MS4H04
Time series and forecasting
Lecturer: Dr Angelica Pachon

Student name: Zainab Khan
Student ID: 23028535

Table of Contents

1. Introduction:

A time series is a series of observations collected over a time period usually within equal intervals. When collected at set intervals it is known as a discrete time data series, otherwise if the data is collected continuously it is known as a continuous time series. Time series analysis looks to gain meaningful characteristics and insightful statistics from a time series. Forecasting can be carried out on a time series based on historic observed values. In this report an ARIMA model (Auto-Regressive Integrated Moving Average) model is used as a forecasting algorithm to predict future values.

There are three main components in a time series:

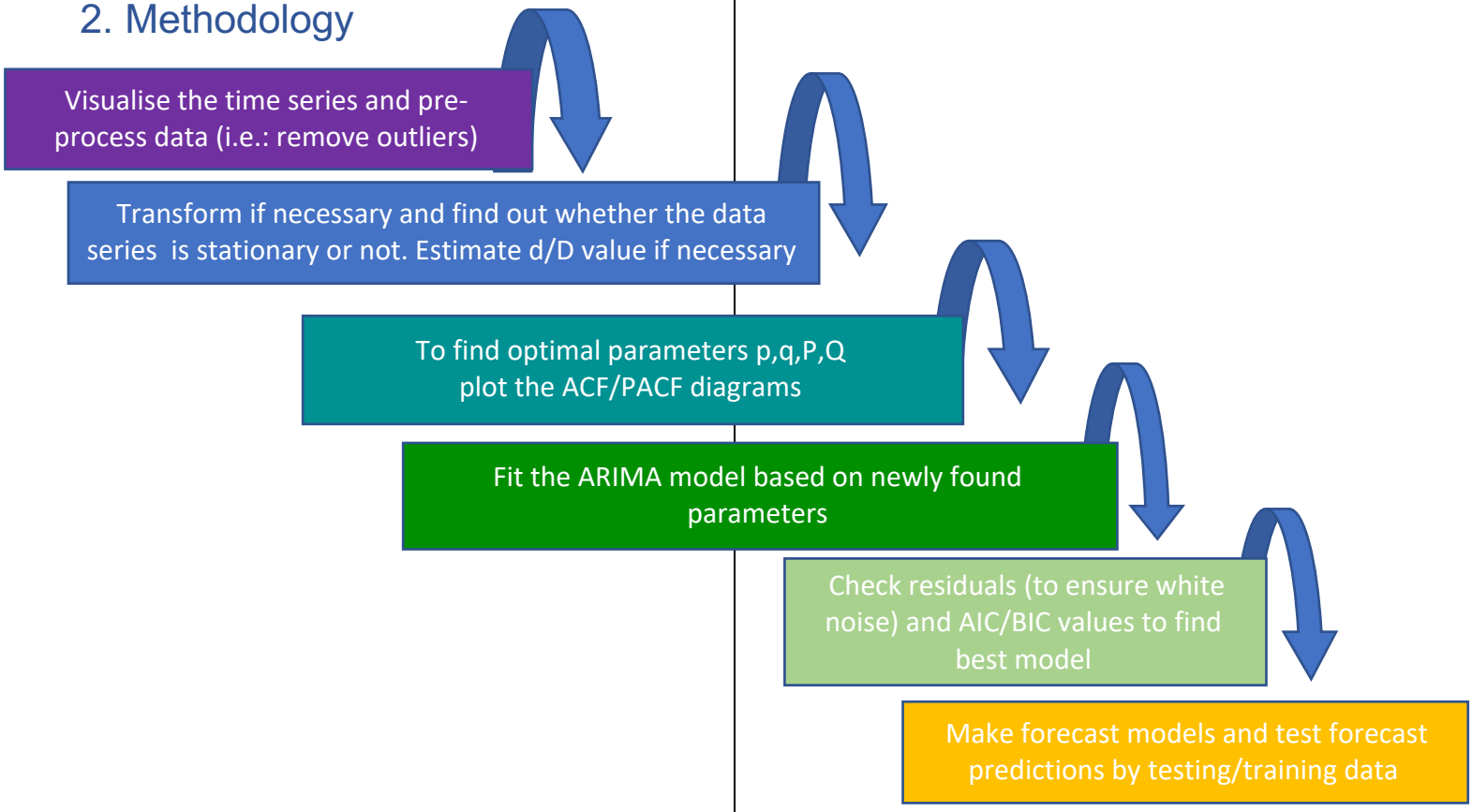
1. **Trend:** Occurs when there is a long-term increase or decrease observed in the data points.
2. **Seasonal:** an observed pattern that is determined by fixed seasonal factors at a known frequency.
3. **Cyclic:** an observed cycle of a rise or fall in data points that aren't of fixed frequency. There are generally associated with economic conditions and occur in a business setting.

The contents of the report will include fitting two time series with a SARIMA and a ARIMA model in R. The SARIMA model will be fitted based on data from ONS regarding the amount of UK passengers travelling abroad whereas the ARIMA data will be fitted on data based on the actual number of hours work in the UK (full time) and is also taken from the ONS time series archive. Both time series will be of different frequencies; monthly and quarterly respectively.

ARIMA models are categorised by three parameters:

1. **'p'** : the order of the Autoregressive term (AR); an AR model is used in forecasting and relates the historic values of the time series variable to the time series itself.
2. **'d'** : the number of times differencing is carried out in order to achieve a stationary time series. This accounts for the integration element (the 'I') in ARIMA
3. **'q'** : the order of the Moving Average term (MA) which looks into forecasting the errors in a model that resembles a linear regression model.

2. Methodology



3. SARIMA model:

Objective: To predict the amount of UK passenger visits' abroad for the next 2 years.

Dataset information: This dataset was sourced from the Office of National Statistics (ONS) and contains information about the number of UK visit's abroad and spans from 1986 January to 2020 March. The data shows numbers in thousands. In order to forecast future visits abroad a SARIMA model will be used, the past values data will be fed into the model.

Assumptions: The data must be stationary in order to use it for modelling, this entails some criteria that must be met. A data series containing a trend or seasonality must be transformed for it to be stationary. Over a given time period these criteria below must be met:

- There should be a constant mean (μ) over time t.
- There should be a constant variance (σ) over time t.
- There autocovariance function between two observed values (X_{t_1} and X_{t_2}) should only depend on the time interval $t_1 and\ t_2$.

3.1. Exploratory data Analysis (EDA):

After the data series is converted to a time series, the start as well as the end of the time series needs to be defined alongside the frequency of the data series. The data for the UK visits' abroad is collected on a monthly basis therefore the frequency is set to be 12. The cycle function is used to provide the position in the cycle of each observation.

```
ts1 <- ts(data['passenger'], start = c(2010,3), end = c(2020,3),frequency = 12)
class(ts1)
start(ts1)
end(ts1)
frequency(ts1)
summary(ts1)
cycle(ts1)
```

```
[1] "ts"
[1] 2010    3
[1] 2020    3
[1] 12
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2010                 3   4   5   6   7   8   9  10  11  12
2011   1   2   3   4   5   6   7   8   9  10  11  12
2012   1   2   3   4   5   6   7   8   9  10  11  12
2013   1   2   3   4   5   6   7   8   9  10  11  12
```

Figure 1: Defining time series and cycle

In figure 2, seasonality is obvious as well as a very slight increase in trend. It can also be observed that there is an increase in variance over time and that there is an outlier in 2020 probably due to Covid.

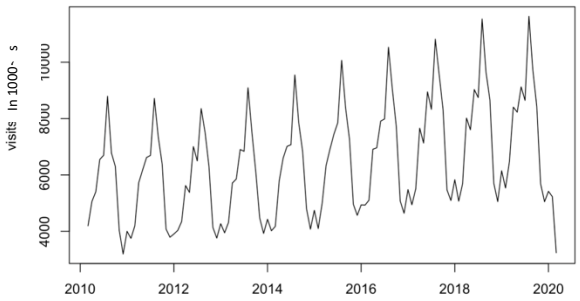


Figure 2: Time series plot

Due to outliers effecting the analysis and forecasting, it is best to remove/replace it. Because it is at the end of the trend it can be removed assuming that such things won't happen again. In a business setting, this shouldn't be ignored and considered if forecasting in the longer term but to make it simple to work with the time series end dates will be changed to just before covid affecting flights.

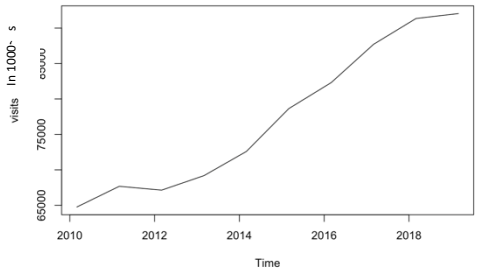


Figure 3 Monthly aggregated plot

When the data is aggregated on a monthly basis, a clear upward trend can be seen (see figure 3)

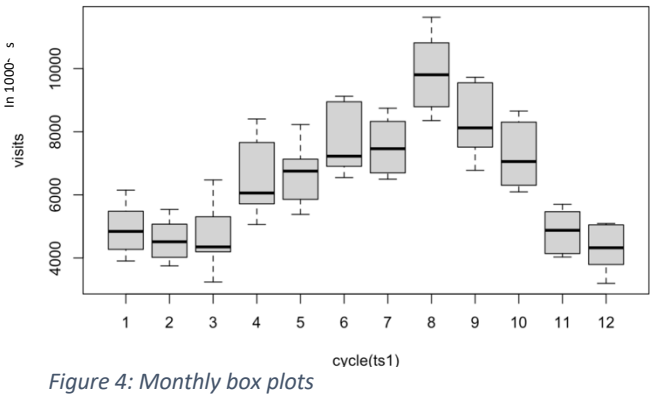


Figure 4: Monthly box plots

The boxplots in figure 4 show that generally the data behaves differently per month suggesting seasonality. For example, in August is most likely when there will be the greatest amounts of visits abroad.

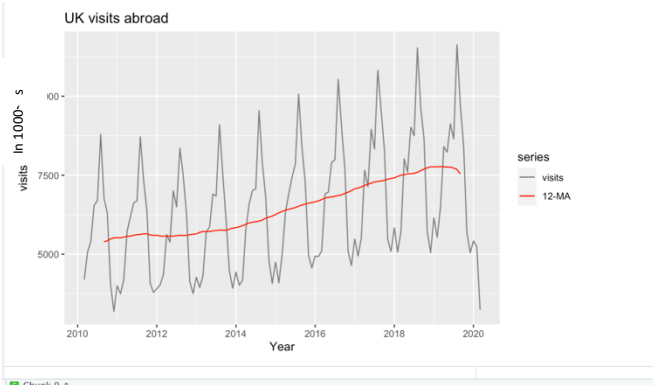


Figure 5: 12-month average plot

The moving average plot also supports the idea of trend in figure 5. A general increase in values can be observed.

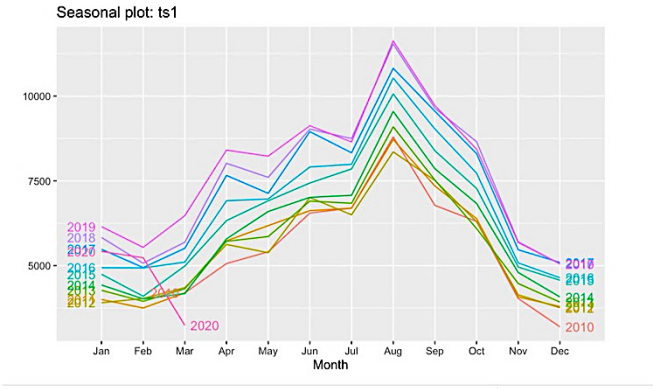


Figure 6 Seasonal plot of time series

The seasonal plot in figure 6 support the idea that there is seasonality present in the series. There is a spike every August in the seasonal plot and every year show peaks in similar trends for each month.

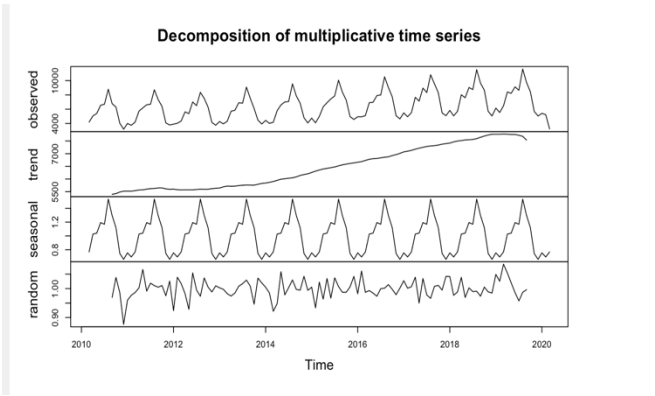


Figure 7 Decomposition of time series

The multiplicative decomposition in figure 7 and the random element of the decomposition looks to be uncorrelated with anything, but this will be investigated as below otherwise it shows trend and seasonality in data series. The multiplicative decomposition is used to look at due the series being non-linear and increasing over time.

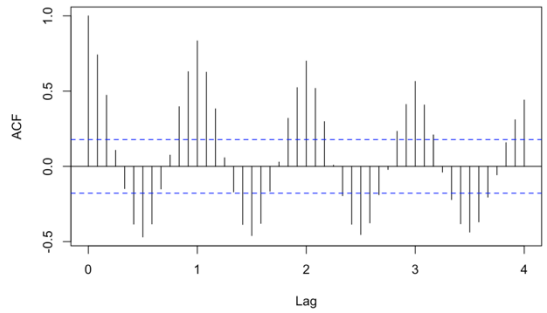


Figure 8 ACF plot

The ACF plot in figure 8 suggests that the series is non stationary due to the fact that the decrease in data values towards 0 is slow. As the values of the ACF in figure 8 slowly decrease as the lags increase, it suggests trend. Also the sinosoidal shape suggests seasonality due to the correlations being larger for the seasonal lags, at multiples of the seasonal frequency, than for other lags.

```
Warning in adf.test(ts1) : p-value smaller than printed p-value

        Augmented Dickey-Fuller Test

data:  ts1
Dickey-Fuller = -8.3963, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary


        KPSS Test for Level Stationarity

data:  ts1
KPSS Level = 0.64201, Truncation lag parameter = 4, p-value = 0.01882
```

Figure 9 Statistical tests for stationarity

Augmented Dicky Fuller test was carried out (see figure 9) to test whether the time series data is stationary by checking if there is a unit root present in the data.

-The null hypothesis (H_0) is that there is a unit root in the time series suggesting a non-stationary time series which consists of some time dependant structure.

-The alternative hypothesis (H_1) is that there is no unit root present in the time series data and the data is stationary.

Figure 9 suggests that we can reject the null the hypothesis that the time series is not stationary due to p value being less than the confidence interval of 0.05 (alpha value), i.e: the time series is showing stationary characteristics which include no autocorrelation between data points. This is questionable as there are some outliers and these tests can be quite sensitive of that.

Therefore another statistical test which is the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) was carrid out.

-The null hypothesis (H_0) is that there is no unit root in the time series suggesting a stationary time series which consists of some time dependant structure.

-The alternative hypothesis (H_1) is that there is no unit root present in the time series data and the data is stationary.

As the P value is less than the alpha value of 0.05, it suggest that the null hypothesis of the time series being stationary should be rejected. This would align with the the assumption that the data is not stationary. Due to the difference in results, it could be suggested that the series is trend stationary and not strict stationary.

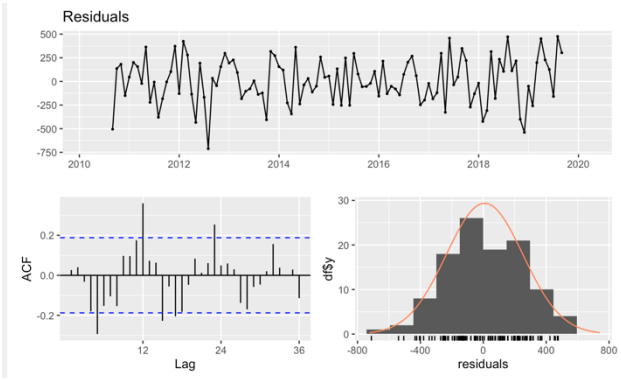


Figure 10 Residual analysis

The residuals were also analysed. The histogram in figure 10 should ideally show a normally distributed plot, the plot is somewhat near a symmetrical normally distributed plot, but this may be affected by outliers. The autocorrelation function plot shows that most the random data meet the 95% interval of confidence that the data is uncorrelated.

```
        Augmented Dickey-Fuller Test

data:  random1
Dickey-Fuller = -4.0825, Lag order = 3, p-value = 0.01746
alternative hypothesis: stationary
```

Figure 11 Dicky Fuller test on residuals

The Augmented Dicky Fuller test was carried (see figure 11) to test whether the random data is stationary by checking if there is a unit root present in the data. The null hypothesis (H_0) is that there is a unit root in the residual element of the time series suggesting a non-stationary data which consists of some time dependant structure. The alternative hypothesis (H_1) is that there is no unit root present in the risidual element of the time series data. Figure 11 suggests that we can reject the null the hypothesis that the random data is not stationary due to p value being less than the confidence interval of 0.05, i.e: the random data is showing stationary characteristics which include no autocorrelation between data points.

3.2 Model Fitting

3.2.1 Pre-processing

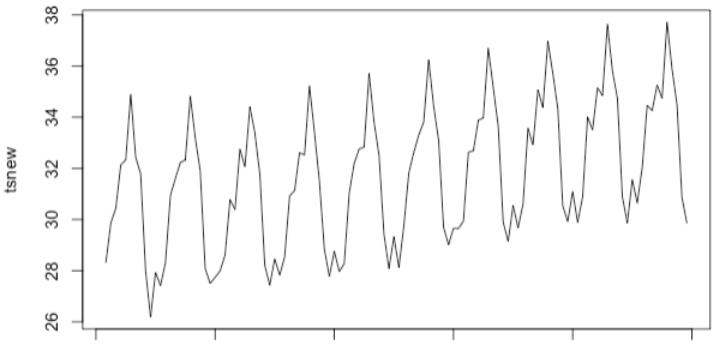


Figure 12 Transformed data to reduce variance

Due to the data having increased variance with time, a BoxCox transformation will be applied to meet the assumptions (see figure 12, compare with figure 2). Not only this but the time series will be shortened by three months to exclude the period of outliers (due to Covid), due to the outliers being at the end and the circumstances being unusual, this is more time efficient then replacing outlier values.

3.2.2 Removing trend and stationarity:

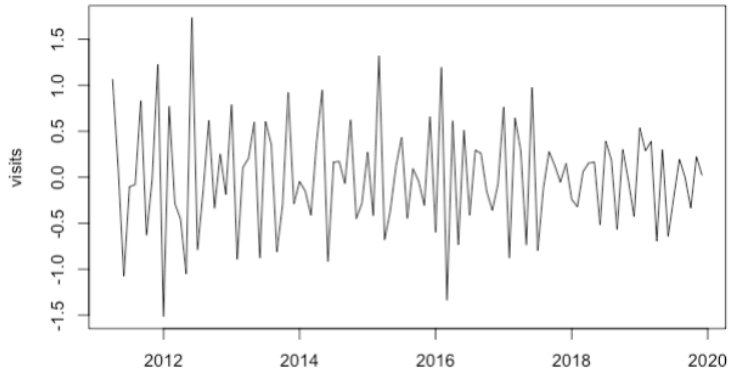


Figure 13 Plot of stationary data

In order to remove trend, the series will be differenced once and to remove seasonality the series will be differenced with the lag-d differencing operator (∇_d). As the data is monthly, the differencing operator will be 12. This will make the parameter 'd' and 'D' both 1. By observing the plot after differencing to remove the seasonality and trend element, the obvious
trend and seasonality pattern is removed, and the time series oscillates around 0.

3.2.3 Fitting the best model

The ACF and PACF plots can be used to

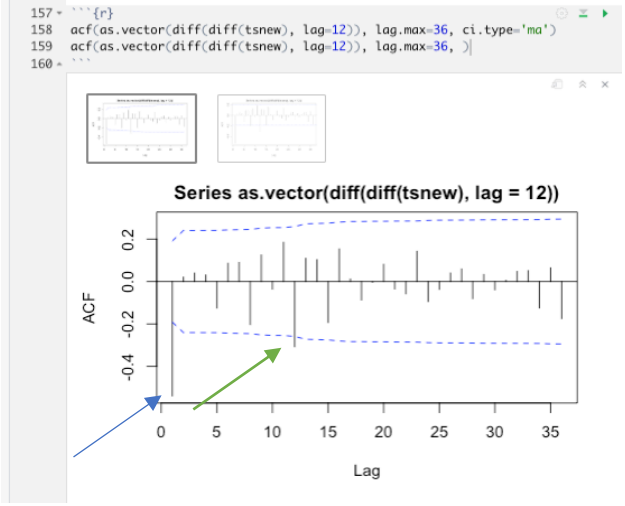


Figure 14 ACF plot

determine 'p', 'q', 'P', 'Q'. By looking look at lags which show correlation, the 4 parameters can be estimated. 'p' could be 1-3, whereas 'P' could be 1-2. 'q' could be 0-1 and 'Q' could be 1. The SARIMA model will be run with different parameters to find the optimum model

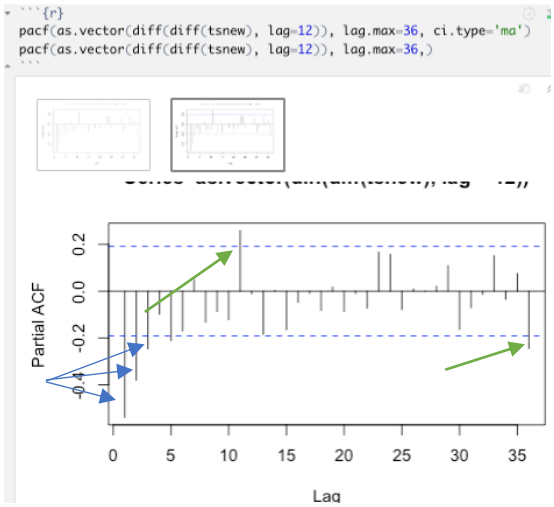


Figure 15 PACF plot

```
f1 <- arima(tsnew, order=c(0,1,1), seasonal=list(order=c(1,1,1), p=12))
f2 <- arima(tsnew, order=c(1,1,1), seasonal=list(order=c(1,1,1), p=12))
f3 <- arima(tsnew, order=c(0,1,1), seasonal=list(order=c(0,1,1), p=12))
f4 <- arima(tsnew, order=c(1,1,1), seasonal=list(order=c(0,1,1), p=12))
f5 <- arima(tsnew, order=c(0,1,2), seasonal=list(order=c(1,1,1), p=12))
f6<- arima(tsnew, order=c(1,1,2), seasonal=list(order=c(1,1,1), p=12))
f7 <- arima(tsnew, order=c(0,1,2), seasonal=list(order=c(0,1,1), p=12))
f8 <- arima(tsnew, order=c(1,1,2), seasonal=list(order=c(0,1,1), p=12))
f9 <- arima(tsnew, order=c(0,1,1), seasonal=list(order=c(1,1,2), p=12))
f10 <- arima(tsnew, order=c(1,1,1), seasonal=list(order=c(1,1,2), p=12))
f11<- arima(tsnew, order=c(1,1,1), seasonal=list(order=c(1,1,2), p=12))
f12 <- arima(tsnew, order=c(0,1,2), seasonal=list(order=c(1,1,2), p=12))
f13 <- arima(tsnew, order=c(0,1,2), seasonal=list(order=c(1,1,2), p=12))
f14 <- arima(tsnew, order=c(0,1,2), seasonal=list(order=c(0,1,2), p=12))
f15<- arima(tsnew, order=c(0,1,1), seasonal=list(order=c(0,1,2), p=12))
f16<- arima(tsnew, order=c(2,1,1), seasonal=list(order=c(0,1,1), p=12))
f17<- arima(tsnew, order=c(3,1,1), seasonal=list(order=c(0,1,1), p=12))
f18<- arima(tsnew, order=c(2,1,2), seasonal=list(order=c(0,1,2), p=12))
f19<- arima(tsnew, order=c(3,1,2), seasonal=list(order=c(0,1,2), p=12))
f20<- arima(tsnew, order=c(2,1,1), seasonal=list(order=c(1,1,1), p=12))
f21<- arima(tsnew, order=c(3,1,1), seasonal=list(order=c(1,1,1), p=12))
```

Figure 16 Various SARIMA models are fitted

Looking at the AIC values the SARIMA model, the SARIMA model (0,1,1) (0,1,1) shows the best fit (see figure 16) due to it having the lowest AIC value but the residuals must also be checked. So, for now the SARIMA model with order (0,1,1) (0,1,1) will be used.

```
[1] 104.2454
[1] 105.2749
[1] 103.8108
[1] 104.3751
[1] 105.2289
[1] 107.2283
[1] 104.3152
[1] 106.3152
[1] 106.1381
[1] 107.182
[1] 107.182
[1] 107.1365
[1] 107.1365
[1] 105.1365
[1] 104.1452
[1] 106.2585
[1] 108.0159
[1] 108.9981
[1] 110.756
[1] 107.1686
[1] 109.8672
```

Figure 17 AIC values for fitted models

3.2.4 Residual analysis

To verify whether the residuals of the model are truly white noise the residuals need to replicate the characteristics of white

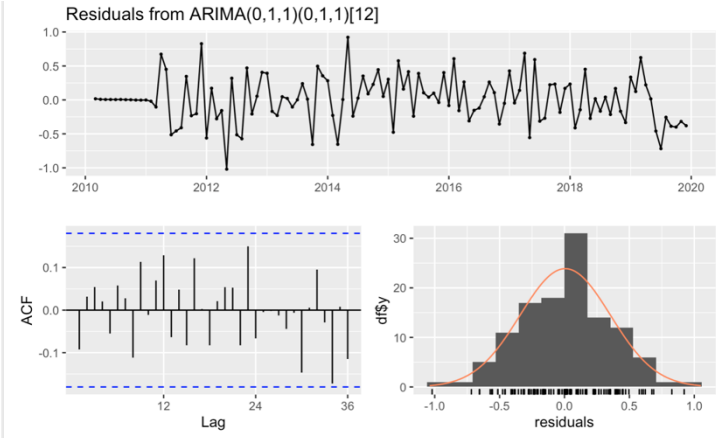


Figure 18 Residual analysis

-The residuals in figure 18 generally show that no trend or obvious pattern, it oscillates around 0 (has a mean of 0), no obvious changes in variance or observations of outliers and show no signs of autocorrelation.
-The histogram in figure 18 slightly deviates from the normal distributed bell shape however it still has some symmetry and is distributed around 0 suggesting normality.
-The ACF plot in figure 18 shows no significant autocorrelations, all the lags are within the 95% confidence interval.

```
        Ljung-Box test

data:  Residuals from ARIMA(0,1,1)(0,1,1)[12]
Q* = 19.38, df = 22, p-value = 0.6218

Model df: 2.   Total lags used: 24
```

Figure 19 Lung Box test

-The p value for the Ljung-Box test in figure 20 is greater than 0.05 alpha level suggesting that the null hypothesis of the error terms being uncorrelated, should be rejected.

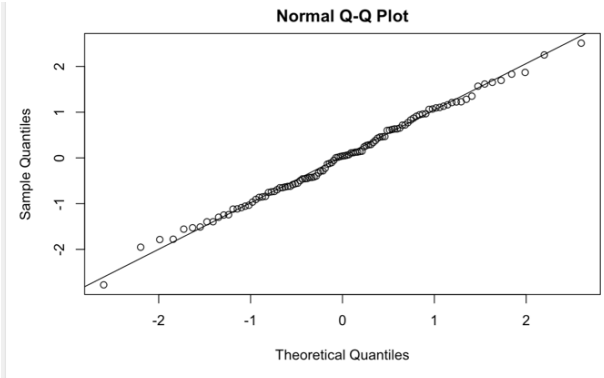


Figure 20 Q-Q plot

-The Q-Q plot in figure 20 shows the data is normally distributed due to the scatter points being roughly in a straight line, this is due to the distribution being evenly aligned with the standard normal variate.

```
        Augmented Dickey-Fuller Test

data:  residuals(f3)
Dickey-Fuller = -4.3198, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary


        Shapiro-Wilk normality test

data:  rstandard(f3)
W = 0.99484, p-value = 0.9453


        One-sample Kolmogorov-Smirnov test

data:  rstandard(f3)
D = 0.072852, p-value = 0.5582
alternative hypothesis: two-sided
```

Figure 21 Statistical tests for stationarity and normality

-The ADF p value of the residuals shown in figure 21, is less than the significant level of 0.05 therefore the null hypothesis that the data is not stationary can be rejected.

-The Shapiro-Wilk test has a p value greater than 0.05, the null hypothesis that the data is normally distributed cannot be rejected and it can be concluded that the data is normal.

-The null hypothesis of the one-sample kolmogorov-smirnov test states that there is no difference between the two distributions (normality and time series), as the p value is greater than 0.05, the null hypothesis cannot be rejected, and normality can be assumed.

All the above suggests that the residuals are truly random and resemble white noise.

3.2.5 Forecasting

Figure 22 shows the forecast of visits of UK national abroad for the next two years. The dotted lines represent the 95% prediction intervals.

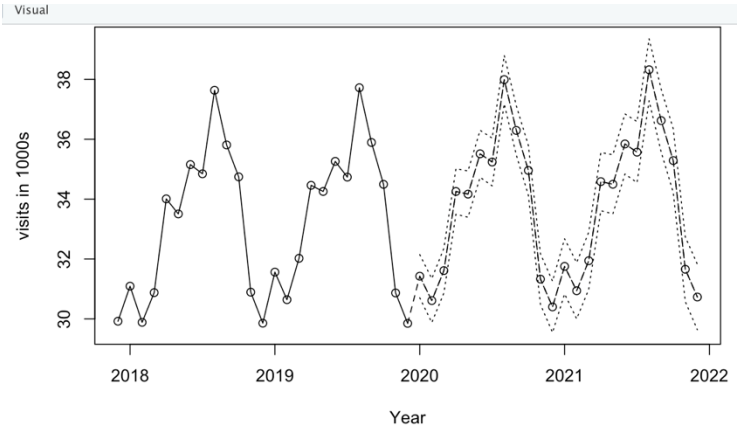


Figure 22 Forecasting for the next 24 months

3.2.6 Forecasting error

The data was split into a ratio of 8:2 for testing and training purposes. Root mean square error will be used to assess the top 3 models. The second SARIMA model with order (0,1,1) (0,1,2) had the lowest RMSE score of 0.243%. The previously selected SARIMA model (0,1,1) (0,1,1) has a higher RMSE but just around by 0.2%. This shows that the lowest AIC score doesn't equate to the best model.

```
> futurf3<- predict(f3_train,n.ahead = 12)
> accuracy(futurf3$pred, test_tsnew)
                    ME      RMSE       MAE         MPE      MAPE      ACF1 Theil's U
Test set -0.0315221 0.2621455 0.22283 -0.1145085 0.6886934 -0.413402  0.135382
>
> futurf15 <- predict(f15_train,n.ahead = 12)
> accuracy(futurf15$pred, test_tsnew)
                    ME      RMSE       MAE         MPE      MAPE     ACF1 Theil's U
Test set -0.06130788 0.2432583 0.2094902 -0.2039864 0.6503312 -0.3661 0.1272559
>
> futurf1<- predict(f1_train,n.ahead = 12)
> accuracy(futurf1$pred, test_tsnew)
                    ME      RMSE       MAE         MPE      MAPE       ACF1 Theil's U
Test set -0.05787535 0.2450449 0.2093297 -0.1936487 0.6500238 -0.3738487 0.1278678
>
```

Figure 23 Forecasting error for top 3 models

3.2. Conclusion

A SARIMA model was created and the steps to create it were outlined. Methods such as differencing and transforming the time series data were used to make the data stationary. Then the p,d,q and P,D,Q values were estimated via the ACF and PACF plots and the model was fitted

based on the estimated parameters. The residuals were checked in order to meet all the assumptions. The series was forecasted and then trained and tested. It was found an alternative model had a lower RMSE even though it had a slightly higher AIC value.
The RMSE value is quite small therefore the model could prove to be very useful in forecasting. Figure 23 shows an MAPE (mean absolute percentage error) value of 0.65%, so the model was around 99% accurate in predicting. Noting that we removed outliers, this model would not have effectively predicted for the unprecedented covid pandemic.

4. ARIMA model

Objective: To predict the number of hours worked in a week by full time workers

Dataset information: This dataset was sourced from the Office of National Statistics (ONS) and contains information about the average actual weekly hours of work for full time workers and spans from 1992 January to 2022. In order to forecast future visits abroad an ARIMA model will be used, the past values data will be fed into the model.

Assumptions: The data must be stationary in order to use it for modelling, this entails some criteria that must be met. A data series containing a trend or seasonality must be transformed for it to be stationary. Over a given time period these criteria below must be met:

- There should be a constant mean (μ) over time t.
- There should be a constant variance (σ) over time t.
- There autocovariance function between two observed values (X_{t_1} and X_{t_2}) should only depend on the time interval $t_1 and\ t_2$.

4.1. Exploratory data Analysis (EDA):

After the data series is converted to a time series, the start as well as the end of the time series needs to be defined alongside the frequency of the data series. The data is collected on a quarterly basis therefore the frequency is set to be 4. The cycle function is used to provide the position in the cycle of each observation.

```
> class(ts1)
[1] "ts"
> start(ts1)
[1] 1992    2
> end(ts1)
[1] 2022    1
> frequency(ts1)
[1] 4
```

```
> cycle(ts1)
     Qtr1 Qtr2 Qtr3 Qtr4
1992         2    3    4
1993    1    2    3    4
1994    1    2    3    4
1995    1    2    3    4
1996    1    2    3    4
1997    1    2    3    4
1998    1    2    3    4
1999    1    2    3    4
2000    1    2    3    4
2001    1    2    3    4
2002    1    2    3    4
2003    1    2    3    4
2004    1    2    3    4
```

Figure 24 Defining time series and cycle

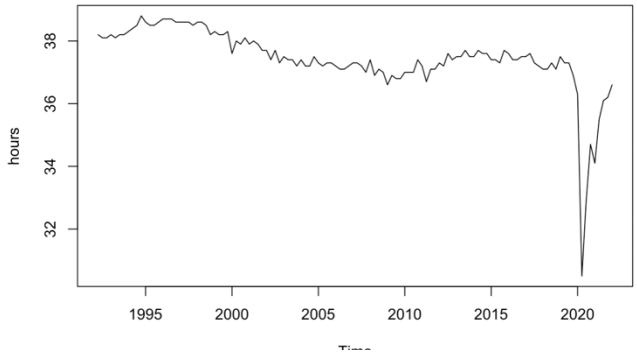


Figure 25: Plot of time series

Figure 25 Shows the plot of the time series. There seems to be a trend and an outlier occurring during the pandemic. The outlier should be dealt with otherwise it may interfere in analysis.

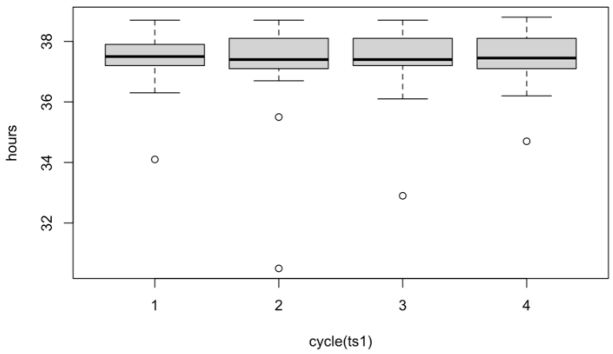


Figure 26 Quarterly box plots

Figure 26 shows the box plots for each quarter of the time series. They show similar statistical characteristics as opposed to a seasonal series boxplot; this suggests there isn't seasonality present in the series.

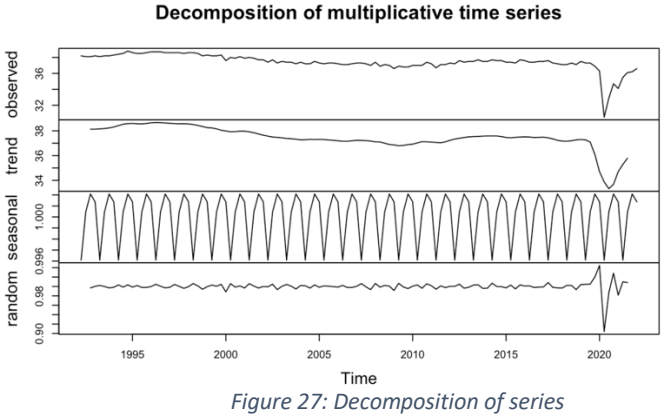


Figure 27: Decomposition of series

The decomposition shows that a trend and the random element of the decomposition looks to be white noise (except for outliers).

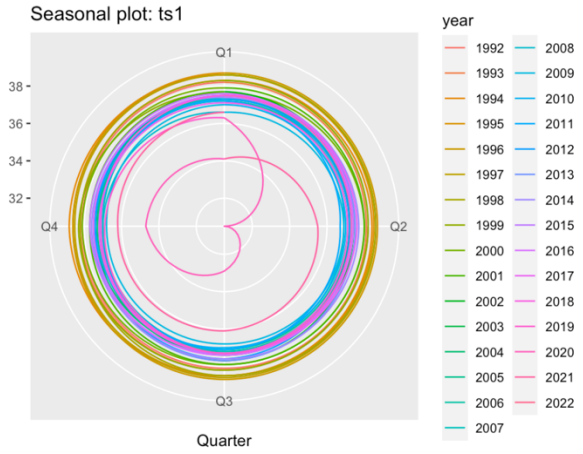


Figure 27: Decomposition of series

The polar seasonal plot in figure 27 does not show any patterns, so it supports the idea there is only trend in the series.

Figure 28: Quarterly moving average of series

The moving average plot supports the idea of a downward trend, it also showcases how much the outliers affects the moving average.

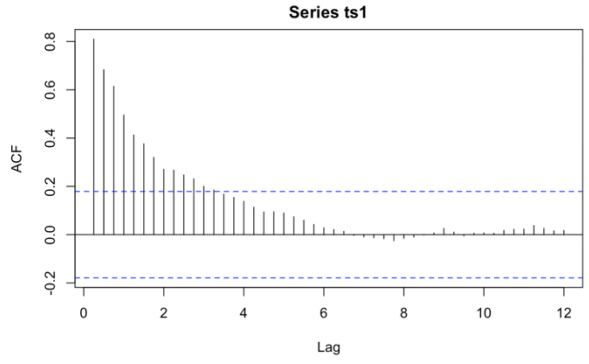


Figure 29: ACF plot

The ACF plot in figure 29 suggests that the series is non stationary due to the fact that the decrease in lag values towards 0 is slow. As the values of the ACF in figure 29 slowly decrease as the lags increase it suggests trend.

```
Warning in kpss.test(ts1) : p-value smaller than printed p-value

        KPSS Test for Level Stationarity

data:  ts1
KPSS Level = 1.2933, Truncation lag parameter = 4, p-value = 0.01


        Augmented Dickey-Fuller Test

data:  ts1
Dickey-Fuller = -3.7864, Lag order = 4, p-value = 0.02202
alternative hypothesis: stationary
```

Figure 30: assessing stationarity of series

-The ADF has a p value of less than 0.05 in figure 30, therefore the null hypothesis that the time series is not stationary is rejected, suggesting a stationary series. This result can be sensitive to outliers and may be the reason why it doesn't support the assumption of non-stationarity.

-The KPSS test has a p value of less than 0.05, therefore the null hypothesis that the series is stationary is rejected, this supports the idea that the series is not stationary. The opposing outputs may change when the outliers are addressed.

4.2 Fitting the ARIMA model

4.2.1 Pre-processing

The outliers will be detected and replaced using the 'tsclean' function available in the forecast package.

The plot still shows a great drop in the height of the pandemic which is ideal as it is more accurate and representative of the pandemic whilst also taking care of outliers. As the variance doesn't increase/decrease significantly over time no other transformations will be applied.

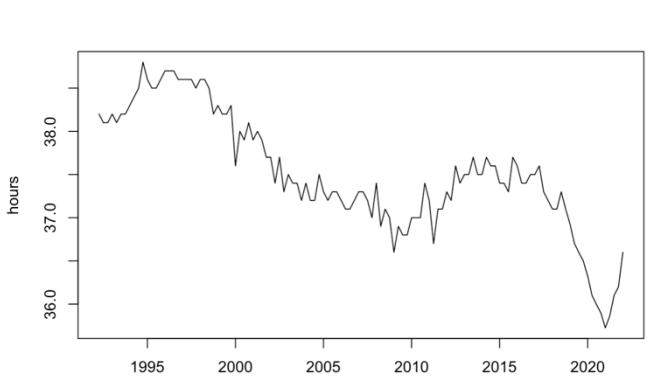


Figure 31: Outliers detected and replaced with tsclean function

4.2.2. Removing trend

To remove the trend the series will be differenced once.

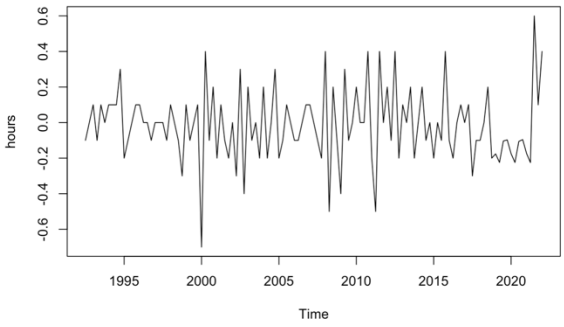


Figure 32: Series differenced to remove trend

It is noticeable that there is no more trend visible, even the outliers seem to be accounted for. The series oscillates around 0 (has a mean of 0) and shows no obvious pattern or correlation suggesting a non-stationary series.

4.2.3 Predicting parameters

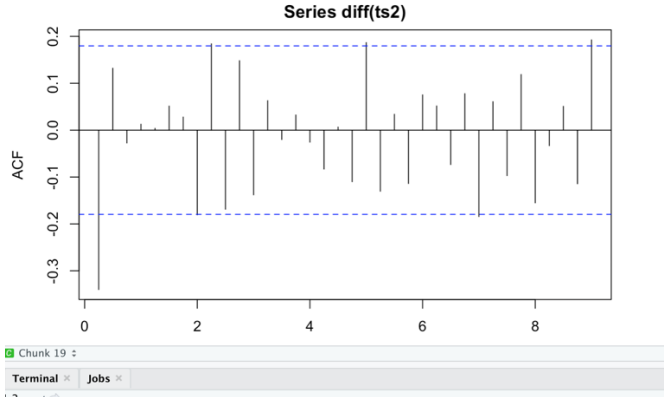


Figure 33: ACF plot

The ACF plot shows only 1 lag out of the interval of confidence to show autocorrelation suggesting that 'q' is 1.

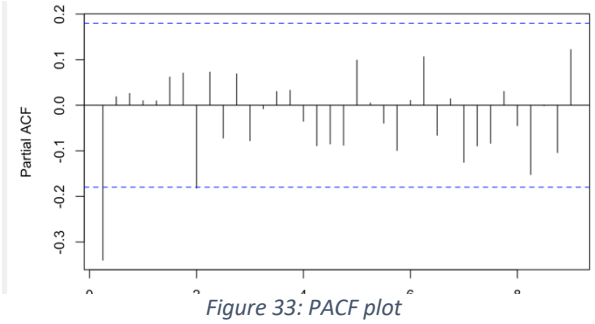


Figure 33: PACF plot

The PACF plot in figure 33 shows only 1 lag that is out of the confidence of interval so 'q' is 1.

The EACF plot in figure 34 suggests models of order (1,0).

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 o o o o o o o o o o o  o  o  o
2 x o o o o o o o o o o  o  o  o
3 x o o o o o o o o o o  o  o  o
4 x o o o o o o o o o o  o  o  o
5 x o o o o o x o o o o  o  o  o
6 x o o o o x o o o o o  o  o  o
7 x x x x x x o o o o o  o  o  o
```

Figure 34: EACF plot

Based on the ACF, PACF and EACF plots, the ideal model should be of order (1,1,0).

4.2.4. Fitting the models

```
161  f1 <- arima(ts2, order = c(0,1,1))
162  AIC(f1)
163  f2 <- arima(ts2, order = c(1,1,1))
164  AIC(f2)
165  f3 <- arima(ts2, order = c(1,1,0))
166  AIC(f3)
167 ```

[1] -48.21583
[1] -48.35314
[1] -50.30157

168
```

Figure 34: AIC values for different ARIMA models

The third model with order (1,1, 0) shows the best AIC number and the residuals will be checked for that model.

4.2.5. Checking residuals

```
        Box-Ljung test

data:  residuals(f1)
X-squared = 0.40711, df = 1, p-value = 0.5234


        Augmented Dickey-Fuller Test

data:  residuals(f1)
Dickey-Fuller = -3.603, Lag order = 4, p-value = 0.03598
alternative hypothesis: stationary


        Shapiro-Wilk normality test

data:  rstandard(f1)
W = 0.98514, p-value = 0.2114


        One-sample Kolmogorov-Smirnov test

data:  rstandard(f1)
D = 0.093116, p-value = 0.2491
alternative hypothesis: two-sided
```

Figure 35: Statistics to test normality and stationarity

-The p value for the Ljung-Box test in figure 35 is greater than 0.05 alpha level suggesting that the null hypothesis of the error terms being uncorrelated, should be rejected.

-The ADF p value of the residuals shown in figure 35, is less than the significant level of 0.05 therefore the null hypothesis that the data is not stationary can be rejected.

- The Shapiro-Wilk test has a p value greater than 0.05, the null hypothesis that the data is normally distributed cannot be rejected and it can be concluded that the data is normal.

- The null hypothesis of the one-sample kolmogorov-smirnov test states that there is no difference between the two distributions (normality and time series) , as the p value is greater than 0.05, the null hypothesis cannot be rejected and normality can be assumed.

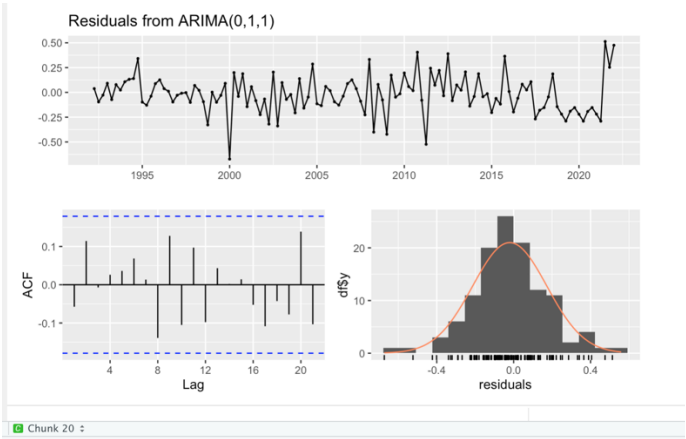


Figure 36: Residual analysis

The lags are within the confidence interval suggesting there is no autocorrelation between the lags. The histogram shows a normally distributed plot with a bell curve and the plot is distributed around 0 suggesting normally distributed data. The residuals don't show an obvious pattern or trend and generally oscillates around 0.

The Q-Q plot is not the most ideal as there are lot of points that are not in line, but this may be due

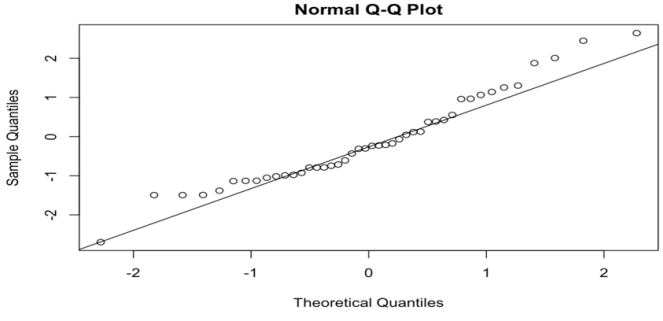


Figure 37: Q-Q plot

to the circumstances around covid even though the outliers were accounted for.

2.3.5. Forecasting

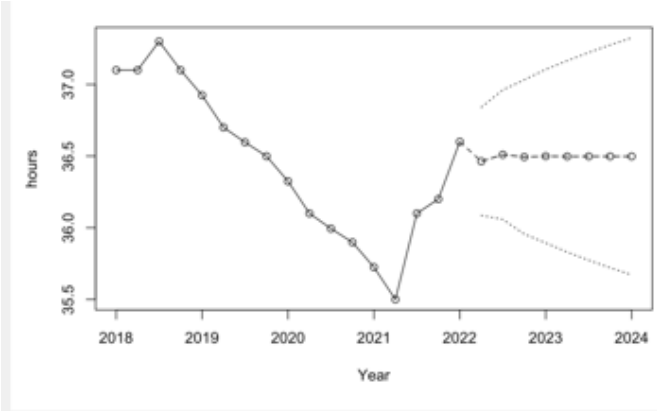


Figure 38: Forecast of hours worked for full time workers

Figure 38 shows the forecast of hours worked by full time workers for the next two years. The dotted lines represent the 95% prediction intervals.

4.3.6: Forecast errors

```
train_ts2 <- window(ts2,end=c(2015,4))
test_ts2 <- window(ts2,start=c(2016,1))
f1_train<-arima(train_ts2,order=c(0,1,1))
f2_train=arima(train_ts2,order=c(1,1,1))
f3_train=arima(train_ts2,order=c(1,1,0))

# predict.Arima: Forecast from ARIMA fits
futurf1<- predict(f1_train,n.ahead = 12)
accuracy(futurf1$pred, test_ts2)

futurf2 <- predict(f2_train,n.ahead = 12)
accuracy(futurf2$pred, test_ts2)

futurf3<- predict(f3_train,n.ahead = 12)
accuracy(futurf3$pred, test_ts2)
```

```
                     ME      RMSE       MAE         MPE      MAPE      ACF1 Theil's U
Test set -0.2040088 0.2721216 0.222117 -0.5486692 0.5968293 0.5180228  1.891315
                    ME      RMSE       MAE         MPE      MAPE      ACF1 Theil's U
Test set -0.19926 0.272481 0.2258211 -0.5360201 0.6066613 0.482382  1.885478
                     ME      RMSE       MAE         MPE      MAPE      ACF1 Theil's U
Test set -0.2243514 0.2925407 0.2447379 -0.6032329 0.6574522 0.4639872  2.027528
```

Figure 39: Forecast errors of three models

The first model with order (0,1,1) has a lower RMSE value than the model selected with the lowest AIC value (ARIMA (1,1,0) therefore it would be wise to select the model with the lower RMSE value as it would be more accurate in predicting the forecast.

4.3 Conclusion

An ARIMA model was created due to the series showing trends, this trend was analysed and removed in order to make it stationary. Outliers were considered as they would greatly affect the analysis and forecasting, even after using a function to remove them, there was still a dip in

the time series. R The ACF and PACF plots were investigated to estimate the ‘p’ and ‘q’ value. Once the model was fitted with the new values, the residuals were analysed, and the forecast error was found. From the forecast error, the model with the lowest AIC value had a slightly higher RMSE value. Therefore, another model (ARIMA (0,1,1) was selected as the final model to be used for predicting a forecast. The MAPE value suggests that the model was 99% accurate in predicting the time series value in the test-train split. As the data from the pandemic was kept it could be said that the forecast is quite representative of actual hours worked and would be realistic in predicting future hours worked by full time workers.