ABDULLAH ANSARI

IDENTIFYING AN OPTIMAL TIME SERIES MODEL FOR ANNUAL MEAN TEMPERATURE IN THE MIDLANDS

INTRODUCTION:

In this report, we analyse a dataset containing the annual mean temperature in degrees Celsius for the Midlands region of England from the years 1900 to 2021. The data was sourced from the UK Meteorological Office Hadley Climate Centre and available fromwww.metoffice.gov.uk/hadobs/hadcet. The analysis aims to develop a suitable time series model for Midlands region temperature data. We'll start by describing the dataset then discuss model identification, justifying our choice and finally fit the model and some diagnostic checks.

DATA DESCRIPTION:

The dataset, named cet_temp.csv, contains two columns: year and avg_annual_temp_C. The year column represents the years from 1900 to 2021. The avg_annual_temp_C column denotes the corresponding annual mean temperature in degrees Celsius for each year in the Midlands region of England. Below are the summary statistics of the dataset calculated in R using the summary() function.

Min	1st Qu	Median	Mean	3rd Qu	Max
12.07	12.73	13.21	13.30	13.85	14.48

Now we will visualise the time series data by producing a time plot to check for stationarity, seasonality and trends, an Autocorrelation Function (ACF) plot which shows the correlation between observations in a time series at different lags and a Partial Autocorrelation Function (PACF) plot that shows the correlation between observations in a time series while removing the effects of intermediate lags.

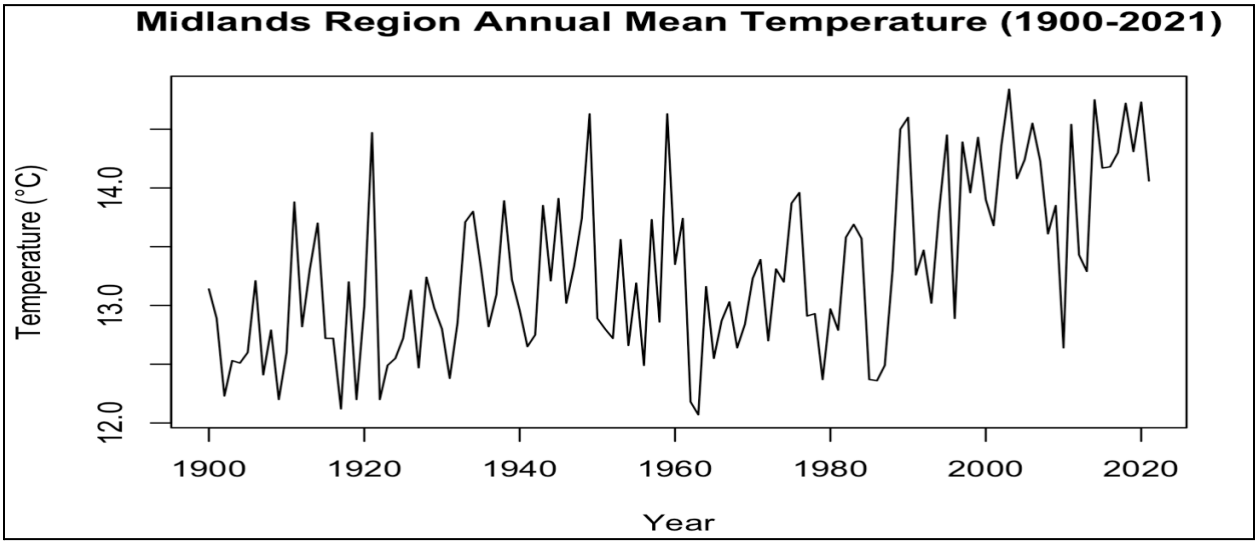


Fig1

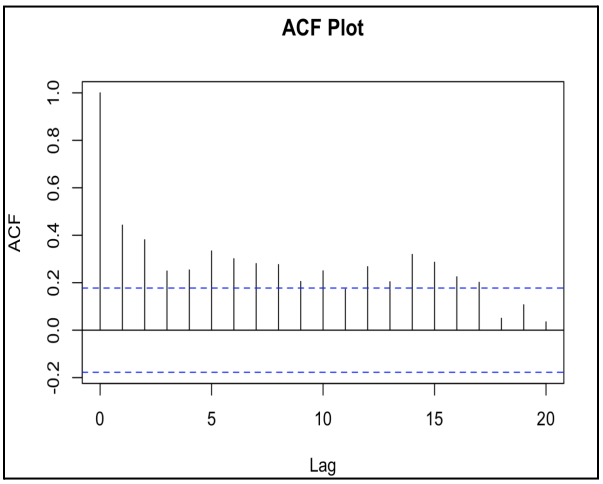


Fig2

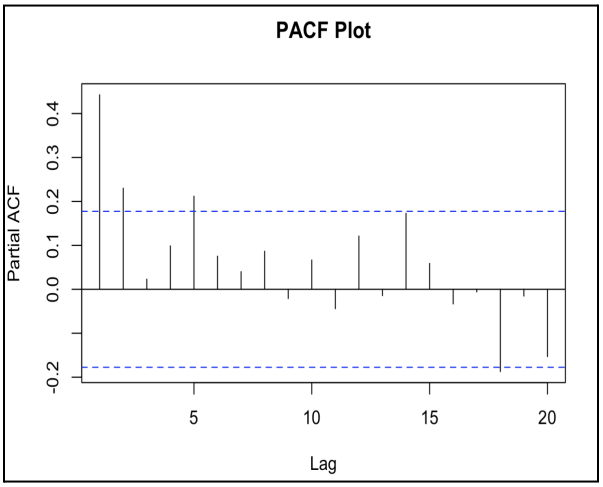


Fig3

Examining the time plot we can see a slightly positive trend and some variations over time. From the ACF and PACF plots, we can see that the sample ACF against the lag decreases initially as the lag increases, the decline to zero is not very rapid we can see a cyclic structure (for example, the ACF peak height at lag 14 is similar to that at lag 5) and so it is not certain that the series is stationary. The PACF plot doesn't offer as much insight into whether the series is stationary as the time plot and ACF plot does.

MODEL IDENTIFICATION:

The time series does not appear stationary since the ACF does not quickly decay to zero. To make it stationary, we difference the data by one season and then analyze the time plot, ACF, and PACF of the differenced series.

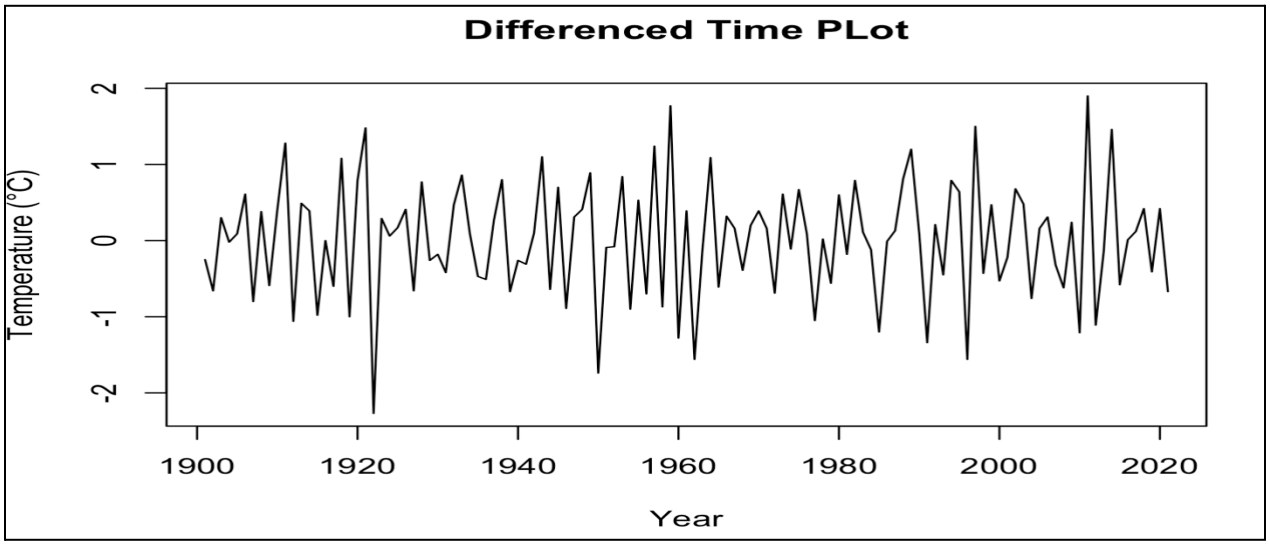


Fig4

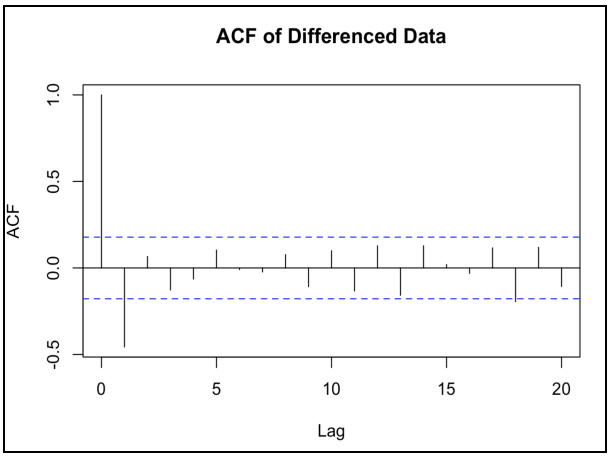


Fig5

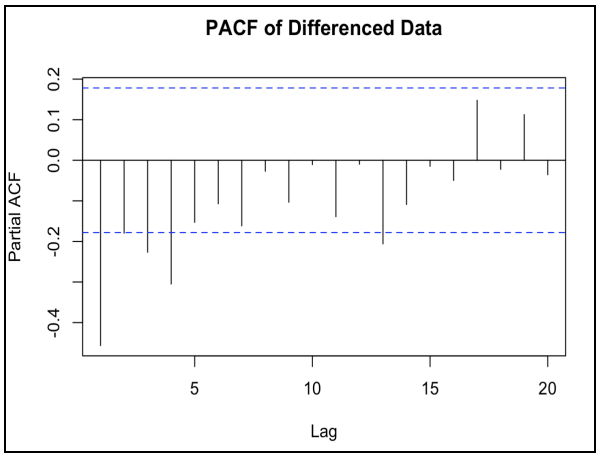


Fig6

These plots provide insights into the stationarity and autocorrelation structure of the transformed time series, guiding the selection of an appropriate time series model. Analysing the above plots, the time plot implies that the process is stationary. The plot of sample ACF cuts off to zero after lag 1 which suggests that a Moving Average model of order 1 (MA(1)) may be appropriate. From the sample PACF plot, we also see that there is a peak at lag 1, which may imply that there is an Auto-Regressive (AR) component in the process which we will further explore.

MODEL FITTING AND DIAGNOSTICS:

We will start by fitting an MA(1) model to the differenced data by using the arima() command in R and then plotting a time plot, an ACF plot of the residuals and a plot of the LjungBox test P-values for some lags.

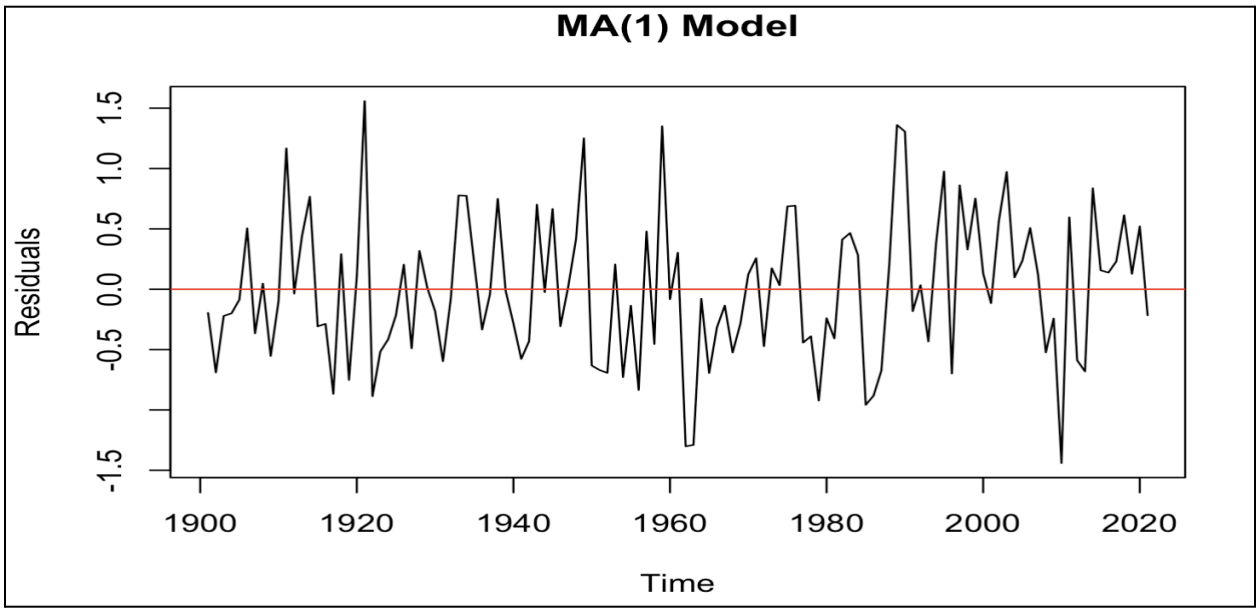
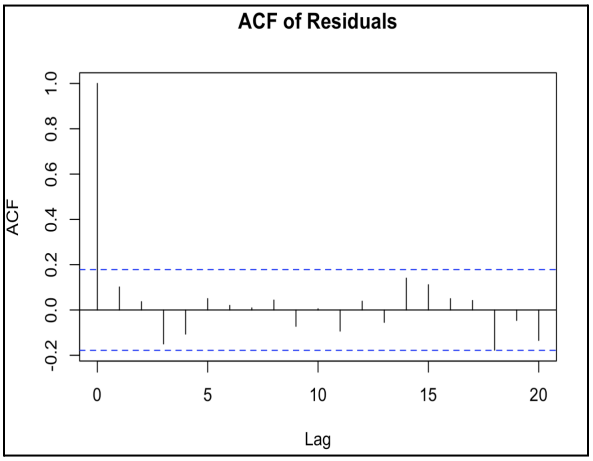


Fig7



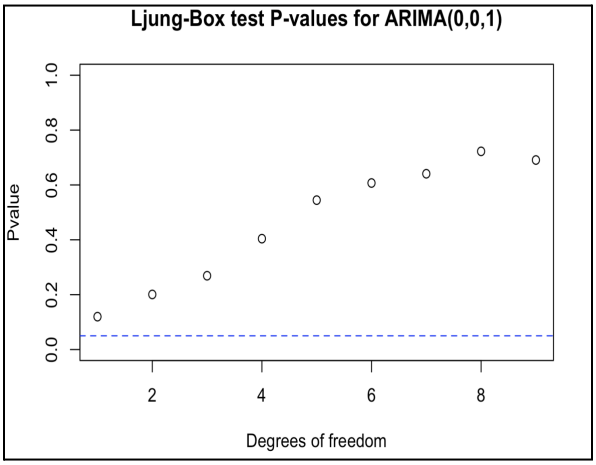


Fig8 **Fig9**

The following observations can be made based on the analysis of the residuals from the fitted ARIMA model. The time plot of the residuals resembles the pattern of white noise, indicating that the residuals do not have any significant patterns or trends. The sample autocorrelation function (ACF) of the residuals is approximately zero for lags greater than zero, suggesting that the residuals are independent and do not have significant autocorrelation. Looking at the plot for the Ljung-Box test we see that all the p-values are greater than **0.5** (represented by the blue dashed line) which suggests that the MA(1) model provides a good fit. The log-likelihood of the model is **-109.87**, and the Akaike Information Criterion (AIC) is **225.74**. These results show that the model appropriately captures the patterns in the data.

After fitting an MA(1) for further exploration we check whether the addition of further parameters would improve the model fit. We start by adding an AR component and fitting an ARMA(1,1) model to the differenced data the model output is given below:

```
arima(x = cet_temp_diff, order = c(1, 0, 1), method = "ML")

Coefficients:
         ar1      ma1  intercept
      0.1309  -0.9246     0.0126
s.e.  0.1027   0.0470     0.0053

sigma^2 estimated as 0.3501:  log likelihood = -109.05,  aic = 226.09
```

From the model output, we see that the AIC (**226.09**) for the ARMA(1,1) is slightly greater than the AIC (**225.74**) for the MA(1) model which may suggest that the MA(1) model is a better fit. Additionally, the test statistic for evaluating the null hypothesis (H0: $\varphi = 0$) against the alternative

hypothesis (H1: $\varphi \neq 0$), which assesses the significance of the autoregressive (AR) term in the model, is calculated as the ratio of the estimated AR coefficient to its standard error.
The resulting value of **0.1309/0.1027 = 1.274** is less than the critical value of **1.96** for a **95%** confidence level. Since the test statistic falls below this threshold, we fail to reject the null hypothesis at the 5% significance level. Therefore, the analysis suggests that the AR term should not be included in the model.

To explore a little further we check if the addition of another MA term may improve the model fit and hence we fit an MA(2) model and the model output is given below:

```
arima(x = cet_temp_diff, order = c(0, 0, 2), method = "ML")

Coefficients:
          ma1      ma2  intercept
      -0.8086  -0.1005     0.0127
s.e.   0.0843   0.0837     0.0054

sigma^2 estimated as 0.3508:  log likelihood = -109.16,  aic = 226.32
```

We can see that the AIC (**226.32**) for the MA(2) model is greater than the AIC (**225.74**) for the MA(1) model and the test statistic for evaluating the null hypothesis (H0: $\varphi = 0$) against the alternative hypothesis (H1: $\varphi \neq 0$) has the resulting value of **-0.1005/0.0837 = -1.200** that is less than the critical value of **-1.96** for a **95%** confidence level. Since the test statistic falls below this threshold, we fail to reject the null hypothesis at the 5% significance level. Therefore we would choose the MA(1) model as our final model for the given dataset.

CONCLUSION:
The annual mean temperature time series for the Midlands region of England exhibits non-stationarity, which was addressed by differencing the data by one season. After evaluating various ARIMA models, the Moving Average model of order 1, MA(1), was selected as the most appropriate model for this dataset. The diagnostic checks on the residuals from the fitted MA(1) model suggest that the model suitably captures the patterns and dependencies in the original time series data. The residuals resemble white noise and do not exhibit significant autocorrelation, supporting the assumption of independent and identically distributed residuals. Furthermore, the information criteria values indicate that the MA(1) model fits the data well. Attempts to include additional autoregressive or moving average terms did not significantly improve the model fit, as higher AIC values showed. Therefore, we have identified a suitable time series model after careful analysis and model fitting. The final fitted model takes the form of a moving average (MA) model of order 1 and the equation of the model can be written as "**$X_t = Z_t + \theta Z_{t-1}$**" where,

- 'Xt' is the time series variable at time t.
- 'θ' is the parameter representing the coefficient of the lagged error term.
- 'Zt' is white noise with mean zero and constant variance.

APPENDIX:

```
# Loading the dataset
cet_temp<- read.csv("cet_temp.csv")
head(cet_temp)

# Converting the dataset to a time series
cet_temp_ts <- ts(cet_temp$avg_annual_temp_C, start = 1900, frequency = 1)

# Summary statistics of the time series
summary(cet_temp_ts)

# Plotting the time series
plot(cet_temp_ts, main = "Midlands Region Annual Mean Temperature (1900-2021)",
    xlab = "Year", ylab = "Temperature (°C)", col = "black")

# ACF and PACF plots
acf(cet_temp_ts, main = "ACF Plot")
pacf(cet_temp_ts, main = "PACF Plot")

# 1st order differencing.
cet_temp_diff = diff(cet_temp_ts)

# Plot the time series for the differenced data
plot(cet_temp_diff, main = "Differenced Time PLot",
    xlab = "Year", ylab = "Temperature (°C)", col = "black")

# ACF and PACF plots of the differenced data
acf(cet_temp_diff, main = "ACF of Differenced Data")
pacf(cet_temp_diff, main = "PACF of Differenced Data")


# Fitting an MA(1) model
MA1_model <- arima(cet_temp_diff, order = c(0,0,1),method = "ML")
MA1_model

# Plotting a time plot of the residuals
plot(MA1_model$residuals, main = "MA(1) Model",
    xlab = "Time", ylab = "Residuals")
abline(h = 0, col = "red")

# ACF plot of the residuals
```

```
acf(MA1_model$residuals, main = "ACF of Residuals")

# LB Test function
LB_test<-function(resid,max.k,p,q){
 lb_result<-list()
 df<-list()
 p_value<-list()
 for(i in (p+q+1):max.k){
  lb_result[[i]]<-Box.test(resid,lag=i,type=c("Ljung-Box"),fitdf=(p+q))
  df[[i]]<-lb_result[[i]]$parameter
  p_value[[i]]<-lb_result[[i]]$p.value
 }
 df<-as.vector(unlist(df))
 p_value<-as.vector(unlist(p_value))
 test_output<-data.frame(df,p_value)
 names(test_output)<-c("deg_freedom","LB_p_value")
 return(test_output)
}

# Performing the Ljung-Box test
MA1_model.LB<-LB_test(MA1_model$residuals,max.k=11,p=0,q=1)

MA1_model.LB

plot(MA1_model.LB$deg_freedom,MA1_model.LB$LB_p_value,xlab="Degrees of
freedom",ylab="Pvalue",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

# Fitting an ARMA(1,1) model for further exploration
arma_model<- arima(cet_temp_diff, order = c(1,0,1),method = "ML")
arma_model

# Fitting an MA(2) model for further exploration
MA2_model <- arima(cet_temp_diff, order = c(0,0,2),method = "ML")
MA2_model

# Conclusion: The MA(1) model turns out to be the best fit for the "cet_temp_ts" dataset.
```

TIME SERIES FORECASTING OF EAST MIDLANDS HOUSE PRICES: A STATISTICAL APPROACH

EXECUTIVE SUMMARY:

This report presents a detailed analysis of monthly average house prices in the East Midlands region of the United Kingdom from January 2010 to December 2019. The dataset used, named "em_house_prices.csv," contains monthly average sale prices represented in British pounds (GBP). The primary objectives of this analysis were to explore the dataset, identify trends, seasonality, and autocorrelation, select an appropriate time series model, and generate forecasts for the first half of 2020.
Key findings from the analysis include:

- **Data Exploration:** Summary statistics revealed a gradual increase in average house prices over the ten years, starting at £136,102 in 2010 and peaking at £195,345 in 2019. Seasonal patterns and potential trends were observed from time series plots.
- **Model Identification and Fitting:** Initial non-stationarity in the time series data required differencing to achieve stationarity. Based on autocorrelation and partial autocorrelation functions, a SARIMA(1,1,1)(0,1,1) model was selected initially to account for non-seasonal autoregressive components and both seasonal and non-seasonal moving average components.
- **Model Diagnostics:** Residual analysis and diagnostic plots were used to refine the model, ultimately selecting SARIMA(1,1,2)(0,1,1) as the best fit. The model was validated using hypothesis testing and Ljung-Box tests, which indicated that residual autocorrelation was adequately addressed.
- **Forecasting:** The chosen SARIMA model was employed to forecast average house prices for the first six months of 2020. Point forecasts suggested a continued upward trend, with the highest forecasted price of £197,933.9 for June 2020. Prediction intervals were provided to gauge the level of uncertainty in the forecasts.

Conclusion: The SARIMA(1,1,2)(0,1,1) model effectively captured the underlying patterns in the dataset and produced reliable forecasts and the equation for the model was **"$y'_t = y'_{t-1} + \phi_1(y'_{t-1} - y'_{t-2}) + (Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}) + \Theta_1 \Theta_{t-12}$"**. Overall, the analysis successfully modelled and forecasted the house prices, providing a foundation for further analysis and decision-making.

INTRODUCTION:

This report aims to analyze monthly average house prices in the East Midlands region of the UK from January 2010 to December 2019 from the given dataset “em_house_prices.csv”. This report will apply statistical techniques and time series analysis methods to model the data and generate forecasts for the monthly average house prices in the East Midlands region for the first six months of 2020. The report will outline data exploration, methodology, model selection, and forecasting results.

DATA DESCRIPTION:

The dataset provided by the local government agency contains monthly average house sale prices in the East Midlands region of the United Kingdom, spanning from January 2010 to December 2019. The data file is named "em_house_prices.csv," with three columns: "month," "year," and "average_price_gbp." The "month" column represents the month of the year, ranging from January to December, while the "year" column indicates the corresponding year, ranging from 2010 to 2019. The "average_price_gbp" column contains each month's mean house sale price, in British pounds (GBP).

We can explore the summary statistics and visualize the time series plot to gain an initial understanding of the data. The summary statistics for the average house prices are as follows:

Summary Statistics	year	average_price_gbp
Min	2010	136102
1st Qu	2012	139900
Median	2014	154469
Mean	2014	160248
3rd Qu	2017	180844
Max	2019	195345

The minimum average price observed in the dataset is £136,102, which occurred in February 2012, while the maximum average price is £195,345, recorded in August 2019. The median average price over the entire period is £172,479, and the mean is £166,048.

Now we will visualise the time series data by producing a time plot to check for stationarity, seasonality and trends, an ACF plot which shows the correlation between observations in a time series at different lags and a PACF plot that shows the correlation between observations in a time series while removing the effects of intermediate lags.

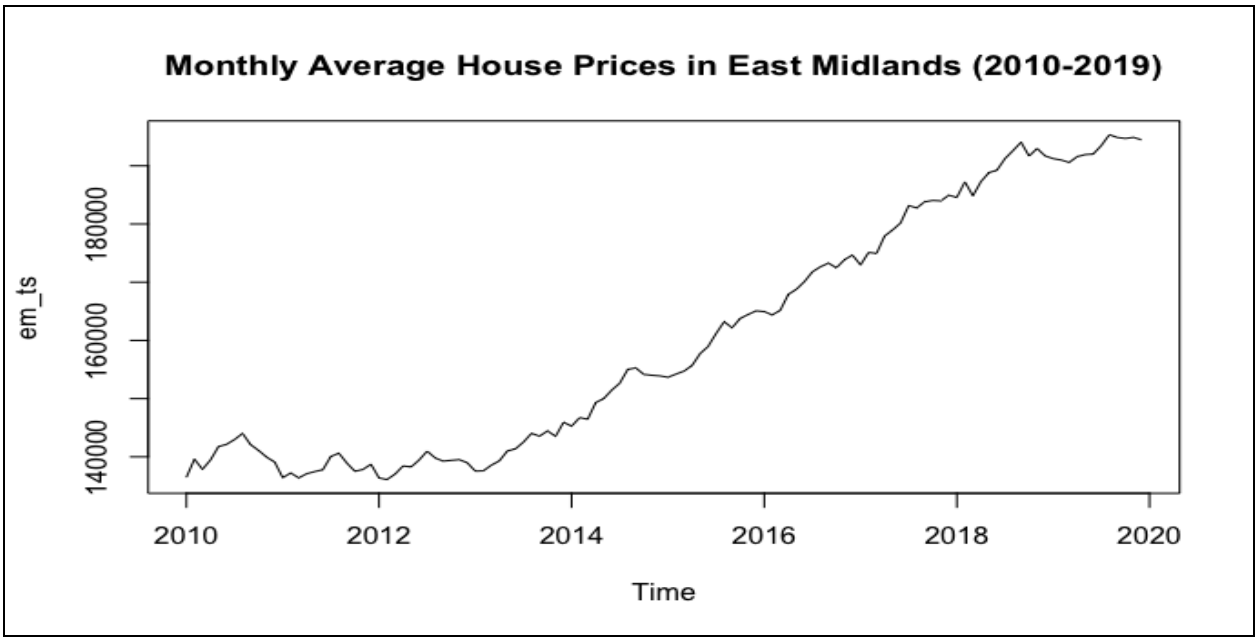


Fig1

We can observe an overall upward trend in the average house prices over the 10 years. The prices started at around 140,000 GBP in 2010 and gradually increased, reaching around 195,000 GBP by the end of 2019. Additionally, the plot suggests the potential presence of seasonal patterns or cyclical behaviour, with regular peaks and troughs visible throughout the time series, indicating that house prices may exhibit some level of seasonality within each year.

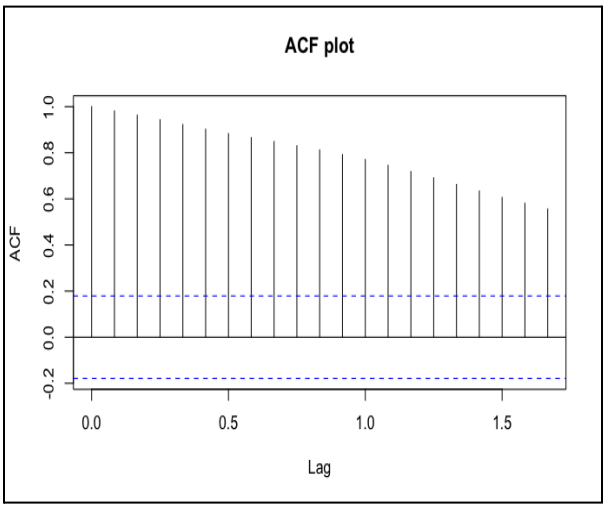


Fig2

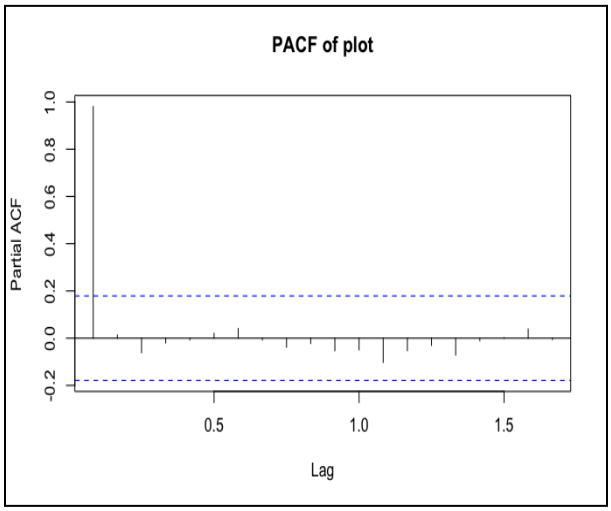


Fig3

The ACF plot shows that the autocorrelation coefficients start with a high positive value at lag 0 (always 1) and then gradually decline. This suggests the presence of autocorrelation in the data up to a certain lag, which indicates potential non-stationarity and the need for differencing to make the series stationary.

In the PACF plot, we can see that the PACFs have a significant spike at lag 1, and then they quickly drop. This suggests the potential presence of an autoregressive (AR) component in the time series model.

MODEL IDENTIFICATION:

Based on the ACF and PACF plots, it is evident that the time series is not stationary. To address this issue, we will use differencing to make the series stationary. This involves transforming the data by subtracting the previous value from each current value to stabilize variance and achieve stationarity. Specifically, the ndiffs() and nsdiffs() function in R are used to determine the required orders of non-seasonal and seasonal differencing, respectively. Based on the R output, we need to perform first-order differencing for both non-seasonal and seasonal components of the data.

After applying the appropriate differencing transformations, the time plot, ACF, and PACF of the differenced series are visualised using the plot(), acf() and pacf() functions in R. This will help assess if stationarity has been achieved and guide the identification of suitable ARIMA models that can capture the remaining patterns in the data.

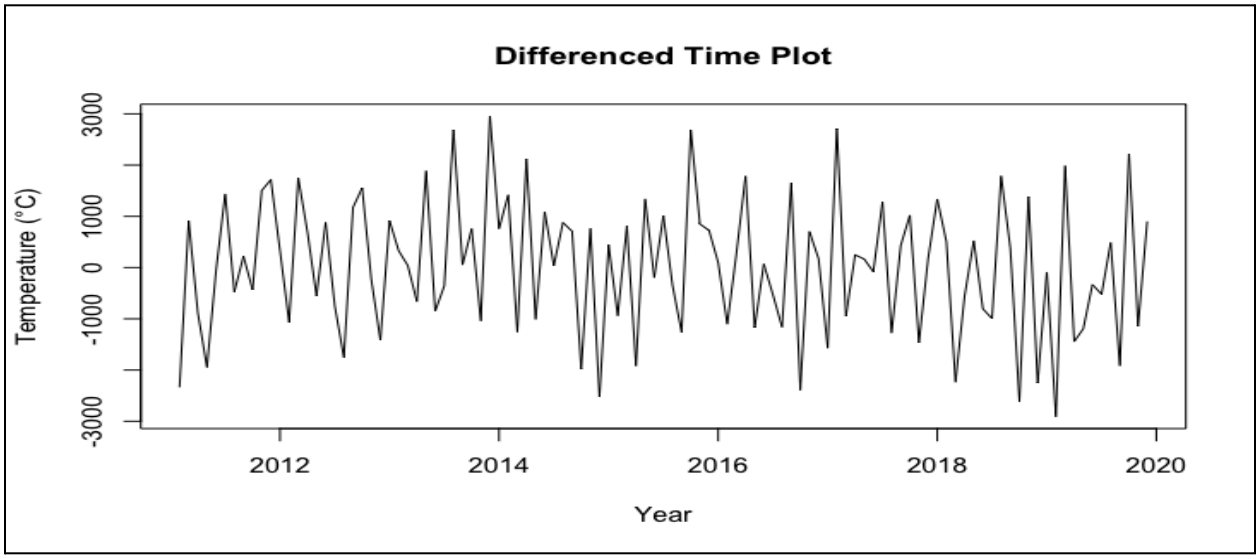


Fig4

After differencing, the time series now fluctuates around a constant mean level, indicating that the trend and seasonality components have been removed. The differenced series oscillates between positive and negative values, suggesting that the remaining patterns are likely stationary. There are still some noticeable peaks and troughs in the differenced series, which could indicate the presence of short-term autocorrelation or moving average components that need to be accounted for in the ARIMA model. However, the overall pattern of the differenced series is relatively stable, with no evident long-term trends or seasonality remaining.

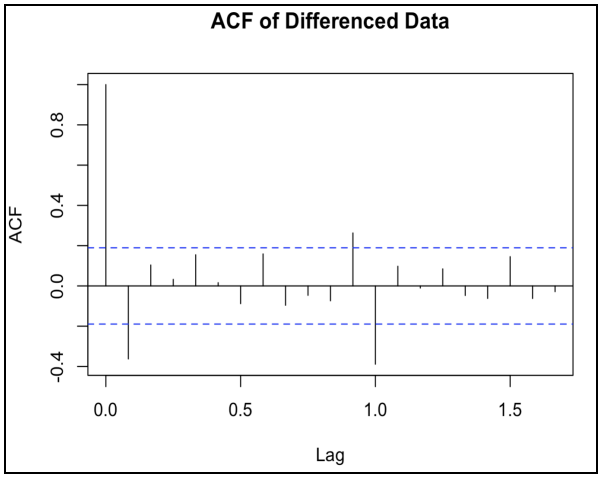


Fig5

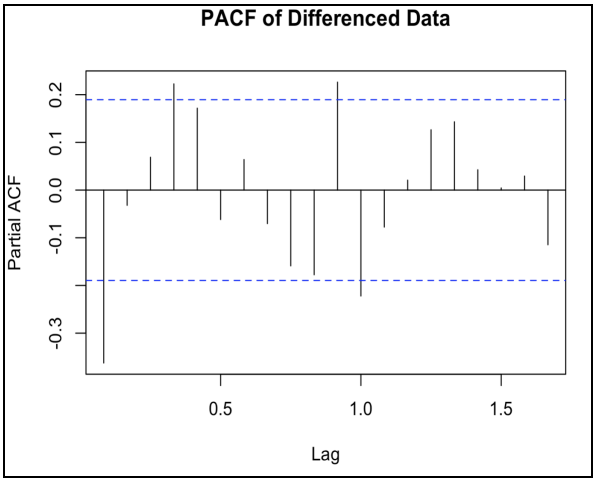


Fig6

From the above ACF and PACF plots, we see that the spike at lag 1 in the ACF suggests a non-seasonal moving average (MA) component of order 1 and the spike at lag 12 in the ACF suggests a seasonal MA(1) component. Also, the PACF has a spike at lag 1 indicating that the differenced data may follow an autoregressive (AR) process of order 1. Therefore, we begin with a SARIMA(1,1,1)(0,1,1) model.

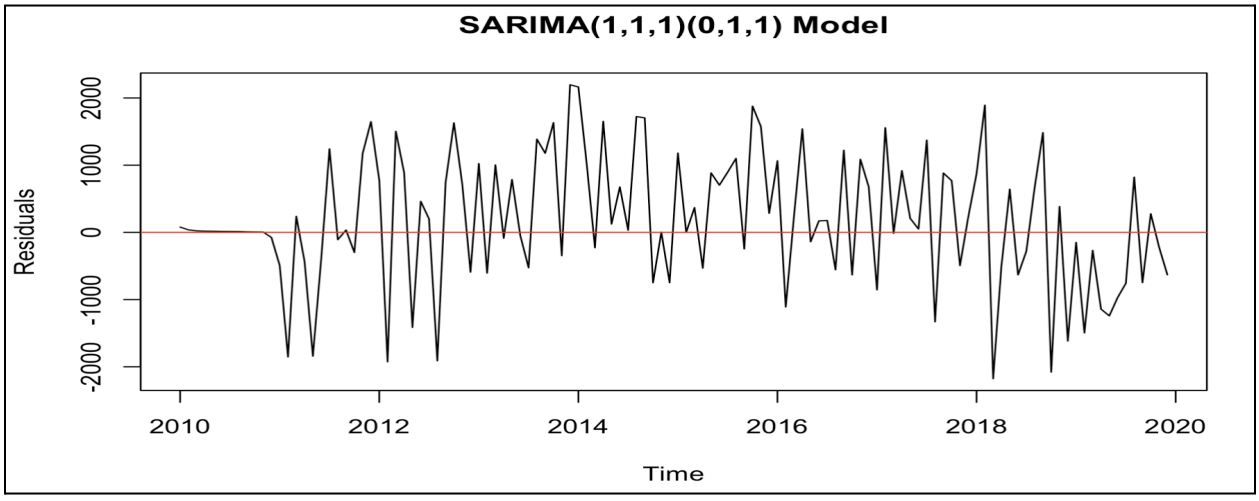
MODEL FITTING AND DIAGNOSTICS:

Based on the analysis provided, a Seasonal Auto-Regressive Integrated Moving Average (SARIMA(p,d,q)(P,D,Q)) model is being fitted, where the model parameters are specified as follows:

- p = 1 (non-seasonal autoregressive component)
- P = 1 (seasonal autoregressive component)
- q = 1 (non-seasonal moving average component)
- Q = 0 (seasonal moving average component)
- d = 1 (non-seasonal differencing)
- D = 1 (seasonal differencing)

After fitting the model we plot a time plot, ACF and PACF plot of the residuals to check if the model fits the data appropriately.

Fig7



From the above time plot, we see that the residuals fluctuate around zero, showing some seasonality with peaks and troughs occurring at regular intervals. There are some larger positive and negative spikes, suggesting events that the model may not have captured adequately. The presence of a seasonal pattern and occasional large spikes suggests that the model may not have fully accounted for all the patterns in the data.

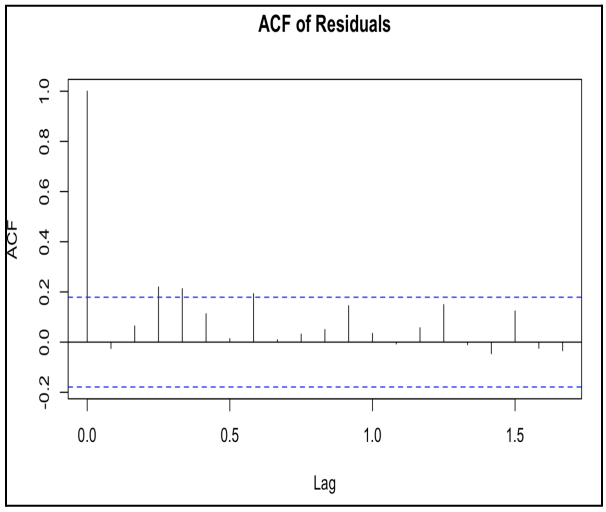


Fig8

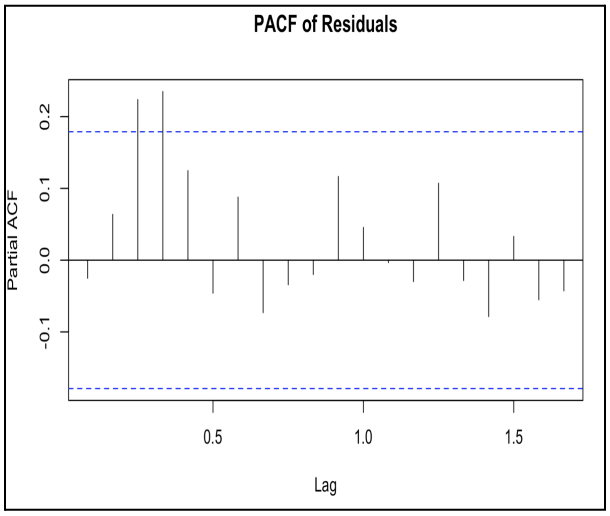


Fig9

Looking at the ACF and PACF plots of the residuals we can see that both the plots have spikes at lags 3 and 4 indicating that some additional non-seasonal terms need to be included in the model.

Considering the outcomes shown in the plot above we can add an additional MA component and check if the model has adequately captured the relevant patterns in the data, and the residuals can be considered random and uncorrelated. Now we fit a SARIMA(1,1,2)(0,1,1) model and produce a time plot, ACF and PACF plots of the residuals.

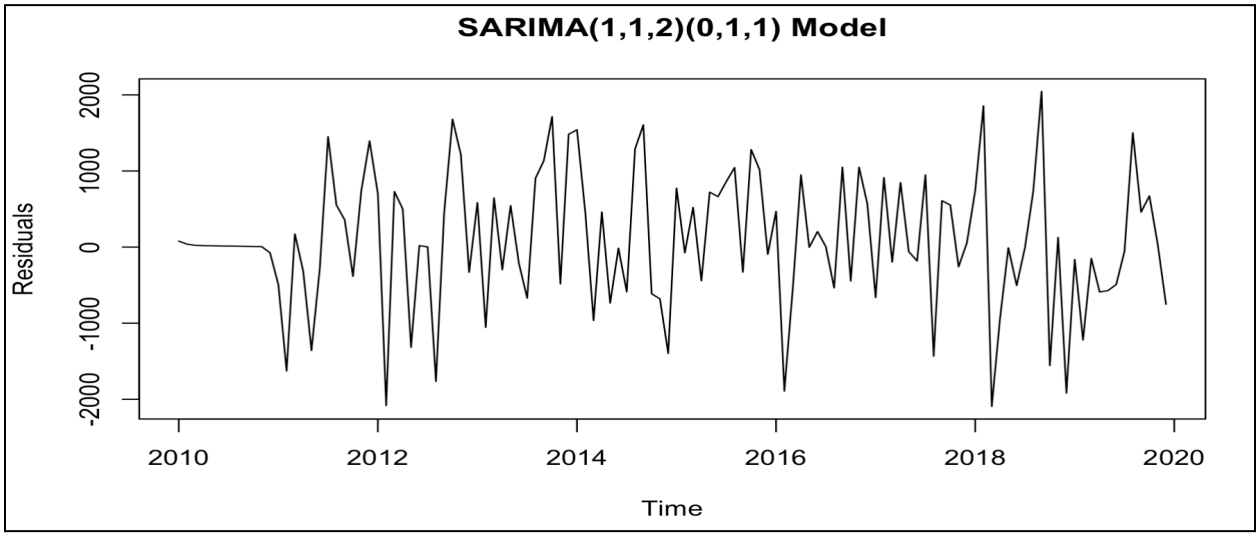


Fig10

In the above time plot of the SARIMA(1,1,2)(0,1,1) model compared to the previous SARIMA(1,1,1)(0,1,1) model, the residuals in this plot seem to have fewer large positive or negative spikes, suggesting an improvement in capturing events. The plot resembles the pattern of white noise, indicating that the residuals do not have any significant patterns or trends.

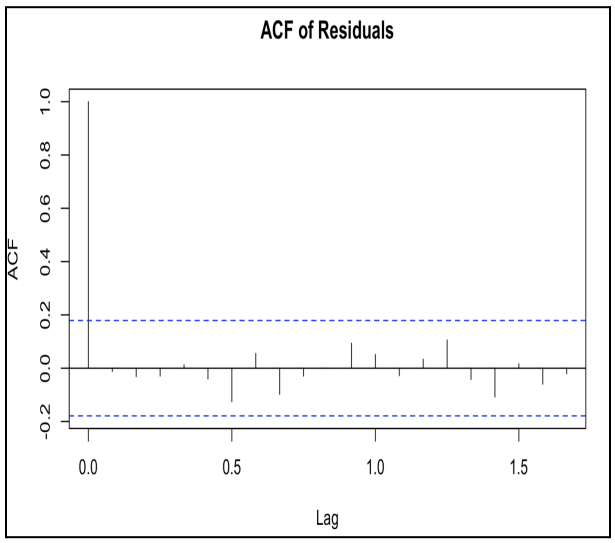


Fig11

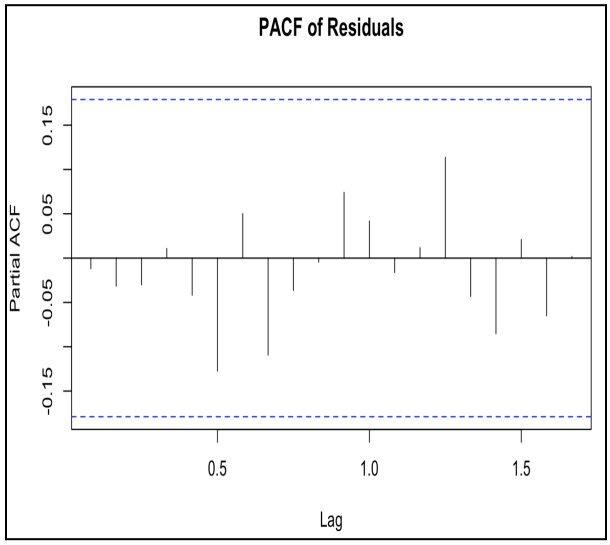
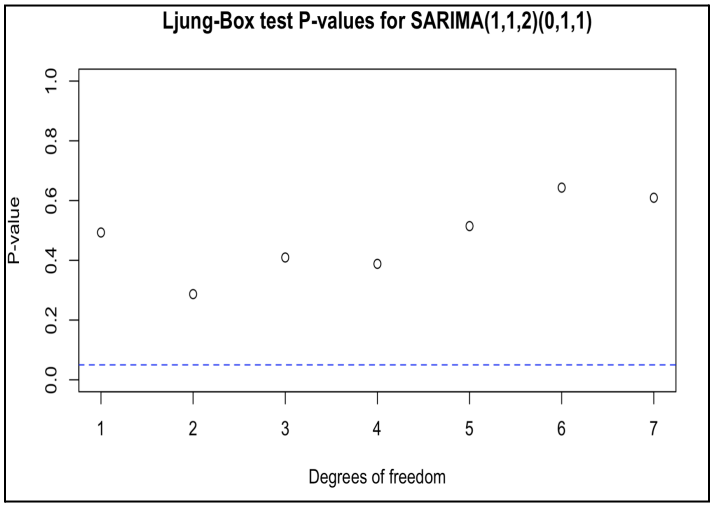


Fig12

The following observations can be made from the sample ACF and PACF plots. In the ACF plot, the ACF values for all lags fall within the confidence intervals represented by the dashed lines. This indicates that there is no significant autocorrelation present in the residuals at any lag. From the PACF plot, we can see that there are no significant spikes at any lags, which indicates that the residuals from the fitted SARIMA model are white noise.

By incorporating an additional non-seasonal MA(2) term in this SARIMA(1,1,2)(0,1,1) model, it appears that the model has successfully accounted for any remaining autocorrelation patterns that may have been present in the residuals of the previous simpler model. Now we can perform the Ljung-Box test of p-values and plot the p-values to gain insights into the adequacy of our model.

Fig(13)



Looking at the above plot we see that for all lags (from 1 to 7), the p-values are relatively high (ranging from approximately 0.29 to 0.64) which are greater than 0.05 (represented by the dashed line). These higher p-values suggest that no significant autocorrelation is detected in the residuals of our SARIMA model at the tested lags. The model output is given below:

```
Series: em_ts
ARIMA(1,1,2)(0,1,1)[12]

Coefficients:
          ar1      ma1     ma2      sma1
      0.855  -1.2235  0.5234  -0.8108
s.e.  0.093   0.0996  0.0901   0.1337

sigma^2 = 911391:  log likelihood = -890.44
AIC=1790.89   AICc=1791.48   BIC=1804.25
```

We can also check the model's AIC, AICc, and BIC values to ensure that the model is a good fit for the data. The Akaike Information Criterion (AIC) is a measure of the relative quality of a statistical model for given data, The Corrected Akaike Information Criterion (AICc) is a modification of the AIC that adjusts for small samples and The Bayesian Information Criterion (BIC) is another criterion for model selection that penalizes model complexity more heavily than AIC. The SARIMA(1,1,2)(0,1,1)[12] model has the AIC (**1790.89**), AICc (**1791.48**), and BIC (**1804.25**) values.

After fitting the SARIMA (1,1,2)(0,1,1)[12] model for further exploration we check whether the addition of further parameters would improve the model fit by checking the AIC, AICc, and BIC values. We can start by adding a seasonal MA component and fitting a SARMA(1,1,2)(0,1,2)[12] model the model output is given below:

```
Series: em_ts
ARIMA(1,1,2)(0,1,2)[12]

Coefficients:
          ar1      ma1     ma2      sma1     sma2
      0.8425  -1.2146  0.5271  -0.7950  -0.0642
s.e.  0.0993   0.1033  0.0902   0.1693   0.1435

sigma^2 = 900979:  log likelihood = -890.34
AIC=1792.68   AICc=1793.52   BIC=1808.72
```

From the model output, we see that the AIC (**1792.68**), AICc (**1793.52**) and BIC (**1808.72**) for the SARMA(1,1,2)(0,1,2)[12] are slightly greater than the AIC (**225.74**), AICc (**1791.48**), and BIC (**1804.25**) for the SARIMA(1,1,2)(0,1,1)[12] model which may suggest that the SARIMA(1,1,2)(0,1,1)[12] model is a better fit.

Additionally, the test statistic for evaluating the null hypothesis (H0: $\varphi = 0$) against the alternative hypothesis (H1: $\varphi \neq 0$), which assesses the significance of the Moving Average (MA) term in the model, is calculated as the ratio of the estimated MA coefficient to its standard error. The resulting value of **-0.064/0.1435 = -0.4459** is less than the critical value of **1.96** for a **95%** confidence level. Since the test statistic falls below this threshold, we fail to reject the null hypothesis at the 5% significance level. Therefore, the analysis suggests that the model should not include the MA term.

To explore a little further we check if the addition of another MA and AR term may improve the model fit and hence we fit a SARIMA(1,1,3)(0,1,1)[12] and a SARIMA(2,1,2)(0,1,1)[12] models and the model's output are given below:

```
Series: em_ts
ARIMA(1,1,3)(0,1,1)[12]

Coefficients:
          ar1      ma1     ma2     ma3     sma1
      0.8511  -1.2179  0.5172  0.0059  -0.8092
s.e.  0.1282   0.1635  0.1689  0.1329   0.1351

sigma^2 = 920870:  log likelihood = -890.44
AIC=1792.89   AICc=1793.73   BIC=1808.92
```

```
Series: em_ts
ARIMA(2,1,2)(0,1,1)[12]

Coefficients:
          ar1      ar2      ma1     ma2      sma1
      0.8602  -0.0065  -1.2275  0.5285  -0.8096
s.e.  0.1797   0.1976   0.1514  0.1763   0.1349

sigma^2 = 920773:  log likelihood = -890.44
AIC=1792.89   AICc=1793.73   BIC=1808.92
```

The test statistic used to evaluate the significance of the Moving Average (MA) and Auto-Regressive (AR) terms in the model resulted in values of **0.044** and **-0.032**, respectively, which are both less than the critical value of **1.96** for a **95%** confidence level. This means that we cannot reject the null hypothesis (H0: $\varphi = 0$) at the **5%** significance level. Therefore, based on this analysis, it is suggested that we would not include the additional MA and AR terms in the model.

As a final measure, we can also compare the information criteria values (AIC, AICc, and BIC) along with the log-likelihood and sigma^2 (residual variance) values. Lower values of AIC, AICc, and BIC are generally preferred. Additionally, higher log-likelihood and lower sigma^2 values are preferable. All of the values of the models that have been analysed are given below:

ARIMA(1,1,2)(0,1,1)[12]:

- AIC = 1790.89
- AICc = 1791.48
- BIC = 1804.25
- Log-likelihood = -890.44
- Sigma^2 = 911391

ARIMA(1,1,2)(0,1,2)[12]:

- AIC = 1792.68
- AICc = 1793.52
- BIC = 1808.72
- Log-likelihood = -890.34
- Sigma^2 = 900979

ARIMA(1,1,3)(0,1,1)[12]:

- AIC = 1792.89
- AICc = 1793.73
- BIC = 1808.92
- Log-likelihood = -890.44
- Sigma^2 = 920870

ARIMA(2,1,2)(0,1,1)[12]:

- AIC = 1792.89
- AICc = 1793.73
- BIC = 1808.92
- Log-likelihood = -890.44
- Sigma^2 = 920773

Based on the above comparison the ARIMA(1,1,2)(0,1,1)[12] is the best model for the given time series data. Now we can use the chosen model to perform forecasting.

FORECASTING:

After fitting the ARIMA(1,1,2)(0,1,1)[12] model to the East Midlands monthly house price data, the model is used to forecast the average house prices for the first six months of 2020. The forecasted values and a plot of those values, along with the associated 80% and 95% prediction intervals, are shown below:

Month	Point Forecast	80% Prediction Interval	95% Prediction Interval
January 2020	£193,930.5	£192,702.6 - £195,158.4	£192,052.6 - £195,808.4
February 2020	£194,836.4	£193,384.6 - £196,288.3	£192,616.0 - £197,056.8
March 2020	£194,401.0	£192,620.8 - £196,181.3	£191,678.3 - £197,123.7
April 2020	£196,204.1	£194,029.6 - £198,378.6	£192,878.5 - £199,529.7
May 2020	£197,202.1	£194,596.7 - £199,807.6	£193,217.5 - £201,186.8
June 2020	£197,933.9	£194,879.2 - £200,988.7	£193,262.1 - £202,605.7

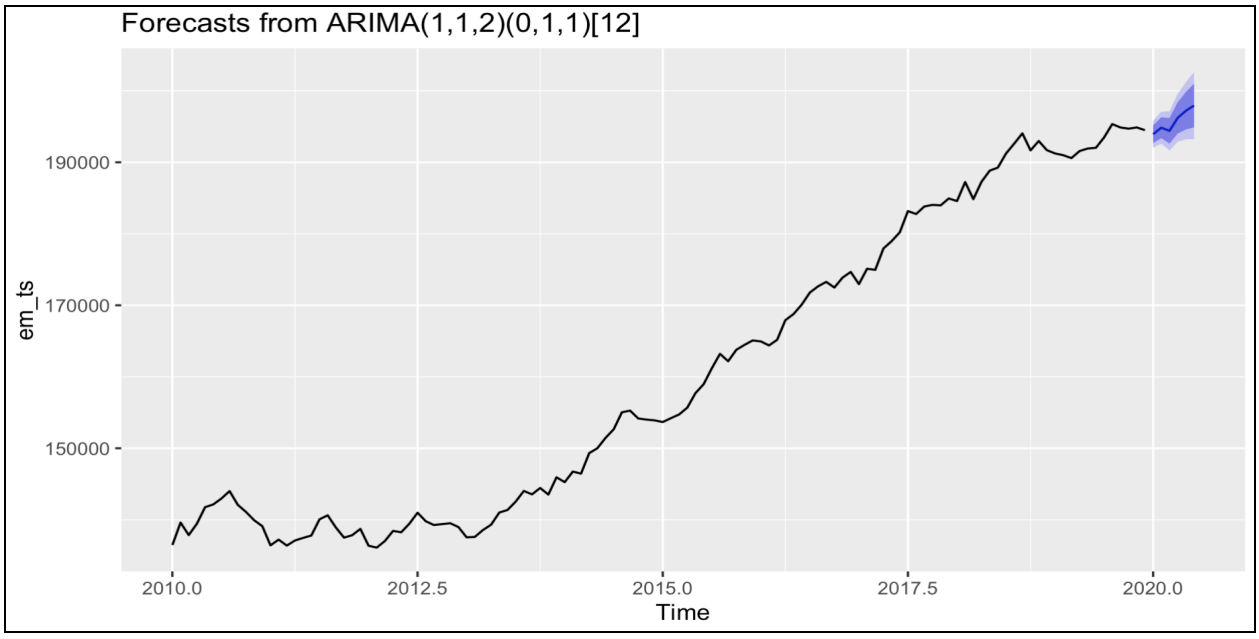


Fig14

The forecasted values suggest a continued upward trend in the average house prices for the East Midlands region during the first half of 2020. The point forecasts show a gradual increase in prices, with the highest forecasted price of £197,933.9 for June 2020.

The associated prediction intervals provide a range of possible values within which future house prices are expected to fall with a certain level of confidence. We can see that the 80% prediction intervals are narrower than the 95% intervals, which suggests a higher level of uncertainty in the latter case.

From the forecasting plot, we can see that the forecasted values and the prediction intervals, along with the historical data used for model fitting. The upward trend of the forecasted values matches the general pattern seen in the given data.

CONCLUSION:

The report conducted a comprehensive time series analysis of the monthly average house prices in the East Midlands region from 2010 to 2019. After detailed model diagnostics, the SARIMA(1,1,2)(0,1,1)[12] model was chosen as the best fit, accounting for non-stationarity, seasonality, autoregressive, and moving average components. Using this model, forecasts were generated for the first six months of 2020, indicating a continued upward trend with the highest forecasted price of £197,933.9 in June. The associated prediction intervals provide a range of possible values within which future house prices are expected to fall, offering insights into the uncertainty surrounding the forecasts. Overall, the analysis successfully modelled and forecasted the house prices, providing a foundation for further analysis and decision-making. The equation for the SARIMA $(1, 1, 2)(0, 1, 1)_{12}$ model can be written as:

$$\text{“}y'_t = y'_{t-1} + \phi_1(y'_{t-1} - y'_{t-2}) + (Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}) + \Theta_1 \Theta_{t-12}\text{”}$$

Where:

- y'_t is the differenced and seasonally differenced series.
- ϕ_1 is the non-seasonal autoregressive parameter.
- θ_1 and θ_2 are the non-seasonal moving average parameters.
- Θ_1 is the seasonal moving average parameter.
- Z_t is the white noise error term at time t.

APPENDIX: R code

```
# Importing the forecast library
library(forecast)

# Reading the dataset
em_house_prices <- read.csv("em_house_prices.csv")

# Statistical summary of the data
summary(em_house_prices)

# Converting to time series data
em_ts <- ts(em_house_prices$average_price_gbp, start = c(2010, 1), frequency = 12)

# Exploratory Data Analysis (EDA)
# Plotting a time plot
plot(em_ts, main = "Monthly Average House Prices in East Midlands (2010-2019)")

# Checking ACF and PACF plots to determine ARIMA orders
acf(em_ts, main = "ACF plot")
pacf(em_ts, main = "PACF of plot")

# Checking the number of non-seasonal and seasonal differences required to make
# the time series stationary
ndiffs(em_ts) # Non-seasonal
nsdiffs(em_ts) # Seasonal

# Performing both non-seasonal and seasonal differencing simultaneously
em_ts_diff <- diff(diff(em_ts, differences = 1), lag = 12)

# Plotting the time plot, ACF and PACF plot of the series after differencing
plot(em_ts_diff, main = "Differenced Time Plot",
    xlab = "Year", ylab = "Temperature (°C)", col = "black")
acf(em_ts_diff, main = "ACF of Differenced Data")
pacf(em_ts_diff, main = "PACF of Differenced Data")

# The spike at lag 1 in the ACF suggests a non-seasonal moving average (MA) of order 1 component
#and the spike at lag 12 in the ACF suggests a seasonal MA(1) component.
# Also the PACF has a spike at lag 1 indicating that the differenced data series may follow an
autoregressive (AR) process of order 1
# Hence, we begin with a SARIMA(1,1,1)(0,1,1) model,
# indicating a first and non-seasonal difference of AR(1) components, and non-seasonal and
seasonal MA(1) components.
```

```
sarima <- Arima(em_ts,order = c(1, 1, 1), seasonal = c(0, 1, 1))
# Plotting a time plot of the residuals
plot(sarima$residuals, main = "SARIMA(1,1,1)(0,1,1) Model",
    xlab = "Time", ylab = "Residuals")
abline(h = 0, col = "red")

# ACF plot of the residuals
acf(sarima$residuals, main = "ACF of Residuals")
pacf(sarima$residuals, main = "PACF of Residuals")

# Both the ACF and PACF show spikes at lag 3 and 4,
# indicating that some additional non-seasonal terms need to be included in the model.

sarima_2 <- Arima(em_ts,order = c(1, 1, 2), seasonal = c(0, 1, 1))
# Plotting a time plot of the residuals
plot(sarima_2$residuals, main = "SARIMA(1,1,2)(0,1,1) Model",
    xlab = "Time", ylab = "Residuals")
abline(h = 0, col = "red")

# ACF and PACF plot of the residuals
acf(sarima_2$residuals, main = "ACF of Residuals")
pacf(sarima_2$residuals, main = "PACF of Residuals")
sarima_2

# Performing the Ljung-Box test for p-values
# LB Test function for SARIMA model
LB_test_sarima <- function(resid, max.k, order, seasonal_order) {
  p <- order[1]  # Non-seasonal AR order
  d <- order[2]  # Non-seasonal differencing
  q <- order[3]  # Non-seasonal MA order
  P <- seasonal_order[1]  # Seasonal AR order
  D <- seasonal_order[2]  # Seasonal differencing
  Q <- seasonal_order[3]  # Seasonal MA order
  m <- seasonal_order[4]  # Seasonal period

  lb_result <- list()
  df <- list()
  p_value <- list()

  for (i in (p + q + P + Q + 1):max.k) {
    lb_result[[i]] <- Box.test(resid, lag = i, type = "Ljung-Box", fitdf = (p + q + P + Q))
    df[[i]] <- lb_result[[i]]$parameter
```

```
   p_value[[i]] <- lb_result[[i]]$p.value
  }

  df <- as.vector(unlist(df))
  p_value <- as.vector(unlist(p_value))

  test_output <- data.frame(deg_freedom = df, LB_p_value = p_value)
  return(test_output)
}

# Specifying SARIMA model parameters
order <- c(1, 1, 2)          # Non-seasonal order (p, d, q)
seasonal_order <- c(0, 1, 1, 12)  # Seasonal order (P, D, Q, m)

# Residuals from SARIMA(1,1,2)(0,1,1) model
sarima_residuals <- residuals(sarima_2)

# Performing the Ljung-Box test for the SARIMA model
sarima_LB <- LB_test_sarima(sarima_residuals, max.k = 11, order = order, seasonal_order =
seasonal_order)

# Displaying Ljung-Box test results
print(sarima_LB)

# Plotting the Ljung-Box test results
plot(sarima_LB$deg_freedom, sarima_LB$LB_p_value, xlab = "Degrees of freedom", ylab = "P-value",
    main = "Ljung-Box test P-values for SARIMA(1,1,2)(0,1,1)", ylim = c(0, 1))
abline(h = 0.05, col = "blue", lty = 2)  # Adds a horizontal line at significance level 0.05

# Further exploring if the addition of other seasonal and non-seasonal components
# would improve the model fit.
fit1 <- Arima(em_ts,order = c(1, 1, 2),seasonal = c(0, 1, 2))
fit1

fit2 <- Arima(em_ts,order = c(1, 1, 3),seasonal = c(0, 1, 1))
fit2

fit3 <- Arima(em_ts,order = c(2, 1, 2),seasonal = c(0, 1, 1))
fit3

# Since the test statistic falls below the threshold (0.5), we fail to reject
# the null hypothesis at the 5% significance level. Therefore, the analysis
# suggests that all the additional terms should not be included in the model.
```

```
# Hence we choose the ARIMA(1,1,2)(0,1,1) model as our final model for forecasting.

# Forecast for the first six months of 2020.
forecast_values <- forecast(fit, h = 6)
forecast_values

# Plotting the forecasted values.
sarima_2 %>%
  forecast(h = 6) %>%
  autoplot()
```