Time Series Analysis of Deaths In The United Kingdom

Table of Contents

1.1 Introduction

In this century, there have been two major turning points in life expectancy, the first in 2011 and the other in 2020 due to the COVID-19 pandemic, resulting in a high rise in mortality rivalling World War II (The King's Fund, 2024). Capturing the evolution of mortality rates over time can provide insights into the trends (Shen et al., 2024). This report aims to analyze and forecast the number of deaths in the United Kingdom annually using time series modelling.

Accurate time series models allow easy identification of trends, seasonality, and irregular components in historical data. For this task, they will provide actionable insights for predicting the number of deaths in the United Kingdom, to help guide public policies and allocate resources effectively.

This report will include three main parts:

- Exploratory Data Analysis (EDA).
- Time Series Modelling.
- Critical Evaluation and Recommendation.

1.2 Data Exploration

The Dataset used in this analysis(Vital Statistics in the UK), was compiled by the Office of National Statistics. It contained the annual number of deaths in the United Kingdom from 1887 to 2021.

1.2.1 Data Preprocessing

Before any analysis can be done, the data had to undergo some cleaning and preparation, as shown below:

Selecting relevant columns: Year and "Number of Deaths: United Kingdom" were selected.

```
# Selecting columns needed (Year,Number of deaths: United Kingdom)
death_uk <- uk_death %>%
  select(Year,`Number of deaths: United Kingdom`)
```

*Figure **Error! No text of specified style in document.**.1 Code selecting relevant columns in the death_uk data frame*

Renaming columns: The "Number of Deaths: United Kingdom" column was renamed to no_of_deaths.

Removing non-numeric values: Non-numeric values like ":" were removed.

Formatting datatype: The datatypes were converted to integers.

Converting to time series: structured the data into a time series object with annual frequency.

```
# Cleaning and preparing the data for time series analysis.
death_uk <- death_uk %>%
  rename(no_of_deaths =`Number of deaths: United Kingdom`) %>%
  filter(no_of_deaths != ':') %>%
  arrange(Year) %>%
  select(no_of_deaths)
```

```
# Converting the data to a time series
death_uk$no_of_deaths <- as.integer(death_uk$no_of_deaths)
death_uk_ts = ts(death_uk, frequency = 1, start = 1887)
death_uk_ts
```

*Figure **Error! No text of specified style in document.**.2 code converting the data frame to a time series*

1.2.2 Exploratory Data Analysis

With the fully preprocessed data, a time series plot was generated.

```
# Plotting the initial number of deaths from year 1887 - 2021
plot.ts(death_uk_ts, main='Time series of number of deaths in UK (1887 -2021)')
```

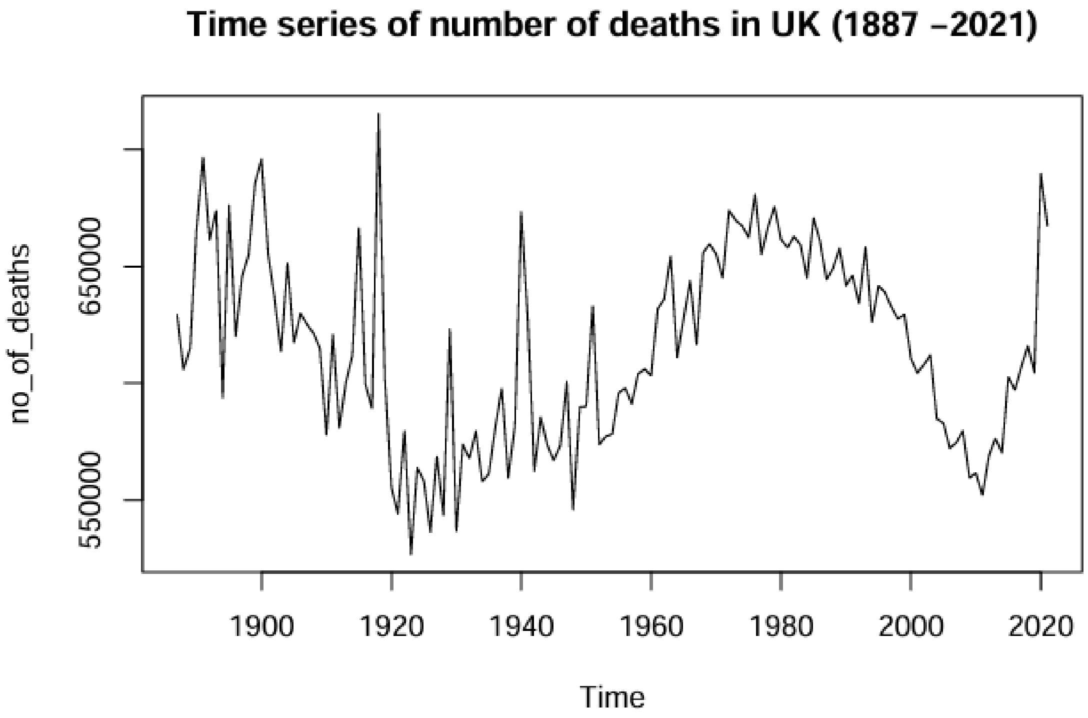

*Figure **Error! No text of specified style in document.**.3 Time series plot of the number of deaths in the UK from 1871 to 2021*

This series appears to be non-seasonal, with long-term trends that show periods of increase and decrease in the number of deaths in the UK. Irregular fluctuations can be noticed that correspond with historical events like WWI (1914 -1918), the influenza pandemic (1918), and WWII (1939 - 1945).

An additive model will be suitable to describe this series because it captures long-term trends and irregular components.

1.2.3 Decomposing Time Series

As discovered in the EDA, the UK Deaths time series is a non-seasonal time series that has both trend and irregular components. Smoothing methods are commonly used to estimate this series type (Coghlan, 2018). A 10-year simple moving average (SMA) was used to smoothen the short-term fluctuations, to reveal the underlying trend.

```
# Estimating trend by smoothing using a simple moving average of order 10.
plot.ts(SMA(death_uk_ts, n=10),
        main='Time series showing trend of number of deaths in UK',
        ylab = 'no of deaths (SMA)')
```

*Figure **Error! No text of specified style in document.**.4 Code for smoothing the time series with an order of 10*

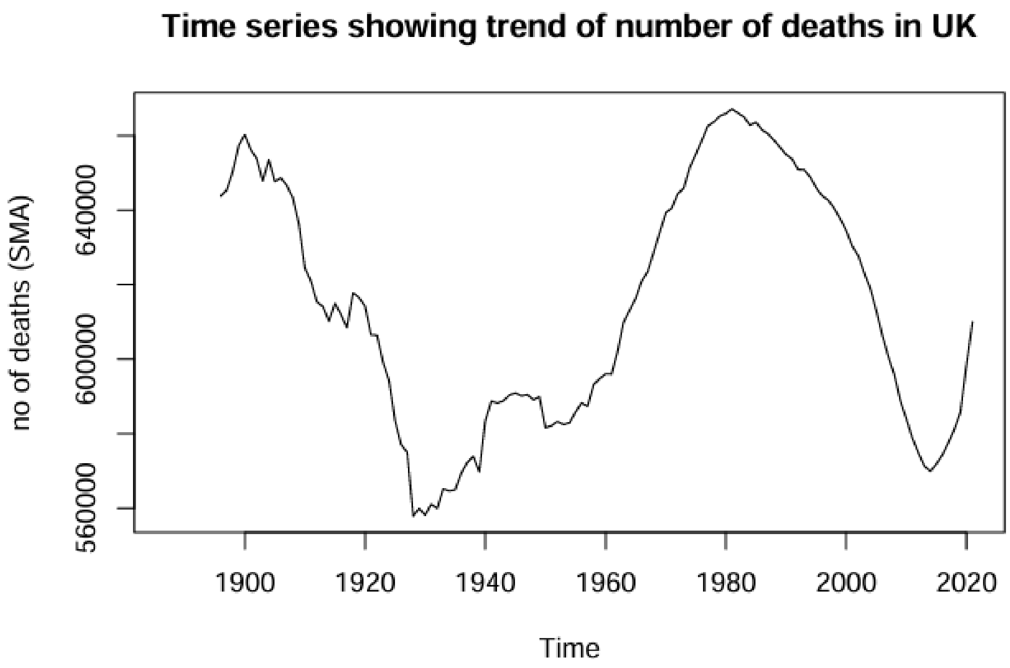

*Figure **Error! No text of specified style in document.**.5 Smoothened time series of deaths in the UK*

After smoothing we can see the following trends:

- Decline in deaths from 1887 – 1920s.
- A sharp spike in deaths in 1914 – 1920 (WWI and Influenza pandemic).
- Gradual increase in deaths from 1920s – 1960s.
- Peaks in 1980 (Aging population of the UK).
- Decline again from 1980s – 2000s (Advancements in healthcare).
- Increase again from the 2010s – 2020s.

1.3 Time Series Modeling

The main aim of this analysis is to forecast the annual number of deaths in the UK, this section focuses on building and evaluating the models that will be used for this forecast. This will be achieved by selecting the appropriate models:

- Holt-Winters Exponential Smoothing
- ARIMA Models

1.3.1 Model 1 – Holt-Winters Exponential Smoothing

Holt-Winters Exponential Smoothing is suitable for this time series data because it can be modelled using an additive approach (Coghlan, 2018). While the data lacks seasonality, it shows clear trends.

```
# Fitting a predictive model using HoltWinters()
death_uk_ts_forcast <- HoltWinters(death_uk_ts, gamma = FALSE)
death_uk_ts_forcast
```

*Figure **Error! No text of specified style in document.**.6 code to fit the predictive model using HoltWinters()*

```
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = death_uk_ts, gamma = FALSE)

Smoothing parameters:
 alpha: 0.4796249
 beta : 0.1534181
 gamma: FALSE

Coefficients:
        [,1]
a 662949.045
b   9829.915
```

Figure ***Error! No text of specified style in document.****.7 Output of the fitted Holt-Winters model*

Fitted Results: An alpha value of approximately 0.48, means that 48% of the weight is given to the most recent observation when estimating the level. A beta value of 0.15 means more weight (85%) is given to past trends during slope estimation. The trend shows an increase of 9,830 deaths/year.

```
# plotting both observed and fitted data from HoltWinters forecast.
plot(death_uk_ts_forcast)
```

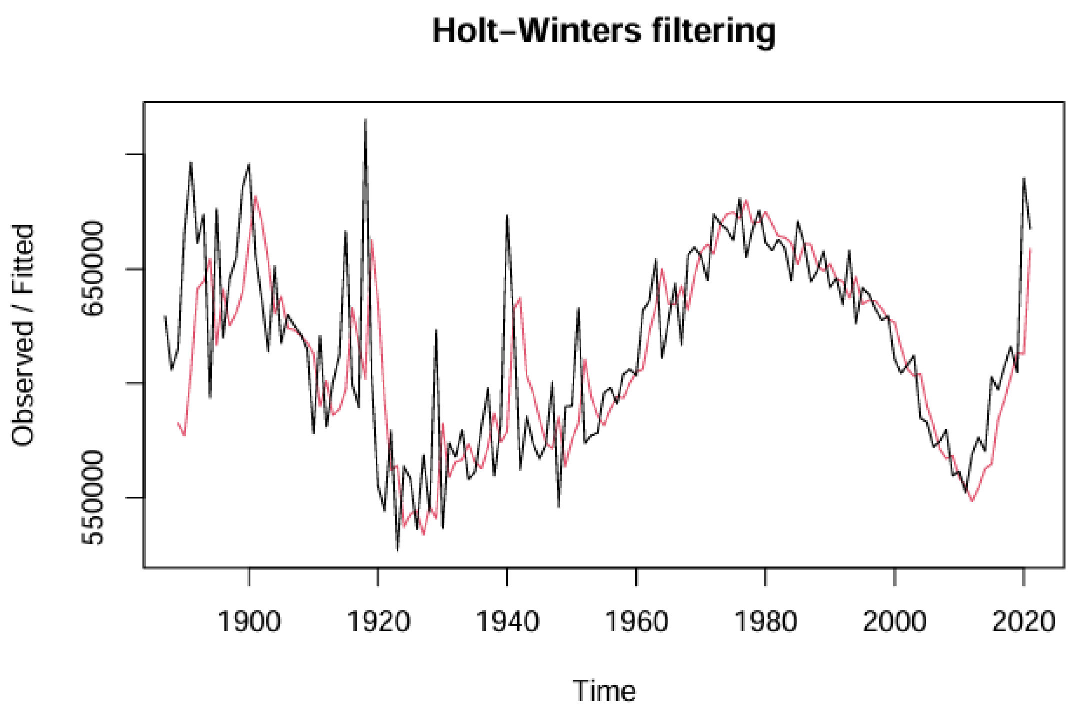

Figure ***Error! No text of specified style in document.****.8 Plot showing observed and fitted data from the Holt-Winters model*

10-year Forecast: The Model predicted a steady rise in deaths, with values ranging from ~673,000 to ~761,000 for the next 10 years. The forecast is represented with a blue line, with 80% and 95% prediction intervals in purple and grey-shaded areas respectively.

```r
# forecasting the next 10 years.
death_uk_ts_forcast2 <- forecast(death_uk_ts_forcast, h=10)
plot(death_uk_ts_forcast2)
```

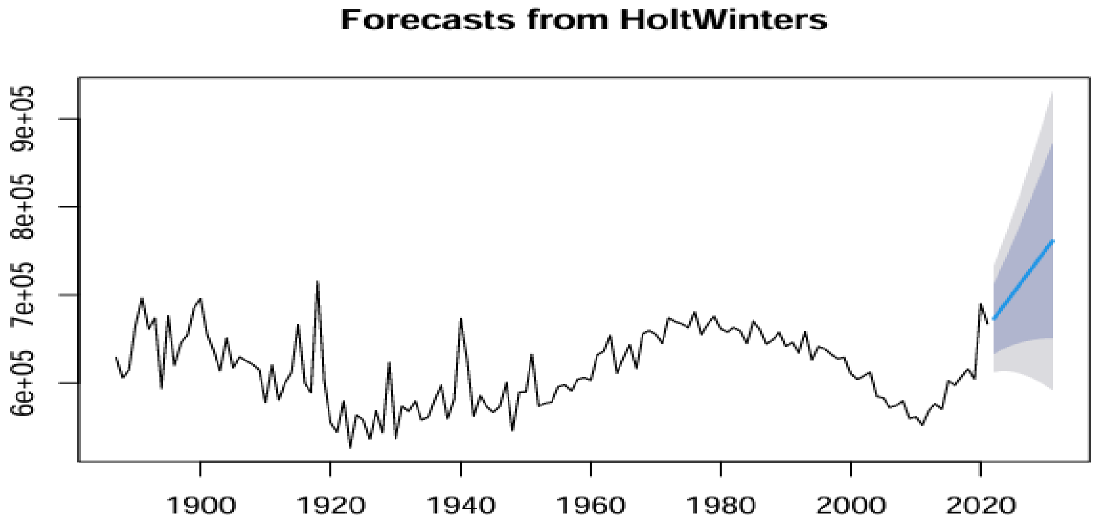

Figure ***Error! No text of specified style in document.****.9 Plot showing the 10-year forecast of the Holt-winters model*

Evaluation Metrics: The model was evaluated by checking the sum of square errors (SSE)and calculating the root mean square error (RMSE) and mean average percentage error (MAPE).

```r
# Sum od square error
death_uk_ts_forcast$SSE
```

```r
# Root Mean Square Error for Holt-Winters
RMSE_HW = sqrt(mean(death_uk_ts_forcast2$residuals^2, na.rm = TRUE))
RMSE_HW
```

```
[1] 30940.96
```

```r
# Mean absolute percentage error MAPE for Holt-Winters
MAPE_HW = mean((abs(death_uk_ts_forcast2$residuals/death_uk_ts)*100), na.rm=TRUE)
MAPE_HW
```

```
[1] 3.537774
```

Figure ***Error! No text of specified style in document.****.10 SSE, RMSE and MAPE of the Holt-Winters model*

The model had an SSE of **127,326,590,931**, RMSE of **30,940.96** and a MAPE of **3.54%,** indicating a good model performance.

1.3.1.1 Model Evaluation

The model was evaluated by checking if the forecast errors show non-zero autocorrelations, and are normally distributed with mean zero and constant variance. This was achieved by:

- Calculating Correlogram using Autocorrelation Function (ACF).
- Testing for significant evidence of non-zero correlations using the Ljung-Box test.
- Plotting the residual time series of the model.
- Plotting a histogram of the forecast errors overlaid with a normal curve with mean zero and constant variance.

ACF: The correlogram shows no significant autocorrelations since none of the lags crossed the significance bounds.

```r
# ACF and Ljung box test
acf(death_uk_ts_forcast2$residuals, lag.max=20 , na.action = na.pass)
```

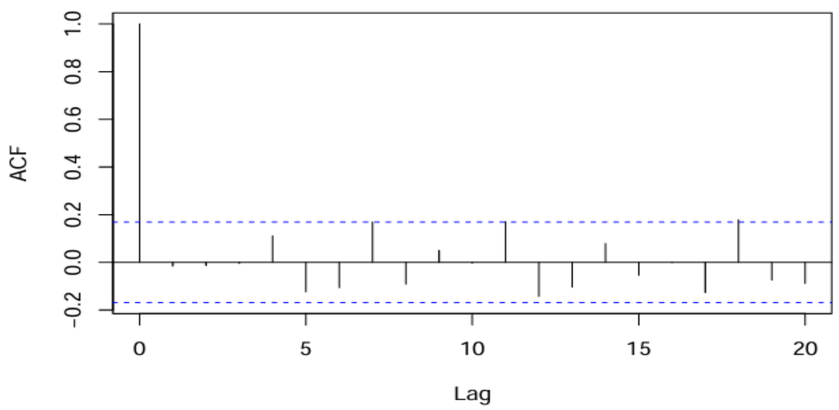

Figure ***Error! No text of specified style in document.****.11 Correlogram of the residuals in the Holt-Winters model*

Ljung–Box Test: The P-value for the LJung-test is 0.055, there is little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

```r
Box.test(death_uk_ts_forcast2$residuals, lag=20, type="Ljung-Box")
```

```
 Box-Ljung test

data:  death_uk_ts_forcast2$residuals
X-squared = 31.015, df = 20, p-value = 0.055
```

Figure ***Error! No text of specified style in document.****.12 code and output of the Ljung-box test for the Holt-winters model*

Residual Time Series plot: the plot shows an approximate mean of zero with roughly constant variance.

```r
#  plotting the forecast errors to check for constant variance
plot.ts(death_uk_ts_forcast2$residuals, main='Holt-Winters Forecast Errors',
        ylab = 'Residuals')
```

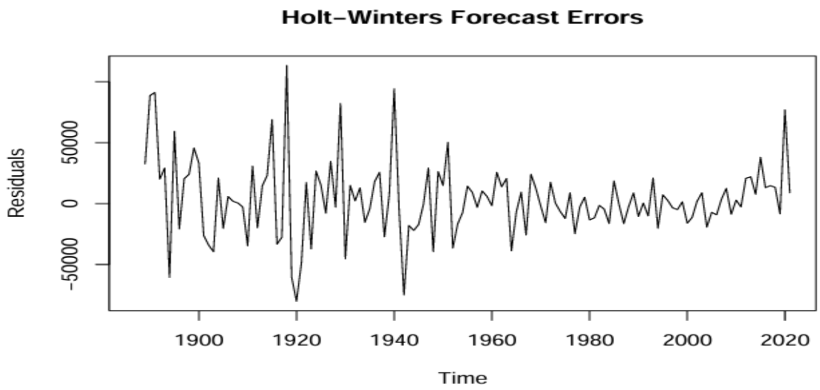

*Figure **Error! No text of specified style in document.**.13 Residual time series plot for the Holt-Winters model*

Histogram of Forecast Errors: The forecast errors appear normally distributed with a relatively small right skew. The mean is closer to zero therefore we can probably assume a normal distribution.

```
# plotting if the forecast errors to check if normally distributed with mean zero
plotForecastErrors(death_uk_ts_forcast2$residuals)
```

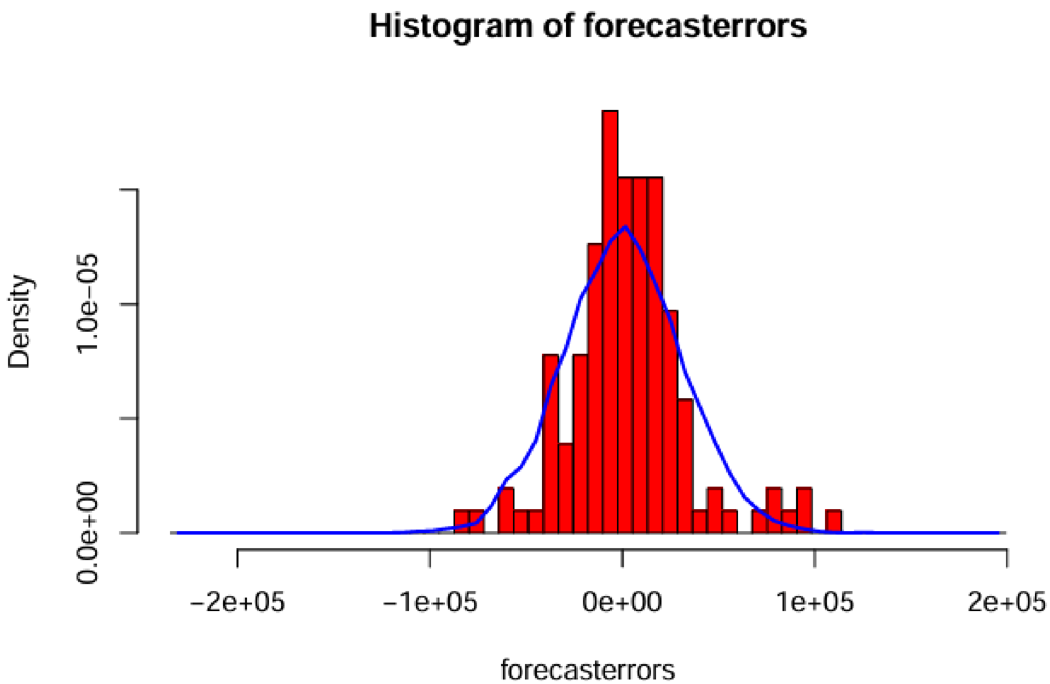

*Figure **Error! No text of specified style in document.**.14 Histogram showing forecast errors of the Holt-Winters model*

Model 1, using Holt-Winters Exponential smoothing provides an adequate predictive model for the number of deaths in the United Kingdom.

1.3.2 Autoregressive Integrated Moving Average (ARIMA) Models

Before using an ARIMA model, a stationarity test must be performed because ARIMA models are built for stationary time series i.e. no trend and seasonality. A stationary time series has a constant mean, variance, and covariance. The augmented Dickey-Fuller test is used to test if the number of deaths in the UK time series is stationary.

```r
# Augmented Dickey-Fuller Test
adf.test(death_uk_ts)
```

```
 Augmented Dickey-Fuller Test

data:  death_uk_ts
Dickey-Fuller = -2.3315, Lag order = 5, p-value = 0.4386
alternative hypothesis: stationary
```

*Figure **Error! No text of specified style in document.**.15 Augmented Dickey-Fuller test*

The augmented Dickey-Fuller test has a null hypothesis that the series is not stationary. Since the P-value of 0.4386 is greater than 0.05, we fail to reject this hypothesis. A differencing function can be used to achieve stationarity in the time series.

```r
# Differencing the time series to make it stationary
death_uk_ts_diff1 <- diff(death_uk_ts, differences = 1)
```

```r
# Plotting the series with difference 1.
plot(death_uk_ts_diff1, main='Time series of number of deaths in UK (DIFF 1)')
```

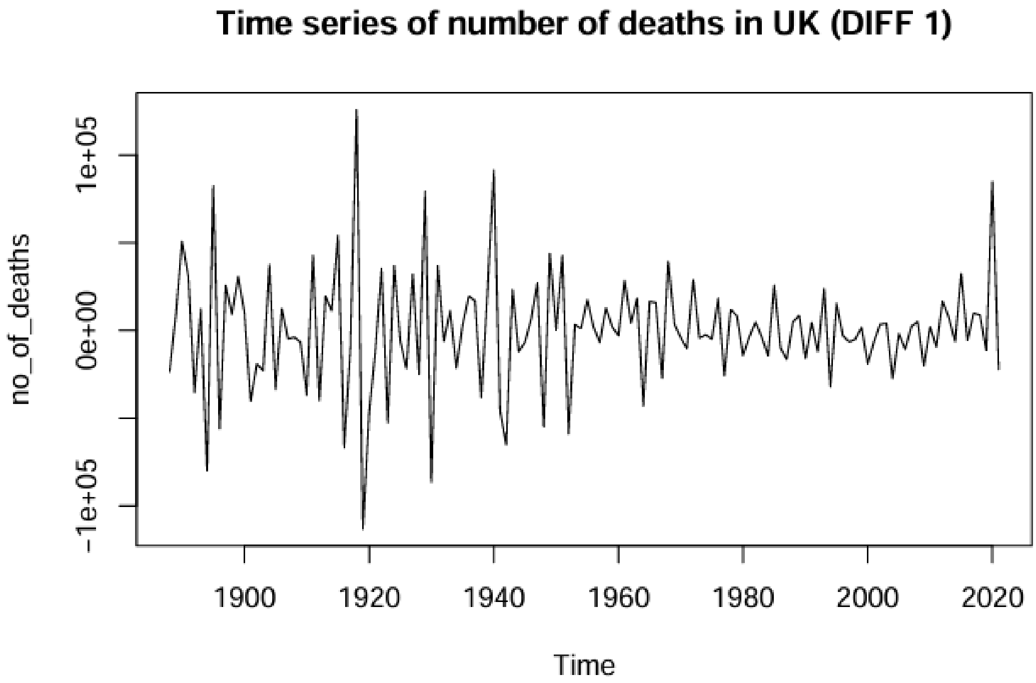

*Figure **Error! No text of specified style in document.**.16 Time series plot of the number of deaths in the UK (diff 1)*

Figure **Error! No text of specified style in document.**.16 Appears to have a stationary mean, the augmented Dickey-Fuller test can be carried out to verify.

```r
# Augmented Dickey-Fuller Test for difference 1
adf.test(death_uk_ts_diff1)
```

```
 Augmented Dickey-Fuller Test

data:  death_uk_ts_diff1
Dickey-Fuller = -6.1896, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

*Figure **Error! No text of specified style in document.**.17 Augmented Dickey-Fuller test of DIFF 1*

The P-value is less than 0.05, therefore we reject the null hypothesis and assume difference 1 is stationary.

1.3.2.1 Selecting ARIMA model

To appropriately select an ARIMA model, it is important to examine the correlogram and partial correlogram of the stationary time series (Coghlan, 2018).

```r
# Plotting the correlogram for diff1
acf(death_uk_ts_diff1, lag.max = 20)
```

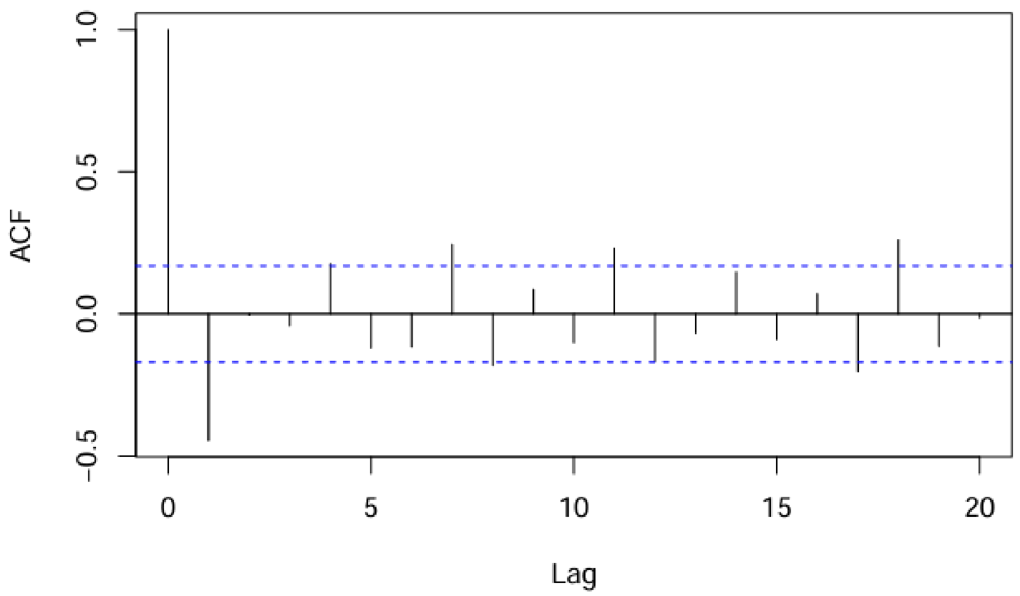

*Figure **Error! No text of specified style in document.**.18 Correlogram of the DIFF 1 time series*

From the correlogram in Figure **Error! No text of specified style in document.**.18, the autocorrelation at lag 1 (-0.446) exceeds the significance bounds. A moving average model of order 1 can be used, which is also an **ARIMA (0,1,1)** that has been differenced once.

```r
# Plotting the partial correlogram for diff1
pacf(death_uk_ts_diff1, lag.max = 20)
```

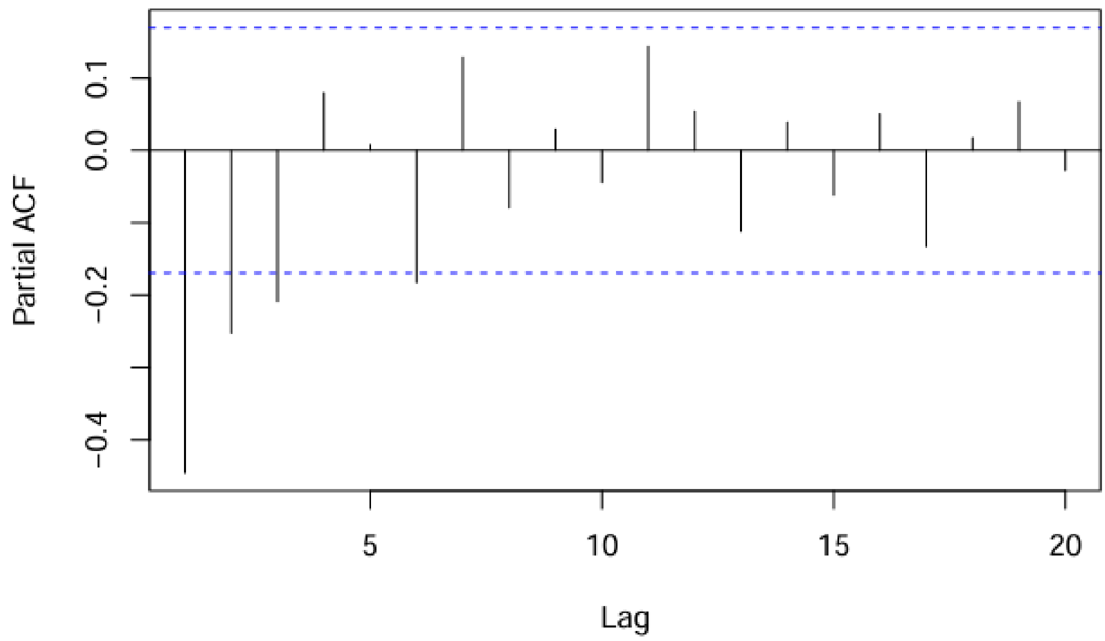

*Figure **Error! No text of specified style in document.**.19 Partial correlogram for the DIFF 1 time series*

The partial autocorrelation at lags 1,2, and 3 exceeds the significance bounds, an auto-regressive model of order 3 is possible. **ARIMA (3,1,0).**

From the principle of parsimony (fewer is better).

1.3.3 Model 2- Moving Average Model of Order 1- ARIMA (0,1,1)

This model ARIMA (0,1,1) is a model with no autoregressive terms, it has been differenced once to make it stationary and indicates that the current value of the time series depends on the past errors.

$$X_t' = \mu + \theta_1 \epsilon_{t-1} + \epsilon_t$$

- X_t': Differenced value, $X_t - X_{t-1}$
- μ: Constant term (mean of zero since it's differenced)
- θ_1: co-efficient for moving average of lag 1.
- ϵ_t : Random error term at time t

```
# Moving average model of order 1 and difference 1.
death_uk_ts_ma <- arima(death_uk_ts, order = c(0,1,1))
death_uk_ts_ma
```

```
Coefficients:
          ma1
      -0.5913
s.e.   0.0695

sigma^2 estimated as 812067647:  log likelihood = -1564.86,  aic = 3133.73
```

*Figure **Error! No text of specified style in document.**.20 ARIMA (0,1,1) model*

After fitting the model, the equation becomes $X_t' = -0.5913\,_1\epsilon_{t-1} + \epsilon_t$

10-year Forecast: The model predicted a stable number of deaths (~650,000) for the next 10 years. The forecast is represented with a blue line, with 80% and 95% prediction intervals in purple and grey-shaded areas respectively.

```
# forecasting the next 10 years using moving average.
death_uk_ts_ma_forecast <- forecast(death_uk_ts_ma, h =10)
death_uk_ts_ma_forecast
```

*Figure **Error! No text of specified style in document.**.21 10-year forecast using ARIMA (0,1,1) model*

```
# 10 year forecast plot
plot(death_uk_ts_ma_forecast)
```

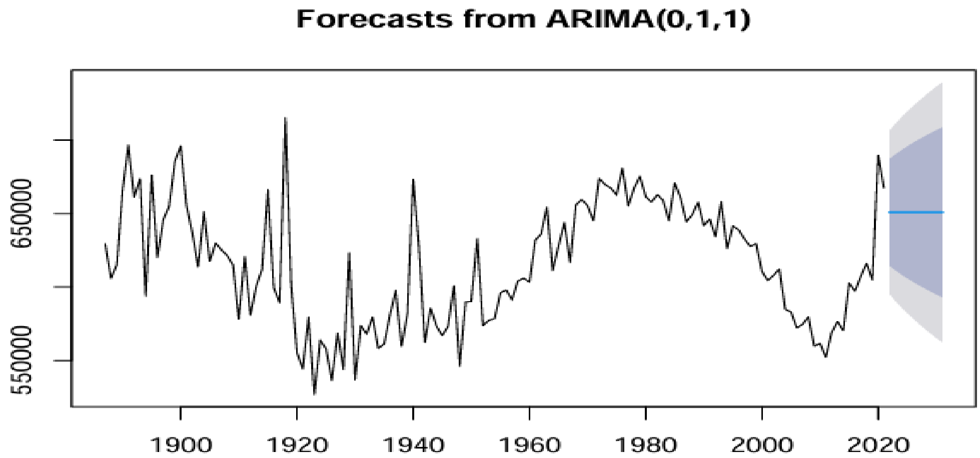

*Figure **Error! No text of specified style in document.**.22 Plot showing the 10-year forecast of the ARIMA (0,1,1) model*

Evaluation Metrics: The model was evaluated by checking the AIC, RMSE, and MAPE.

```
# Evaluation for ARIMA(0,1,1)
AIC(death_uk_ts_ma)
```

```
[1] 3133.728
```

```
accuracy(death_uk_ts_ma)
```

```
       ME     RMSE      MAE         MPE     MAPE     MASE
456.7854 28391.11 20199.13 -0.09581522 3.280053 0.858626
```

*Figure **Error! No text of specified style in document.**.23 code and output of the evaluation of ARIMA (0,1,1)*

The model has an AIC of **3133**, RMSE of **28391.11**, and MAPE of **3.28%** indicating a good model performance.

1.3.3.1 Model Evaluation

ACF: The correlogram shows no significant autocorrelations since none of the lags crossed the significance bounds.

```r
# ACF and Ljung box test
acf(death_uk_ts_ma_forecast$residuals, lag.max=20 , na.action = na.pass)
```

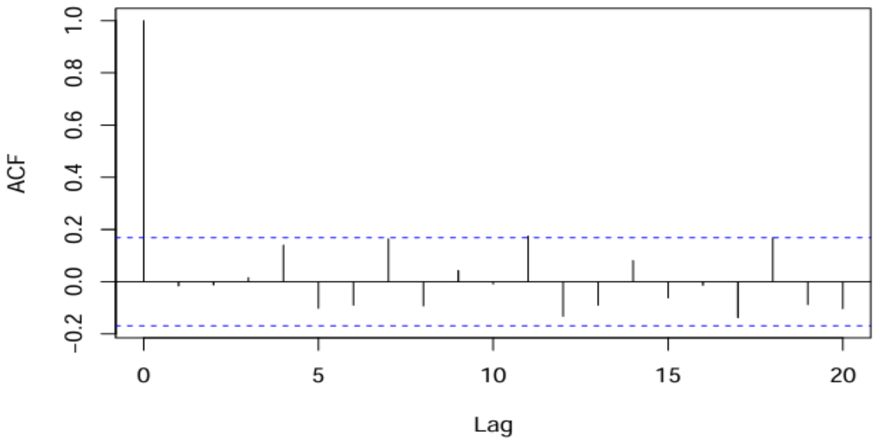

*Figure **Error! No text of specified style in document.**.24 Correlogram of the residuals in the ARIMA (0,1,1) model*

Ljung–Box Test: The P-value for the LJung-test is 0.052, there is little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

```r
Box.test(death_uk_ts_ma_forecast$residuals, lag=20, type="Ljung-Box")
```

```
 Box-Ljung test

data:  death_uk_ts_ma_forecast$residuals
X-squared = 31.243, df = 20, p-value = 0.05206
```

*Figure **Error! No text of specified style in document.**.25 Ljung box test for ARIMA (0,1,1)*

Residual Time Series plot: the plot appears to have mean zero and constant variance

```r
#  plotting the forecast errors to check for constant variance
plot.ts(death_uk_ts_ma_forecast$residuals, main = 'ARIMA(0,1,1) Forecast Errors',
        ylab = 'Residuals')
```

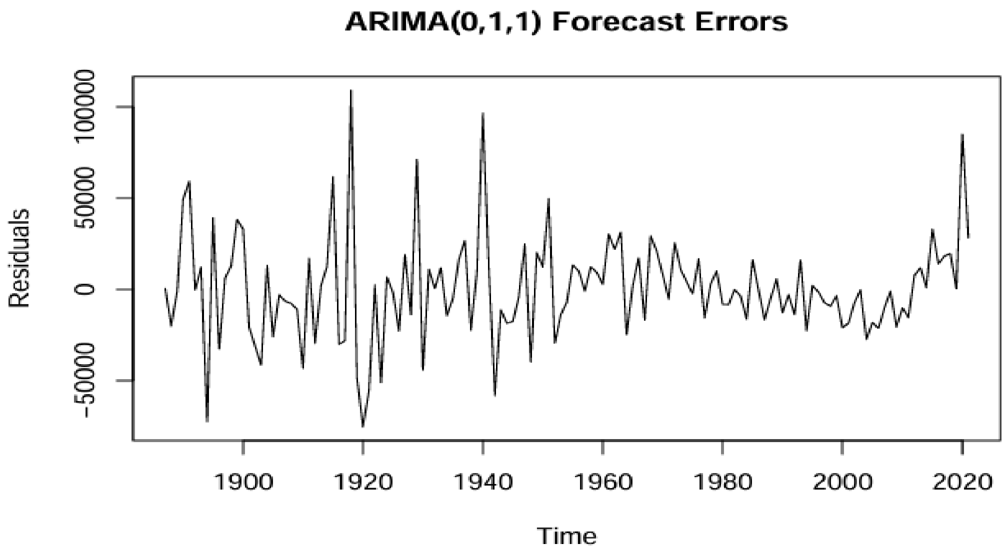

Figure ***Error! No text of specified style in document.****.26 plot showing residual forecast for ARIMA (0,1,1)*

Histogram of Forecast Errors: The forecast errors appear normally distributed with a mean of zero therefore we can probably assume a normal distribution.

```
# plotting if the forecast errors to check if normally distributed with mean zero
plotForecastErrors(death_uk_ts_ma_forecast$residuals)
```

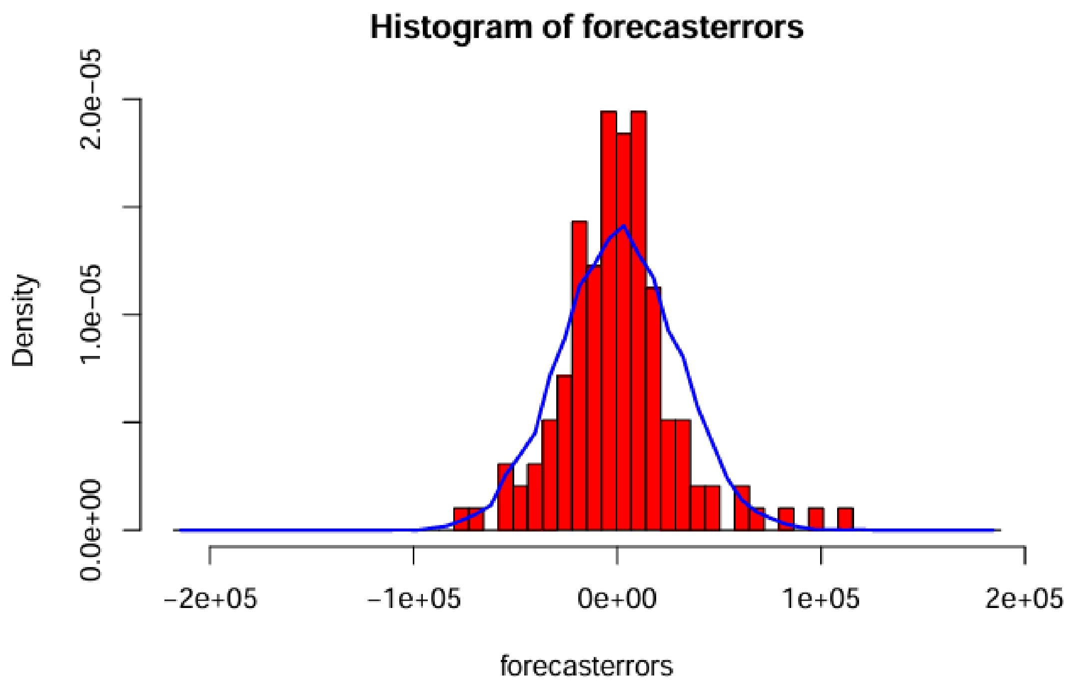

Figure ***Error! No text of specified style in document.****.27 Histogram of the residual forecast for the ARIMA (0,1,1) model*

1.3.4 Model 3- Autoregressive Model of Order 3- ARIMA (3,1,0)

This model ARIMA (3,1,0) is an autoregressive model with no moving average that focuses on the relationships between the past 3 values, of the differenced time series to forecast.

$$X_t' = \phi_1 X_{t-1}' + \phi_2 X_{t-2}' + \phi_3 X_{t-3}' + \epsilon_t$$

- X_t': Differenced value, $X_t - X_{t-1}$
- ϕ_1, ϕ_3, ϕ_3 : Coefficients of first, second, and third lags.
- ϵ_t : Random error term at time t

```
# ARIMA model (3,1,0)
death_uk_ts_ar <- arima(death_uk_ts, order = c(3,1,0))
death_uk_ts_ar
```

```
Coefficients:
          ar1      ar2      ar3
      -0.6108  -0.3798  -0.2166
s.e.   0.0844   0.0968   0.0871

sigma^2 estimated as 794265312:  log likelihood = -1563.41,  aic = 3134.83
```

*Figure **Error! No text of specified style in document.**.28 ARIMA (3,1,0) model*

The equation becomes $X_t' = -0.6108X_{t-1}' - 0.3798X_{t-2}' - 0.2166X_{t-3}' + \epsilon_t$ after fitting the model.

10-year Forecast: The model predicted an increase in number of deaths (~650,000 - ~659,000) for the next 10 years. The forecast is represented with a blue line, with 80% and 95% prediction intervals in purple and grey-shaded areas respectively.

```
# forecasting the next 10 years.
death_uk_ts_ar_forecast <- forecast(death_uk_ts_ar, h =10)
death_uk_ts_ar_forecast
```

```
# 10 year forecast plot
plot(death_uk_ts_ar_forecast)
```

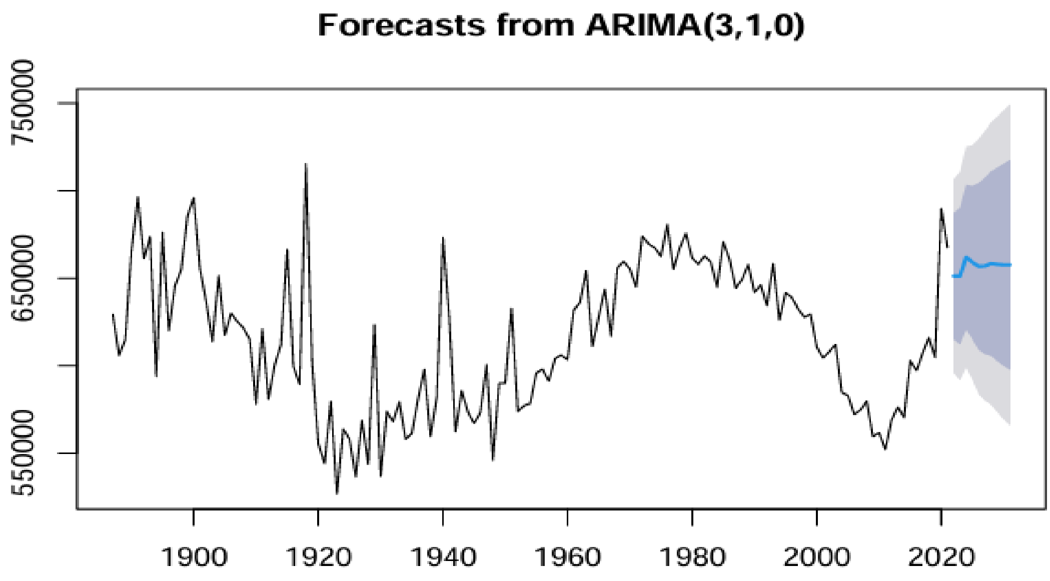

*Figure **Error! No text of specified style in document.**.29 10-year forecast plot for the ARIMA (3,1,0) model*

Evaluation Metrics: The model was evaluated by checking the AIC, RMSE, and MAPE.

```r
# Evaluation for ARIMA(3,1,0)
AIC(death_uk_ts_ar)
```

```
[1] 3134.827
```

```r
accuracy(death_uk_ts_ar)
```

```
     ME     RMSE      MAE         MPE     MAPE      MASE
535.029 28078.19 20038.88 -0.07547298 3.256082 0.8518145
```

The model has an AIC of **3134**, RMSE of **28078.19**, and MAPE of **3.26%** indicating a good model performance.

1.3.4.1 Model Evaluation

ACF: The correlogram shows no significant autocorrelations since none of the lags crossed the significance bounds.

```r
# ACF and Ljung box test
acf(death_uk_ts_ar_forecast$residuals, lag.max=20 , na.action = na.pass)
```

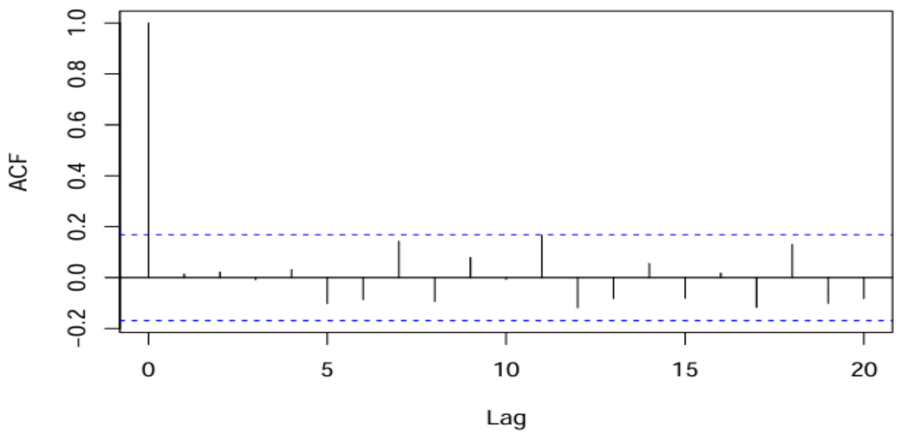

*Figure **Error! No text of specified style in document.**.30 Correlogram of the residuals in the ARIMA (3,1,0) model*

Ljung–Box Test: The P-value for the LJung-test is 0.2342, there is little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

```r
Box.test(death_uk_ts_ar_forecast$residuals, lag=20, type="Ljung-Box")
```

```
 Box-Ljung test

data:  death_uk_ts_ar_forecast$residuals
X-squared = 24.19, df = 20, p-value = 0.2342
```

*Figure **Error! No text of specified style in document.**.31 Ljung box test for ARIMA (3,1,0)*

Residual Time Series plot: the plot appears to have mean zero and constant variance

```r
# plotting the forecast errors to check for constant variance
plot.ts(death_uk_ts_ar_forecast$residuals,main ='ARIMA(3,1,0) Forecast Errors',
        ylab = 'Residuals')
```

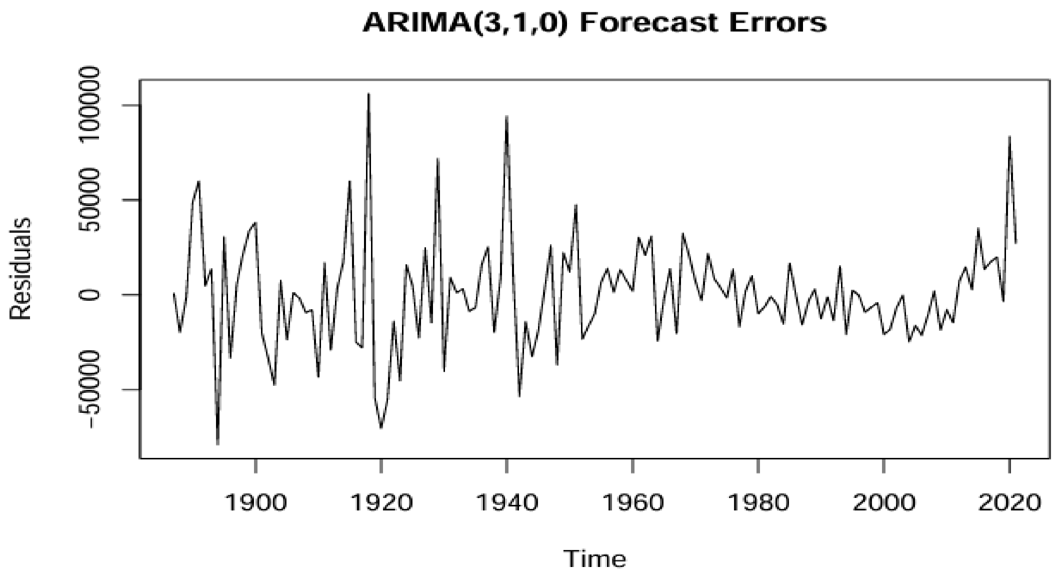

*Figure **Error! No text of specified style in document.**.32 plot showing residual forecast for ARIMA (3,1,0)*

Histogram of Forecast Errors: The forecast errors appear normally distributed with a mean of zero therefore we can probably assume a normal distribution.

```
# plotting if the forecast errors to check if normally distributed with mean zero
plotForecastErrors(death_uk_ts_ar_forecast$residuals)
```

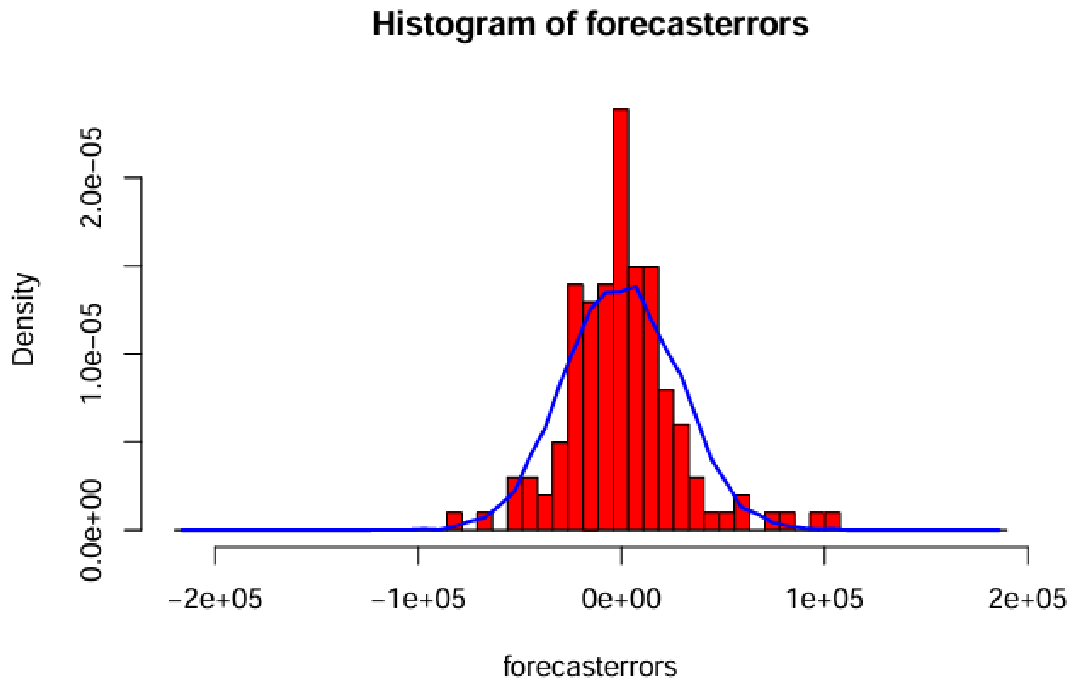

*Figure **Error! No text of specified style in document.**.33 Histogram of the residual forecast for the ARIMA (3,1,0) model*

1.4 Model Comparison

*Table **Error! No text of specified style in document.**.1 Table comparing the three models*

MODEL	RMSE	MAPE	AIC
HOLT-WINTERS	30,940	3.54%	N/A
ARIMA(0,1,1)	28,391	3.28%	**3,134**
ARIMA(3,1,0)	**28,078**	**3.26%**	3,135

1.4.1 Detailed Comparison

1.4.1.1 Holt-Winters Exponential Smoothing

- Strengths: Very effective at capturing long-term trends and simple to interpret.
- Weakness: Higher RMSE (30,940), and MAPE (3.54%) compared to other models.

1.4.1.2 ARIMA (0,1,1)

Strengths:

- Lower RMSE (28,391) than Holt-Winters
- Simple model with very minimal parameters (MA(1)), following the principle of parsimony (Coghlan, 2018).
- Best AIC score among the two ARIMA models.
- No autocorrelation in residuals.

Weaknesses: Higher RMSE (28,391), and MAPE (3.28%) compared to ARIMA(3,1,0).

1.4.1.3 ARIMA (3,1,0)

Strengths:

- The lowest RMSE and MAPE of the three models
- AIC (3135) is very close to ARIMA (0,1,1)
- No autocorrelation in residuals.

Weaknesses: Slightly more complex, with three autoregressive terms.

1.5 Key findings, Recommendations and conclusion

1.5.1 Key Findings

The forecast of the annual deaths in the UK from 2022 to 2031 predicts a steady increase in the number of deaths.

*Table **Error! No text of specified style in document.**.2 10-year forecast for each model*

MODEL	PREDICTIONS
HOLT-WINTERS	~673,000 to ~761,000
ARIMA (0,1,1)	~650,000
ARIMA (3,1,0)	~650,000 to ~659,000

Although ARIMA(3,1,0) has a slightly lower RMSE and MAPE than ARIMA(0,1,1), ARIMA(0,1,1) was chosen as the best model because its performance was very similar and comparable to ARIMA(3,1,0) and it has a simpler model structure. By producing a similar forecasting accuracy with fewer parameters, ARIMA(0,1,1) follows the principle of parsimony (Coghlan, 2018), making it the more efficient and interpretable choice for this analysis.

1.5.2 Recommendations

- Forecast results can be used in health care planning, by effectively allocating resources to combat rising death rates.
- Preventive measures can be implemented via policies tailored to reduce mortality rates.

1.5.3 Conclusion

This study provided a solid 10-year forecast of deaths in the UK, with the ARIMA (0,1,1) model identified as the best-performing model. These results can guide healthcare planning and policies to address the increase in mortality rates.