**Time Series Analysis**

**Predicting the Number of Marriages in the UK**

**Contents**

[1. Time Series Modelling 1](#_Toc208236950)

[1.1 Introduction 1](#_Toc208236951)

[1.2 Exploratory Data Analysis 1](#_Toc208236952)

[1.3 Data Cleaning and Preprocessing 2](#_Toc208236953)

[1.4 Time Series Modelling and Evaluation 5](#_Toc208236954)

[1.5 Conclusion 15](#_Toc208236955)

[References 16](#_Toc208236956)

# Time Series Modelling

## Introduction

This task involves examining patterns in historical data organised sequentially over time using time series analysis to identify components such as trend and seasonality to accurately predict the number of marriages in the UK.

## Exploratory Data Analysis

Exploratory data analysis was carried out, finding a dataset comprising 167 rows spanning the years 1855 to 2021, with columns representing the number of marriages in the different nations of the United Kingdom (Figure 1).

**Figure 1**

*UK Marriage Data Overview****A screenshot of a computer

Description automatically generated***

## Data Cleaning and Preprocessing

The data cleaning and preparation involved addressing the presence of character data with ":" indicating missing values and conversion of the marriage counts to a numeric format, as well as arranging them in ascending order before converting the data into a time series object with annual frequency (Figure 2).

**Figure 2**

***A screenshot of a computer

Description automatically generated****Data Cleaning and Preprocessing Summary*

Subsequently, the time series was plotted, showing notable patterns that were consistent with sociocultural changes, such as an increase in marriages around the two world wars and a continuous downward trend in the number of marriages since the early 1970s. (Figure 3)

**Figure 3**

*Historical Plot of UK MarriagesA graph with lines and numbers

Description automatically generated*

Then we carried out smoothing, a technique to reduce random fluctuations caused by noise in the data, making it easier to observe its underlying trends and patterns and model, using the Simple Moving Averages (SMA) technique which produces a smoother line by calculating the average of data points within a defined window size (n) and then shifting this window across the time series (Tsay, 2005). We plotted the smoothed line with three different “n” values, including 3, 5 and 10, finding that 5 produced a balanced smoothed line capturing the main trends without unnecessary noise (Figure 4).

**Figure 4**

*Smoothing with Different Window SizesA screenshot of a computer screen

Description automatically generated*

Lastly, we visualised the original time series data alongside the smoothed trend line (Figure 5)

**Figure 5**

*Comparison of Smoothed and Original DataA graph of marriage in the united kingdom

Description automatically generated*

## Time Series Modelling and Evaluation

Moving to the modelling stage, we chose ARIMA (AutoRegressive Integrated Moving Average) to model the time series due to its suitability as our data series in the previous steps showed to be linear, non-stationary and lacking seasonality. ARIMA models the time series by generating a function to capture the patterns produced by its data points. This function has three components, AutoRegression (AR), capturing the relationship between a data point and previous observations, Integration (I), referring to the differencing that is required to make the time series stationary and Moving Average (MA), corresponding to the relationship between observations and the prediction errors from past values incorporating them to adjust the accuracy of the model. (Newbold, Carlson, & Thorne, 2013)

We proceeded to confirm the assumptions of ARIMA models and their suitability for our time series, starting by assessing and concluding that the time series data, comprising annual data points, does not have a seasonality component, as this is defined as short-term fluctuations within one year. Additionally, we tested for stationarity using the Augmented Dickey-Fuller (ADF), which evaluates the null hypothesis that the time series is non-stationary finding a Dickey-Fuller statistic of -1.0651 and a p-value of 0.9247, greater than the threshold of 0.05 indicating that the null hypothesis cannot be rejected hence the time series was not stationary, requiring a transformation to make the data suitable for modelling (Figure 6).

**Figure 6**

*Augmented Dickey-Fuller Test for Stationarity*

A screenshot of a computer

Description automatically generated

Then we addressed the issue of non-stationarity to make the time series suitable for modelling by applying first-order differencing (Figure 7), a technique that converts the time series into a sequence of changes between consecutive observations, effectively stabilising the mean and variance over time. The technique was chosen because the raw data plot showed no indications of more complex patterns such as parabolic curves indicating changes in the rate of change, but was more consistent with a linear trend, not requiring higher orders of differencing. The plot of the differenced series (Figure 8) showed fluctuations around a constant mean with no obvious trend, consistent with the expected behaviour of a stationary time series. Additionally, this was confirmed using an Augmented Dickey-Fuller test yielding a Dickey-Fuller statistic of -6.221 and a p-value of 0.01, rejecting the null hypothesis of non-stationarity (Figure 86).

**Figure 7**

*First-Order Differencing for Stationarity****A screenshot of a computer

Description automatically generated***

**Figure 8**

*Differenced Time Series Plot*

A graph showing marriage records

Description automatically generated

To determine the function terms corresponding to the ARIMA model components, we generated and analysed the ACF (Autocorrelation Function) (Figure 9) and PACF (Partial Autocorrelation Function) (Figure 10) plots. The ACF plot, measuring the correlation between the time series and its lagged values, the values corresponding to previous observations delayed by some time steps, showed a large spike at lag 1, followed by a decline, which indicated the need for one MA term. Similarly, the PACF plot, which isolates the direct relationship between the series and its lagged values by taking away the influence of intermediate lags, showed a large spike at lag 1 followed by a gradual decline, indicating the need for one AR component.

**Figure 9**

***A graph with lines and numbers

Description automatically generated****Autocorrelation Function (ACF) Plot*

**Figure 10**

***A graph with lines and numbers

Description automatically generated****Partial Autocorrelation Function (PACF) Plot*

Subsequently, we proceeded to fit an ARIMA model with the time series data (Figure 11)

**Figure 11**

***A screenshot of a computer code

Description automatically generated****ARIMA Model Fitting Results*

To assess the efficiency of the model in capturing the patterns of the time series, we plotted its residuals (Figure 12), the difference between observed and the model’s fitted values, as they are expected to show randomness with no clear patterns, fluctuating around a zero mean, which is what was found in the plot. Similarly, we explored the correlation between the residuals, as they should show no significant correlation, which was confirmed by the AFC plot (Figure 13), indicating that most of the autocorrelations were noise, as they were within the 95% confidence interval bands. Furthermore, to confirm that the residuals were consistent with noise, we carried out the Ljung-Box test (Figure 14), which evaluates the null hypothesis that the residuals are uncorrelated, finding a p-value of 0.04289 suggesting marginal rejection of the null hypothesis, indicating some degree of autocorrelation which was deemed to be reasonably adequate for forecasting purposes given the overall results considering the patterns observed in the plots. Consequently, it was concluded that the ARIMA(1,1,1) appropriately captures the time series flow and patterns and is well-suited for the prediction of marriage levels.

**Figure 12**

*ARIMA Model Residuals Plot****A graph showing time and time

Description automatically generated***

**Figure 13**

***A graph with a line

Description automatically generated****ACF Plot for ARIMA Residuals*

**Figure 14**

***A screenshot of a computer program

Description automatically generated****Ljung-Box Test for Residual Independence*

Lastly, we plotted a ten-year prediction of marriage levels, which indicates a flat trend of marriage levels in the future (Figure 15).

**Figure 15**

*Forecast of UK Marriages (ARIMA Model)****A graph on a screen

Description automatically generated***

To compare the results of the model with a different modelling approach, we replicated the exploratory data analysis and preprocessing steps we followed for the UK as a whole marriage data, but on this occasion for the England & Wales marriage levels data, arriving at a similar time series plot (Figure 16)

**Figure 16**

*England & Wales Marriage* ***A graph on a computer screen

Description automatically generated****Historical Time Series Plot*

Given the non-stationary and non-seasonal characteristics of the data, we opted for implementing an Exponential Smoothing State Space Model (ETS) on the time series. ETS models the time series as a combination of components, including error trend and seasonality, which can be added or multiplied in a function defining the time series data points (Newbold, Carlson, & Thorne, 2013). The implementation was carried out using the “ets()” function, which automatically assigns the components based on the characteristics of the time series (Figure 17). In this case, the selected parameters indicated a multiplicative error component, no trend and no seasonality.

**Figure 17**

*ETS Model Parameters SummaryA screenshot of a computer

Description automatically generated*

To assess the quality of the model, we obtained the Residuals Plot showing no scattered residuals, randomly distributed across a zero-mean line (Figure 18) and the ACF plot indicating most of the residuals are consistent with noise, not associated with underlying patterns of the time series and within the 95% confidence bands (Figure 19). Similarly, we carried out the Ljung-Box Test, obtaining a p-value of 0.0076, indicating some residual and minor room for improvement, but not severe enough to disqualify the model, which otherwise aims to model the time series in a simple manner, capturing its most essential components, avoiding unnecessary complexities (Figure 20). Consequently, the model was deemed appropriate for forecasting the level of marriages in England & Wales.

**Figure 18**

***A graph showing the results of a model

Description automatically generated****ETS Model Residuals Plot*

**Figure 19**

*ACF Plot for ETS Residuals****A graph with a line

Description automatically generated***

**Figure 20**

*Ljung-Box Test for ETS Residuals****A screenshot of a computer

Description automatically generated***

Lastly, we performed a forecast of the marriage levels over the next 10 years, obtaining a plot consistent with a predicted flat trend (Figure 21)

**Figure 21**

*Forecast of England & Wales Marriages (ETS Model)A graph of marriage in england

Description automatically generated*

## Conclusion

We developed an ARIMA model applied to the UK marriage levels time series data and an ETS model applied to the England & Wales time series data. Both models were found to model their corresponding time series appropriately and make predictions, with residuals approximating noise and minor correlation of residuals. When comparing the Mean Standard Error (MEA) and Root Mean Square Error (RMSE), both metrics that measure the differences between observed and fitted values in the models and consequently their accuracy, the ETS model MAE of 11,786.35 and RMSE of 20,386.86 represented an advantage for this model in contrast to the ARIMA model's MAE of 14,118.03 and RMSE of 24,067.84. indicating that the ETS model was more accurate in fitting its corresponding data. This difference, perhaps, was due to the simpler approach of the ETS model, which avoided the need for differencing.

# References

Newbold, P., Carlson, W. L., & Thorne, B. M. (2013). *Statistics for Business and Economics* (8th ed.). Pearson.

Tsay, R. S. (2005). *Analysis of Financial Time Series.* Wiley.