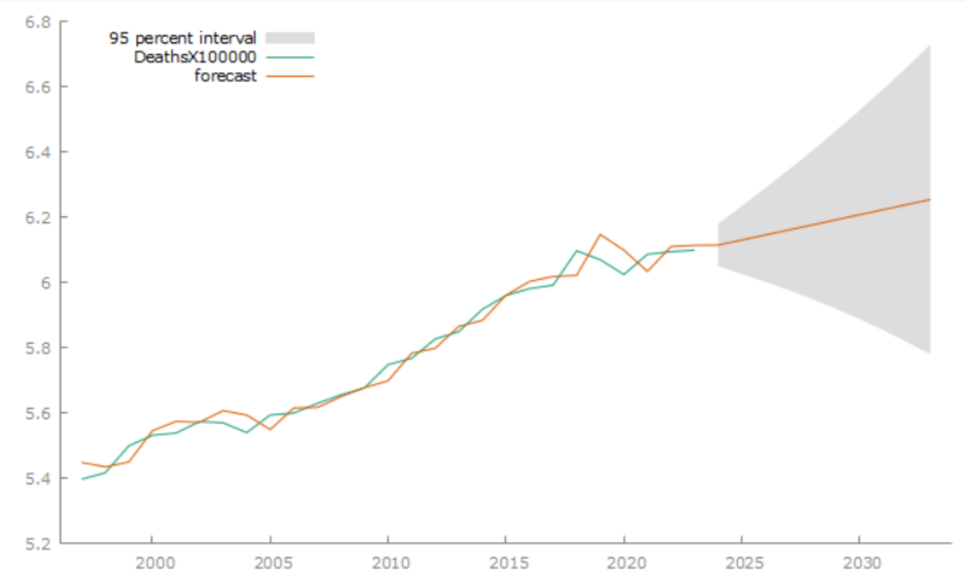


Figure 21: Forecast of 2024-2033

6 Conclusion

In conclusion, the findings of this study reveal a projected increase in cancer-related mortality in the USA over the next decade. The table below provides an overview of the predicted number of deaths by cancer in the country.

Year	Death by Cancer (X100000)	standard error	95% interval
2024	6.11373	0.03301	6.04902 - 6.17843
2025	6.12925	0.05322	6.02495 - 6.23356
2026	6.14478	0.07334	6.00103 - 6.28853
2027	6.16031	0.09425	5.97557 - 6.34504
2028	6.17583	0.11618	5.94813 - 6.40354
2029	6.19136	0.13919	5.91856 - 6.46417
2030	6.20689	0.16329	5.88685 - 6.52693
2031	6.22242	0.18847	5.85303 - 6.59180
2032	6.23794	0.21470	5.81714 - 6.65874
2033	6.25347	0.24195	5.77925 - 6.72769

Table 1: Forecasted Death by Cancer in the USA between 2024-2033.

Reference

[1] Michaels IH, Pirani SJ, Carrascal A. Visualizing 50 Years of Cancer Mortality Rates Across the US at Multiple Geographic Levels Using a Synchronized Map and Graph Animation. Prev Chronic Dis 2020;17:190286. DOI: https://doi.org/10.5888/pcd17.190286.