

# Final Project Report

Declaration of contributions:

No	Name of Team Member	Contribution to the project
1	Matthew John Bentham	0.5
2	John Fergus Murrowood	0.5
	TOTAL	1

RMIT University,

School of Science

2022

# Energy Demand Forecast Victoria

MATH1318: Final Project

Statistically Significant Squad

Matthew Bentham 3923076, John Murrowood 3923075



## Contents

Contents.....	2
1. Introduction.....	3
1.1 Functions used.....	3
2. Preliminary Analysis .....	5
2.1 Import and Analyse Dataset.....	5
2.2 Impact of previous days.....	9
3. Identifying seasonality .....	10
4. Data transformations .....	12
5. Simple deterministic model (Linear + Seasonal model) .....	16
5.2 Forecast.....	18
6. SARIMA model.....	19
6.1 Residual Approach.....	20
6.2 Model Fitting and Diagnostic Checking .....	27
6.2.1 Parameter significance .....	27
6.2.2 BIC/AIC values.....	44
6.2.3 Accuracy Scores .....	45
6.3 Summary .....	46
6.4 Model Diagnostics.....	47
6.5 Forecasting.....	48
7. Conclusion.....	49
8. References .....	50

# 1. Introduction

**Research Question:** What are the most accurate forecasts for monthly averaged total daily electricity demand in Victoria for the next 10 months?

With the resurrection of the State Electricity Commission (SEC), announcement of offshore wind farms in southeast Victoria and Victoria's renewable energy target of 50% by 2030 [1], being able to accurately predict energy demand on a continual basis is becoming increasingly important. As renewable energy sources like wind and solar do not produce constant energy outputs, accurate forecasting of energy demand throughout the day is essential to ensure renewable energy sources are used efficiently. Demand forecasting has become increasingly relevant when incorporating renewable into the grid and in general reduces over and underproduction of energy and minimises energy waste as a whole.

**Data Source:** <https://www.kaggle.com/datasets/aramacus/electricity-demand-in-victoria-australia> [2]

This data set contains the total daily energy demand across Victoria in MWh from 1st Jan 2015 to 6 Oct 2020 which consists of 2016 days. Although the additional features of this data set may not be directly relevant to this report, below is a list of variables contained in the data and their description:

- **date:** datetime, the date of the recording.
- **demand:** float, a total daily electricity demand in MWh.
- **RRP:** float, a recommended retail price in AUD\$ / MWh.
- **demand\_pos\_RRP:** float, a total daily demand at positive RRP in MWh.
- **RRP\_positive:** float, an averaged positive RRP, weighted by the corresponding intraday demand in AUD\$ / MWh.
- **demand\_neg\_RRP:** float, a total daily demand at negative RRP in MWh.
- **RRP\_negative:** float, an average negative RRP, weighted by the corresponding intraday demand in AUD\$ / MWh.
- **frac\_at\_neg\_RRP:** float, a fraction of the day when the demand was traded at negative RRP.
- **min\_temperature:** float, minimum temperature during the day in Celsius.
- **max\_temperature:** float, maximum temperature during the day in Celsius.
- **solar\_exposure:** float, total daily sunlight energy in MJ/m<sup>2</sup>.
- **rainfall:** float, daily rainfall in mm.
- **school day:** boolean if students were at school on that day.
- **holiday:** boolean if the day was a state or national holiday.

*Note:* All code provided is written in Rstudio using R 4.2.0

## 1.1 Functions used:

```
sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}
```

```
Normalitytests <- function(data,title) {
```

```

# Create a QQ plot of the data

qqnorm(y=data, main=title)
qqline(y=data, col=2, lwd=1, lty = 2)

# Perform a Shapiro-Wilk test on the data
shapiro.test(as.numeric(data))
}

Plot_Model <- function(data,model_num,fignum,p,d,q,P,D,Q) {

  layout(matrix(c(1,1,2,3), nrow = 2, ncol = 2, byrow = TRUE))

  m.ts = Arima(data,order=c(p,d,q),seasonal=list(order=c(P,D,Q), period=12))
  res.m = residuals(m.ts);

  plot(res.m,xlab='Time',ylab='Residuals',main=combine_words(c("Figure ", fignum, ":
model " , model_num, " residuals "),sep = "",
and = ""))

  acf(res.m,lag.max=48, main=paste("ACF plot"))
  pacf(res.m,lag.max=48, main=paste("PACF plot"))

  return(m.ts)
}

# This function was extracted from MATH1318 tutorial 8 (Haydar Demirhan)
residual.analysis <- function(model, fignum,std = TRUE,start = 2, class =
c("ARIMA", "GARCH", "ARMA-GARCH", "fGARCH")[1],method="",o1="",o2=""){
  library(TSA)

  if (class == "ARIMA"){
    if (std == TRUE){
      res.model = rstandard(model)
    }else{
      res.model = residuals(model)
    }
  }else if (class == "GARCH"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "ARMA-GARCH"){
    res.model = model@fit$residuals
  }else if (class == "fGARCH"){
    res.model = model@residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }
  par(mfrow=c(2,3), mar = c(1, 1,6, 2))
  if (o1 != ""){
    plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of
standardised residuals")
    mtext(combine_words(c("Figure ", fignum, ": SARIMA",o1,o2," Residual Analysis
(",method,"))",sep = "",

```

```

    and = ""),
    side = 3,
    line = -2,
    outer = TRUE)}
else{plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of
standardised residuals")
mtext(combine_words(c("Figure ", fignum, ": Residual Analysis") ,sep = "",
and = ""),
    side = 3,
    line = -2,
    outer = TRUE)

}
abline(h=0)
hist(res.model,main="Histogram of standardised residuals")
qqnorm(res.model,main="QQ plot of standardised residuals")
qqline(res.model, col = 2)
acf(res.model,main="ACF of standardised residuals")
print(shapiro.test(res.model))
k=0
LBQPlot(res.model, Lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)

par(mfrow=c(1,1))
}

Parameter. estimation <- function(data,order1,order2,method,fignum){
  model = Arima(data,order=order1,seasonal=list(order=order2, period=12),method =
method)

  residual.analysis(model = model,fignum =
fignum,method=method,o1=combine_words(c('(',combine_words(c(order1),sep = ",",
and = ""),')'),sep="","",and=""),o2=combine_words(c('(',combine_words(c(order2),sep = ",",
and = ""),')'),sep="","",and=""))
  print(coeftest(model))
  return(model)
}

Stationary_tests <- function(data){
  print(adf.test(data))
  print(pp.test(data))
  print(kpss.test(data))
}

```

## 2. Preliminary Analysis

### 2.1 Import and Analyse Dataset

To further investigate the time series data, the data set is first imported and converted to a time series object. This time series objects are then plotted using the plot() function to visualise the main characteristics.

```
Data <- read_csv("complete_dataset.csv")
```

```
# Convert Dataframe to Timeseries:
```

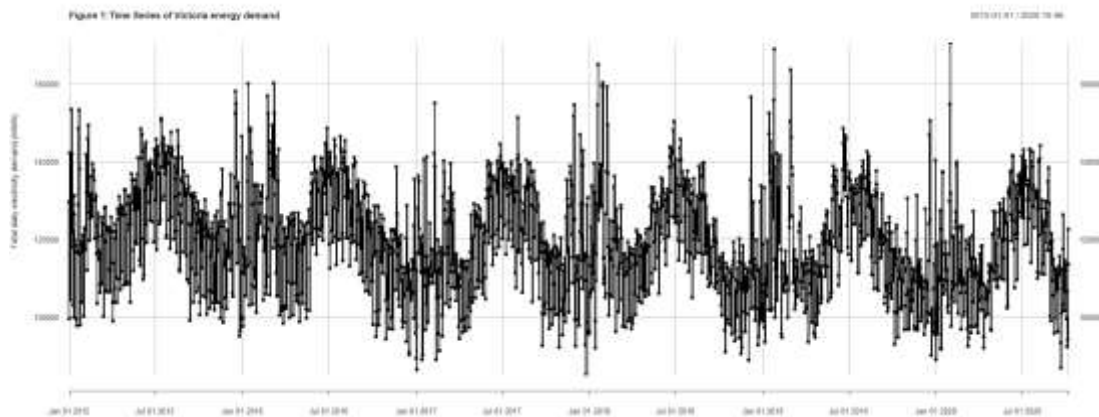
```
Data_xts <- xts(Data$demand, Data$date)
```

```
kable(head(Data))
```

date	demand	RRP	demand_pos_RRP	RRP_positive	demand_neg_RRP	RRP_negative	frac_at_neg_RRP	min_temperature	max_temperature	solar_exposure	rainfall	school_day	holiday
2015-01-01	99635.0	25.6	97319.2	26.4	2315.8	-7.2	0.0	13.3	26.9	23.6	0.0	N	Y
2015-01-02	129606.0	33.1	121082.0	38.8	8524.0	-47.8	0.1	15.4	38.8	26.8	0.0	N	N
2015-01-03	142300.5	34.6	142300.5	34.6	0.0	0.0	0.0	20.0	38.2	26.5	0.0	N	N
2015-01-04	104330.7	25.0	104330.7	25.0	0.0	0.0	0.0	16.3	21.4	25.2	4.2	N	N
2015-01-05	118132.2	26.7	118132.2	26.7	0.0	0.0	0.0	15.0	22.0	30.7	0.0	N	N
2015-01-06	130672.5	31.3	130672.5	31.3	0.0	0.0	0.0	17.7	26.0	31.6	0.0	N	N

```
#Plot the Time Series:
```

```
plot(Data_xts, type="o", ylab="Total daily electricity demand (MWh)", xlab="Year",  
main="Figure 1: Time Series of Victoria energy demand")
```



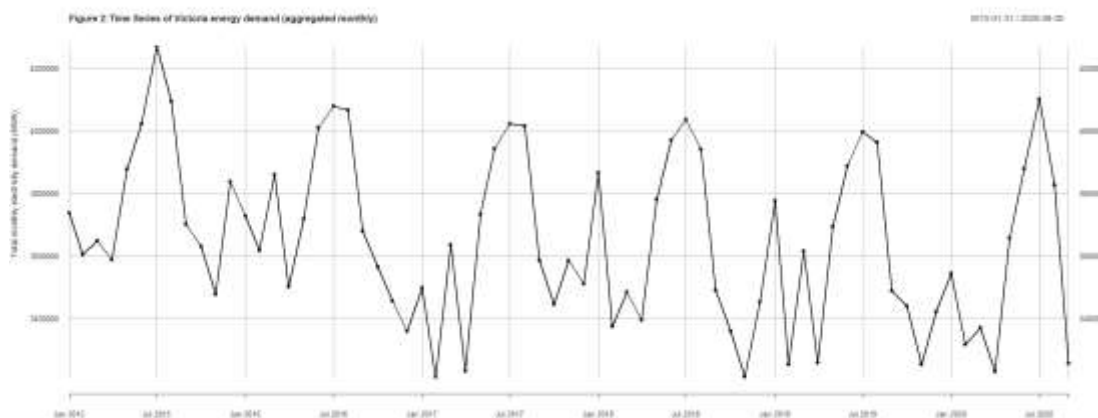
- **Trend:** A very slight downward trend.
- **Variance:** No obvious changing variance.
- **Seasonality:** It can be seen all the peaks occurring around July each year with short spikes occurring every January. This is as expected as generally more electricity is used during winter with the exception of very hot days during January which is likely due to the large uptake in cooling systems.
- **Change Point:** No clear change point observed.

As you can see in the plot above the time series data has relatively high granularity due to the fact daily data points are used. Because of this, this data is likely to contain multiple seasonal components on the day, week, and month level. As the main objective of this report is to identify and forecast trends in electricity demand on a monthly basis, this time series data will be aggregated on a monthly basis.

```
# remove last 6 data points as they belong to separate month and will skew the summations
n<-dim(Data)[1]
df<-Data[1:(n-6),]
Data_xts <- xts(df$`demand`, df$date)

Data_xts_agg <- apply.monthly(Data_xts,sum)

#Plot the Time Series:
plot(Data_xts_agg,type="o",ylab="Total monthly electricity demand (MWh)",xlab="Year",
     main="Figure 2:Time Series of Victoria energy demand (aggregated monthly)")
```



Time series characteristics observed from the above plot:

- **Trend:** A very slight downward trend.
- **Variance:** No obvious changing variance.
- **Seasonality:** Needs to be further investigated, however annual repeating patterns can be seen around every January and July each year which strongly suggests the existence of seasonality.
- **Change Point:** No clear change point observed.

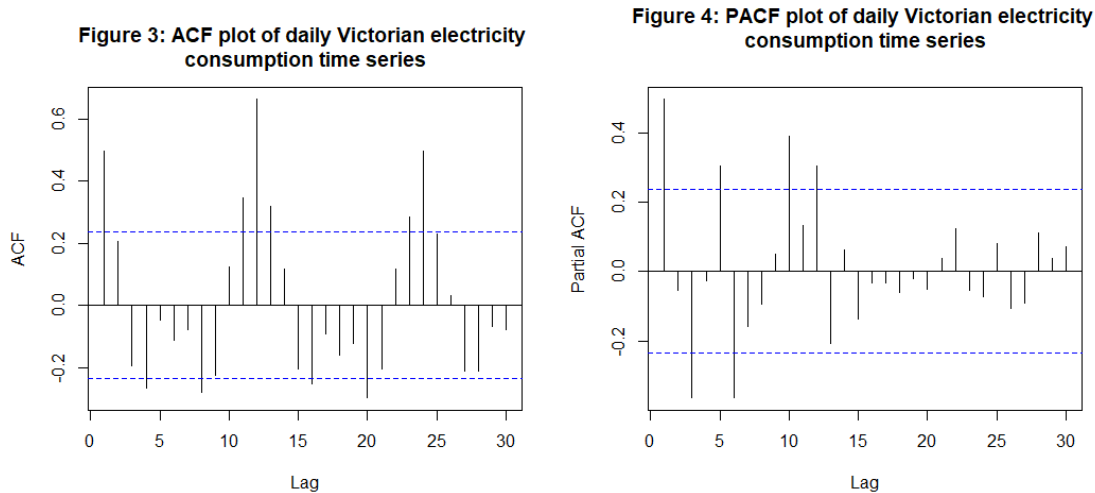
```
# Summary Stats
summary(Data_xts_agg)
```

Index	Data_xts_agg
Min. :2015-01-31	Min. :3209217
1st Qu.:2016-06-30	1st Qu.:3453621
Median :2017-11-30	Median :3631206
Mean :2017-11-29	Mean :3654578
3rd Qu.:2019-04-30	3rd Qu.:3877235
Max. :2020-09-30	Max. :4273538

As seen in the summary data above, the data has a mean of 3654578MWh and a median of 3631206MWh, which is relatively similar indicating that the distribution is roughly symmetric if not slightly right skewed. The min and max have a difference of 1064321MWh which is relatively large, indicating that there is significant variability in the data. The IQR of the data set is 423614MWh which indicates a moderate spread of data points around the mean. As both q1 and q3 are similar distance away from the median it can be concluded the data is relatively symmetric, however the slightly greater difference between q3 and the median does further indicate a slight positive skew.

## # ACF and PACF

```
par(mfrow=c(1,2))
acf(Data_xts_agg, lag.max=30, main="Figure 3: ACF plot of daily Victorian electricity
\consumption time series")
pacf(Data_xts_agg, lag.max=30, main="Figure 4: PACF plot of daily Victorian electricity
\consumption time series")
```



From the ACF plot above we can see a very clear wave-like pattern which indicates the presence of seasonality in our data. The PACF exhibits a slowly decreasing pattern which suggests a gradual decay of autocorrelation, suggesting the presence of a long-term/weak auto-regressive trend in the data. From this one can roughly deduce that a SARIMA model is likely going to be the best model-type for this data.

## # ADF test to test for stationarity

```
Stationary_tests(Data_xts_agg)
```

```
##
## Augmented Dickey-Fuller Test
##
## data: data
## Dickey-Fuller = -3.2417, Lag order = 4, p-value = 0.08841
## alternative hypothesis: stationary
##
##
## Phillips-Perron Unit Root Test
##
## data: data
## Dickey-Fuller Z(alpha) = -36.579, Truncation lag parameter = 3, p-value
## = 0.01
## alternative hypothesis: stationary
##
##
## KPSS Test for Level Stationarity
##
## data: data
## KPSS Level = 0.31195, Truncation lag parameter = 3, p-value = 0.1
```

For the ADF test, a p-value less 0.05 indicates that under the 95% confidence interval the null hypothesis of non-stationarity can be rejected. The ADF test above shows a p-value of 0.088 which is greater than 0.05, meaning the null hypothesis cannot be rejected and the time series is assumed too NOT be stationary. Phillips-

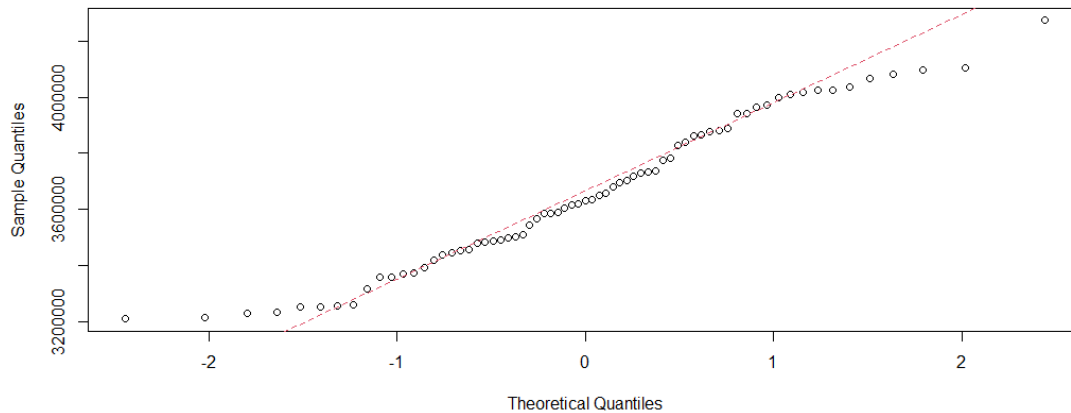


Perron also supported this with a p-value less than 0.05, meaning under the general consensus, non-stationarity can be assumed.

# Check the degree of normality

```
Normalitytests(Data_xts_agg, "Figure 5: QQ plot of Victorian electricity consumption time series")
```

Figure 5: QQ plot of Victorian electricity consumption time series



```
##
## Shapiro-Wilk normality test
##
## data: as.numeric(data)
## W = 0.96648, p-value = 0.06078
```

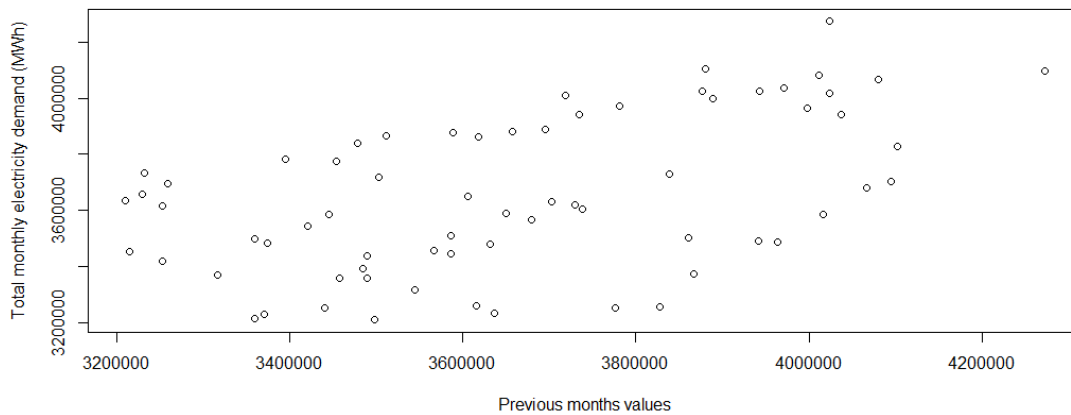
To test for normality in the data, both a shapiro-wilk test and Q-Q plot were used. The shapiro-wilk test achieved a p-value of 0.06 and as this is greater than 0.05, we cannot reject the null hypothesis of normality. The Q-Q plot does show that the data follows the centreline for the majority of points, however, there still is deviation at the ends of the data meaning this data likely skirts on the edge of normality.

## 2.2 Impact of previous days

Another useful initial analysis to conduct is to determine whether or not previous years share a strong correlation or not.

```
# plot lag 1 scatterplot
plot(y=Data_xts_agg, x=zlag(Data_xts_agg), yLab="Total monthly electricity demand (MWh)",
     xLab='Previous months values', main = "Figure 6: Lag 1 Plot.")
```

Figure 6: Lag 1 Plot.



```
y = Data_xts_agg
x = zlag(Data_xts_agg) # Generate first lag of the series
index = 2:length(x) # Create an index to get rid of the first NA value in x
cor(y[index],x[index]) # Compute correlation coefficient of the data with its first lag

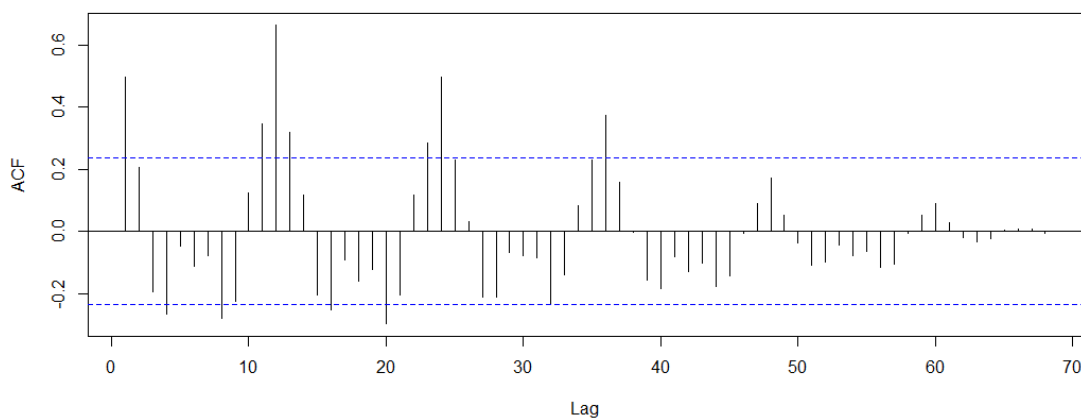
##          [,1]
## [1,] 0.5041434
```

The scatter plot above shows that there is a weak positive correlation between the energy demand on a given month and its succeeding months. The correlation coefficient of 0.50 shows that the correlation is relatively weak and confirms that the time series data exhibits some form of positive autocorrelation, however not a significant amount and the data is NOT randomly distributed.

### 3. Identifying seasonality

```
#Plot acf plot of time series data
acf(Data_xts_agg, lag.max = 70, main="Figure 7: ACF plot of Victorian electricity
consumption time series")
```

Figure 7: ACF plot of Victorian electricity consumption time series



Because a wave pattern with clear significant lags was seen in the ACF plot above and because the general time series plot shows clearly repeating patterns in July and January each year it and therefore non-stationary which will be further investigated. Using the ACF plot the number of lags between each peak in the observed oscillation pattern can be used to determine frequency.

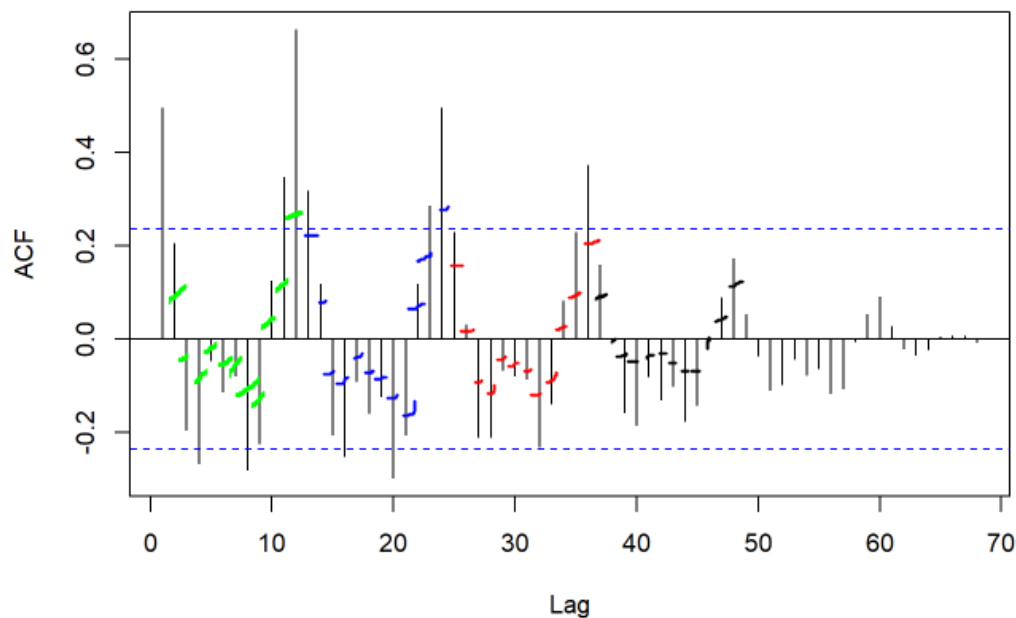


Figure 8: Calculating frequency.

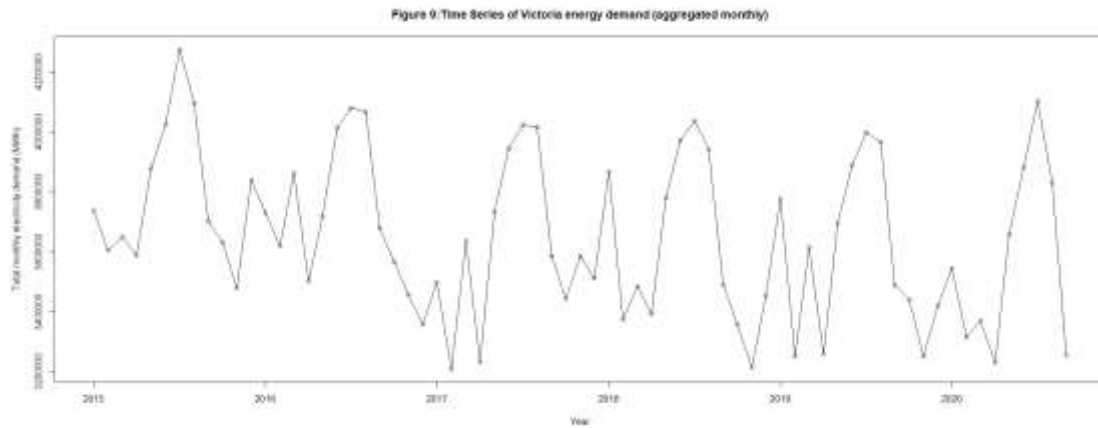
As shown in figure 8, peaks occur at lags 1, 12, 24 etc. which suggests the seasonality of the series exhibits a frequency of approximately 12. This is also expected in terms of the context of the problem as energy usage is likely to go up every winter and start of summer. Because of this the frequency of 12 will be embedded into the time series itself.

```
# Convert to ts object so that we can add frequency:
Data <- df                                # Duplicate data
Data$year <- strftime(Data$date, "%Y")    # Create year column
Data$month <- strftime(Data$date, "%m")   # Create month column

Data_agg <- aggregate(`demand` ~ month + year,      # Aggregate data
                      Data,
                      FUN = sum)

# Convert data into a time series object
Data_ts_agg = matrix(Data_agg$`demand`)
Data_ts_agg = as.vector(t(Data_ts_agg))

Data_ts_agg <- ts(Data_ts_agg, frequency=12, start = c(2015, 1), end = c(2020, 9) )
plot(Data_ts_agg, type="o", ylab="Total monthly electricity demand (MWh)", xlab="Year",
     main="Figure 9: Time Series of Victoria energy demand (aggregated monthly)")
```

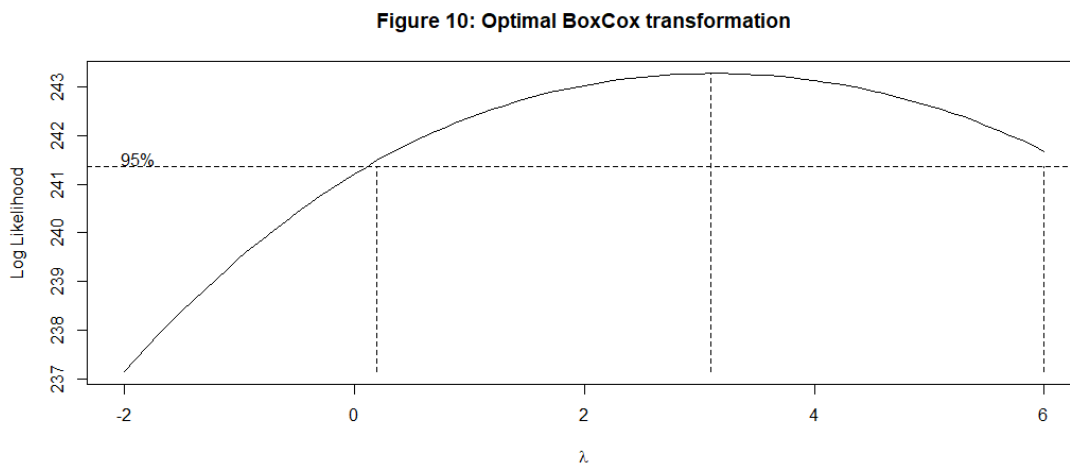


## 4. Data transformations

Because the ACF and PACF plots indicate the presence of trends in the data (downward trends) and because the data exhibited both non-stationarity and non-normality, transformations will be tested in the hopes of removing this trend. Because Log transformations are generally used for variance stabilisation, a more general transformation called Box-Cox was tested.

# As the data is not normal a box cox transformation will be used to see if this improves normality

```
BC = BoxCox.ar(Data_ts_agg, lambda = seq(-2, 6, by = 0.1))
title(main = "Figure 10: Optimal BoxCox transformation")
```



```
Lambda <- BC$Lambda[which(max(BC$Loglike) == BC$Loglike)]
Lambda
```

```
## [1] 3.1
```

The optimal value of lambda in this instance is 3.1.

# Apply box cox transformation to time series and see if normality has improved

```
Data_ts_agg_BC = (Data_ts_agg^lambda-1)/lambda
```

#Plot the Time Series:

```
plot(Data_ts_agg_BC, type="o", ylab="Total monthly electricity demand (MWh)", xlab="Year",
     main="Figure 11: Time Series of Victoria energy demand (aggregated monthly)")
```



As can be seen in figure 11 above, the BoxCox transformation does not appear to have had much effect on the time series.

```
Stationary_tests(Data_ts_agg_BC)
```

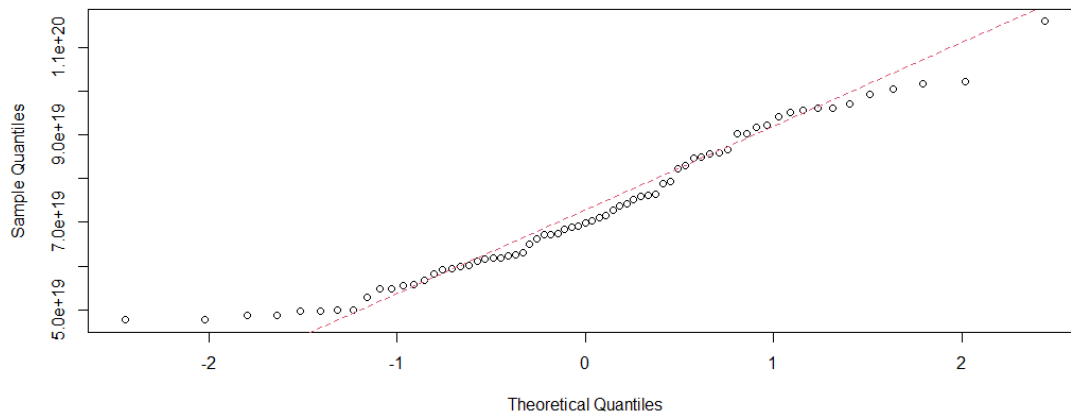
```
##
## Augmented Dickey-Fuller Test
##
## data: data
## Dickey-Fuller = -3.3056, Lag order = 4, p-value = 0.07819
## alternative hypothesis: stationary
##
##
## Phillips-Perron Unit Root Test
##
## data: data
## Dickey-Fuller Z(alpha) = -34.509, Truncation lag parameter = 3, p-value
## = 0.01
## alternative hypothesis: stationary
##
##
## KPSS Test for Level Stationarity
##
## data: data
## KPSS Level = 0.28542, Truncation lag parameter = 3, p-value = 0.1
```

It can be seen in both the ADF and pp test above, the BoxCox transformation slightly increased the stationarity of the time series.

```
# Check if normality has been improved
```

```
Normalitytests(Data_ts_agg_BC, "Figure 12: QQ plot of BC Victorian electricity consumption time series")
```

Figure 12: QQ plot of BC Victorian electricity consumption time series

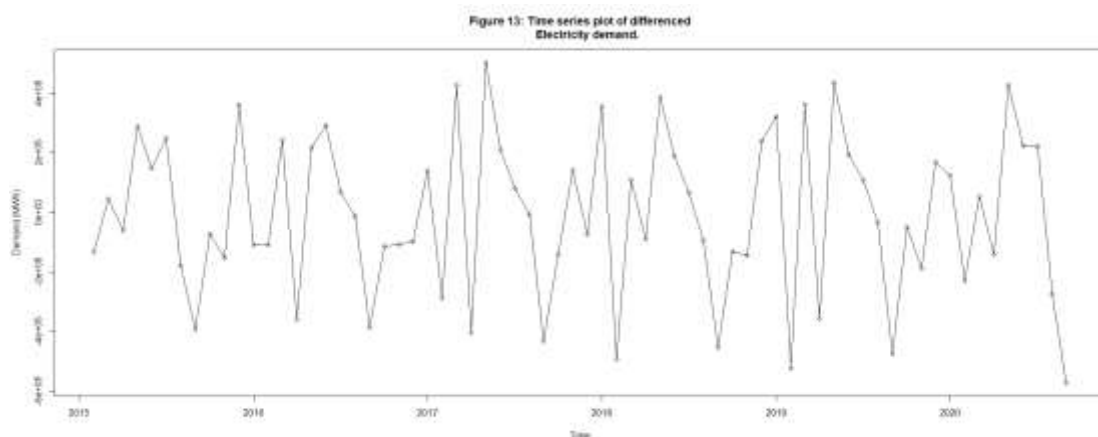


```
##
## Shapiro-Wilk normality test
##
## data: as.numeric(data)
## W = 0.95668, p-value = 0.01762
```

Based on the results of the Shapiro-Wilk test, it is observed that the Box-Cox transformation has made the time series less normal, as indicated by a p-value of 0.02. Consequently, since the transformation has had minimal impact on achieving stationarity and improving normality, it will not be utilized for the model.

Given that the model was not stationary, an initial first difference will be employed. This different approach aims to remove any underlying trend present in the data and make it stationary, which is a crucial requirement for many time series models.

```
# applying first differencing
diff.Data_ts_agg <- diff(Data_ts_agg)
par(mfrow=c(1,1))
plot(diff.Data_ts_agg, type='o', ylab = "Demand (MWh)", main='Figure 13: Time series plot
of differenced
Electricity demand.')
```



Upon initial inspection of the first differenced time series plot, it is observed that there are no apparent trends or changing variances. This suggests that the first difference alone might be sufficient to make the time series stationary.

```
# Use adf test on differenced series
Stationary_tests(diff.Data_ts_agg)

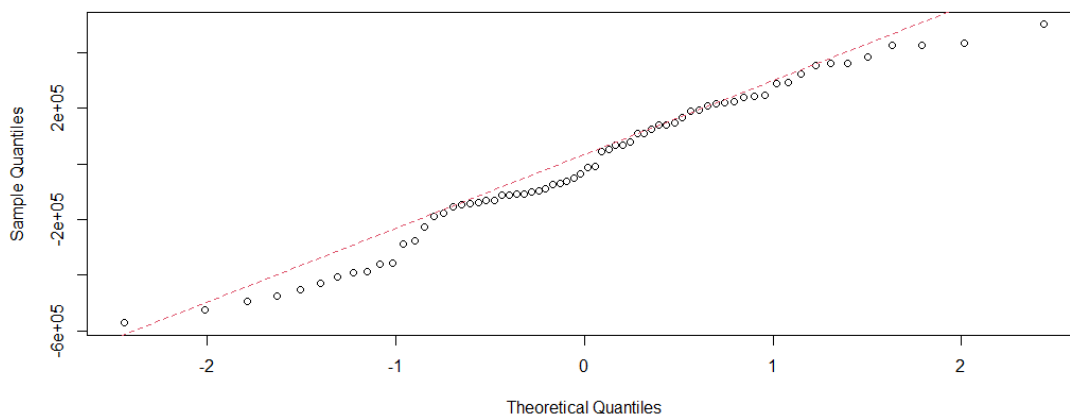
##
## Augmented Dickey-Fuller Test
##
## data: data
## Dickey-Fuller = -4.345, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## Phillips-Perron Unit Root Test
##
## data: data
## Dickey-Fuller Z(alpha) = -74.581, Truncation Lag parameter = 3, p-value
## = 0.01
## alternative hypothesis: stationary
##
##
## KPSS Test for Level Stationarity
##
## data: data
## KPSS Level = 0.028624, Truncation Lag parameter = 3, p-value = 0.1
```

Based on the results of the Augmented Dickey-Fuller (ADF) test, the p-value of 0.01 is less than the significance level of 0.05. Thus, the null hypothesis of non-stationarity can be rejected, and the differenced series can be considered stationary at a 95% confidence level.

While the Phillips-Perron (PP) test may have yielded a p-value less than 0.05, it is important to note that the ADF and KPSS tests both agree on the stationarity of the differenced series. When multiple tests produce conflicting results, it is common practice to rely on the consensus among different tests. In this case, since the ADF and KPSS tests align in indicating stationarity, it is reasonable to assume the series is stationary.

```
# Test for normality of difference plot
Normalitytests(diff.Data_ts_agg, "Figure 14: QQ plot of difference time series")
```

Figure 14: QQ plot of difference time series

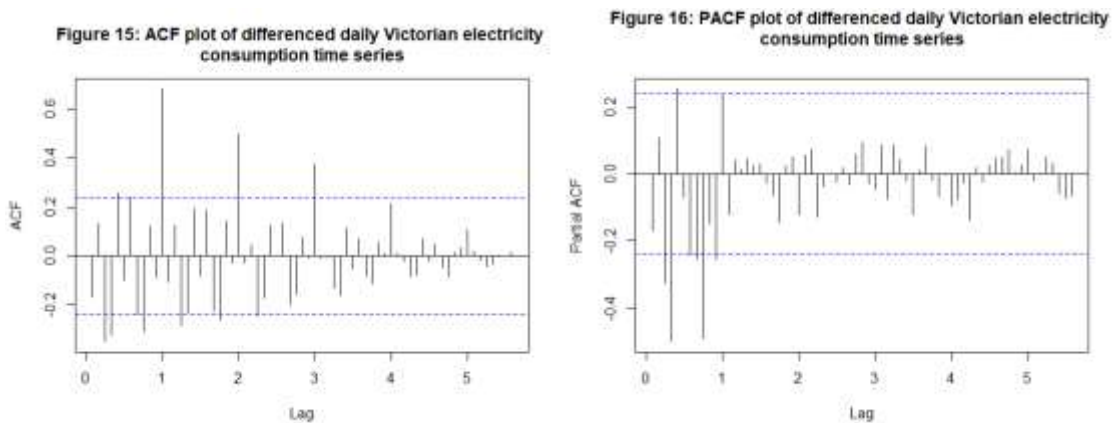


```
##
## Shapiro-Wilk normality test
```

```
##
## data: as.numeric(data)
## W = 0.97446, p-value = 0.1759
```

The Shapiro-Wilk test resulted in a p-value greater than 0.05, indicating that, under the 95% confidence level, we can assume normality of the data. This is beneficial when fitting time series models because many models, such as SARIMA (Seasonal Autoregressive Integrated Moving Average), assume normality of the residuals.

```
# ACF and PACF to find period of differenced series
par(mfrow=c(1,2))
acf(diff.Data_ts_agg, lag.max = 70, main="Figure 15: ACF plot of differenced daily Victorian electricity consumption time series")
pacf(diff.Data_ts_agg, lag.max = 70, main="Figure 16: PACF plot of differenced daily Victorian electricity consumption time series")
```



From the ACF and PACF plots above you can see that significant lags are still occurring with slightly reduced wave-like pattern. The significant lags in the ACF plot indicate that past values of the time series are correlated with the current value, further indicating the presence of a seasonal pattern. Considering both the ACF and PACF, it becomes evident that the data still exhibits seasonality and trend components. These components need to be properly accounted for in subsequent modelling steps to ensure accurate forecasting and capturing the underlying patterns.

Therefore, it is crucial to incorporate appropriate seasonal and trend models, such as SARIMA (Seasonal Autoregressive Integrated Moving Average) or other techniques designed to manage time series data with seasonal and trend components.

## 5. Simple deterministic model (Linear + Seasonal model)

Before delving into more complex SARIMA models, we will start by fitting and analysing a simpler deterministic model. The chosen model for this initial analysis is a hybrid additive model, which combines a seasonal model to capture yearly seasonality and a linear model to capture the decreasing trend.

```
t <- time(Data_ts_agg)
month.=season(Data_ts_agg) # period added to improve table display and this line sets up indicators
Hybridmodel=Lm(Data_ts_agg~month.+t-1) # -1 removes the intercept term
summary(Hybridmodel)

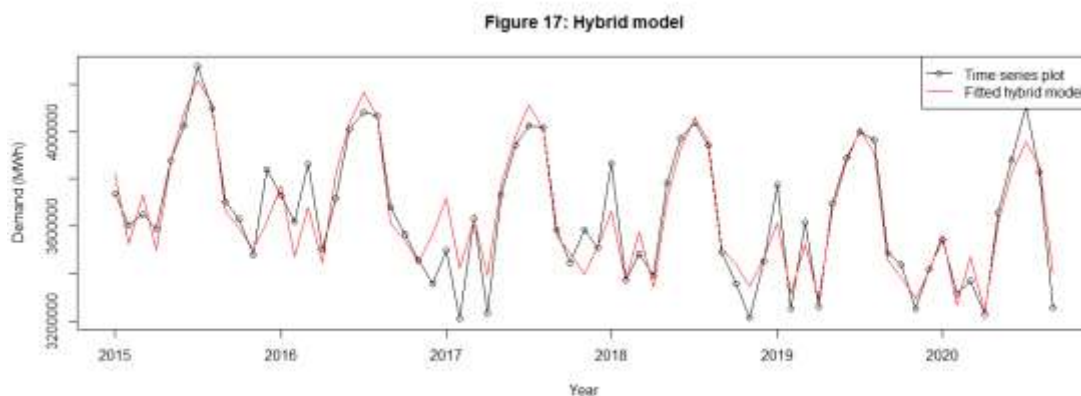
##
## Call:
## lm(formula = Data_ts_agg ~ month. + t - 1)
```



```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -220520  -41803    3449   52567  218496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## month.January  108231300  14946905  7.241 1.38e-09 ***
## month.February  107939393  14947522  7.221 1.49e-09 ***
## month.March     108150783  14948139  7.235 1.41e-09 ***
## month.April     107919614  14948757  7.219 1.50e-09 ***
## month.May       108300373  14949374  7.244 1.36e-09 ***
## month.June      108514093  14949992  7.258 1.29e-09 ***
## month.July      108650996  14950609  7.267 1.25e-09 ***
## month.August    108554887  14951226  7.261 1.28e-09 ***
## month.September 108107398  14951844  7.230 1.44e-09 ***
## month.October   108040260  14948768  7.227 1.46e-09 ***
## month.November  107953793  14949386  7.221 1.49e-09 ***
## month.December  108077317  14950003  7.229 1.45e-09 ***
## t              -51816      7409  -6.994 3.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101400 on 56 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9992
## F-statistic: 6921 on 13 and 56 DF, p-value: < 2.2e-16
```

This hybrid model archived an adjusted R squared score of 0.9992, this overly high score is very worrying and is a good indication of over fitting. Additionally, all the coefficients archive p-values higher than 0.05, meaning they are found to be statistically significant.

```
plot(Data_ts_agg, main = "Figure 17: Hybrid model", yLab="Demand (MWh)", xLab='Year'
     , type='o')
points(t, predict.lm(Hybridmodel), type='l', col='red') # superimpose the fit of model reg0
on the plot of the data
# Add legend
legend("topright", legend = c("Time series plot", "Fitted hybrid model"),
      col = c("black", "red"), lwd = 1, pch = c(1, NA))
```



The plot of the trend model shows that the hybrid model fits the time series to a moderate extent. It has some trouble fitting the seasonal trends in the summer months which is likely due to the stochastic nature of these trends.

## 5.1 Residual analysis

```
# Get standardized residuals
```

```
Hybridmodel.res=rstudent(Hybridmodel)
```

```
residual.analysis(Hybridmodel, '18')
```

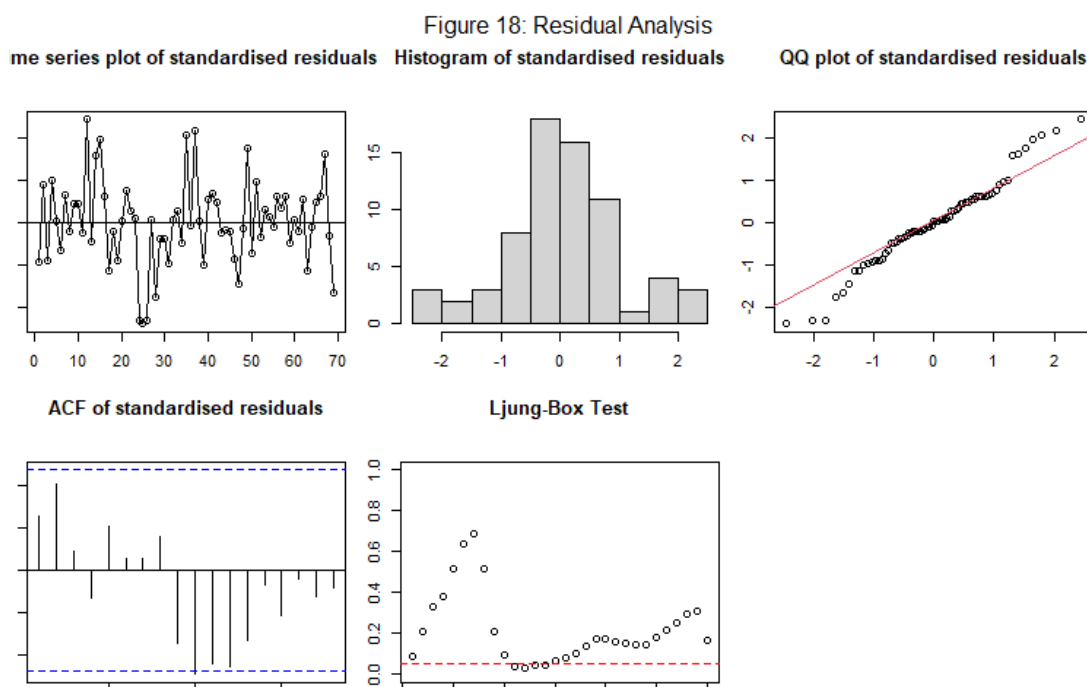
```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: res.model
```

```
## W = 0.97373, p-value = 0.1542
```



- **Histogram:** The histogram is relatively normal.
- **QQ Plot:** The qqplot observed in this model at the tails of the data still significantly deviates from the centre line.
- **ACF:** the lack of a wave-like pattern and significant lags shows that the models were able to significantly capture the seasonality and correlation present in the data.
- **Shapiro-wilk test:** The shapiro-wilk normality test produced a p-value of 0.154. This value is greater than 0.05, meaning under a 95% confidence level and the null hypothesis is NOT rejected, meaning no significant departure from normality was found.
- **Ljung-Box:** Not all values are greater than the confidence interval, suggesting there may be some trend that is not being tracked by the model.

## 5.2 Forecast

```
h <- 0.8333333333333334 # 10 months ahead forecasts
```

```
increment <- 12 # Frequency is 7
```

```
original.t <- t
```

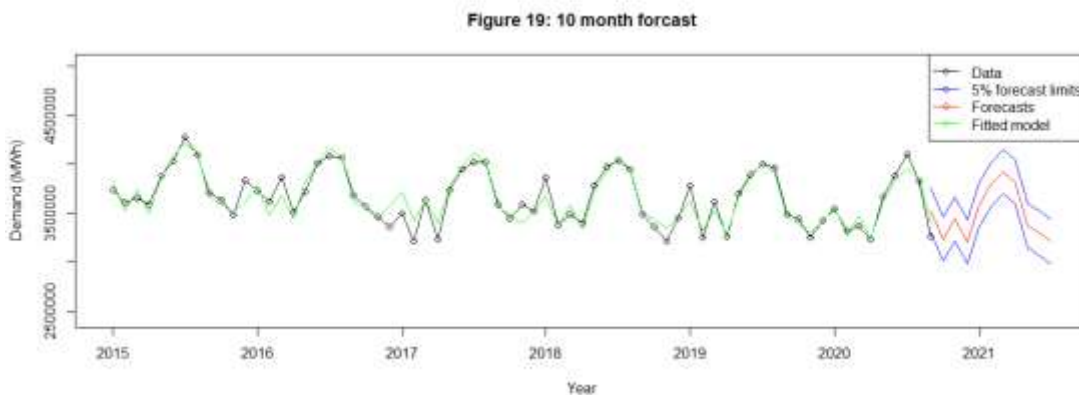
```

n <- length(original.t) # Find the length of observation period
lastTimePoint <- original.t[n] # Find the last time point
ts <- seq(lastTimePoint ,
lastTimePoint + h, 1/12)
new<- factor(rep(month.,1))[1:11] # Introducing the seasonal value for forecasting\\
daysahead <- data.frame(month. = new,t=ts)
pred <- predict.lm(Hybridmodel,daysahead,interval='prediction') # Computing the
prediction as well as prediction interval

# plot forecasted values
par(mfrow=c(1,1))
plot(Data_ts_agg, main = "Figure 19: 10 month forecast",ylab="Demand (MWh)",xlab='Year'
, type='o',ylim = c(min(c(fitted(Hybridmodel))-800000, as.vector(Data_ts_agg))),
max(c(fitted(Hybridmodel)+800000,
as.vector(Data_ts_agg)))),xlim=c(2015,2021.500))

points(t,predict.lm(Hybridmodel),type='l',col='green') # superimpose the fit of model
reg0 on the plot of
lines(ts(as.vector(pred[,1]), start = c(2020.667,1),frequency = 12), col="red", type="l")
lines(ts(as.vector(pred[,2]), start = c(2020.667,1),frequency = 12), col="blue",
type="l")
lines(ts(as.vector(pred[,3]), start = c(2020.667,1),frequency = 12), col="blue",
type="l")
# Add legend
legend("topright", legend = c("Data","5% forecast limits", "Forecasts","Fitted model"),
col = c("black", "blue","red","green"), lwd = 1, pch = 1)

```



From the above plot we can see that the predictions do seem to follow the same very slow downward trend and the rough seasonal variations in the data. The red line represents the fitted values for the next 10 months, while the blue lines indicate the upper and lower confidence intervals for the fitted values. For the simplicity of the model this is an adequate forecast, however its precision is affected by the stochastic nature of the seasonal trends as the difference in peaks between years does not seem to follow a specific trend.

## 6. SARIMA model

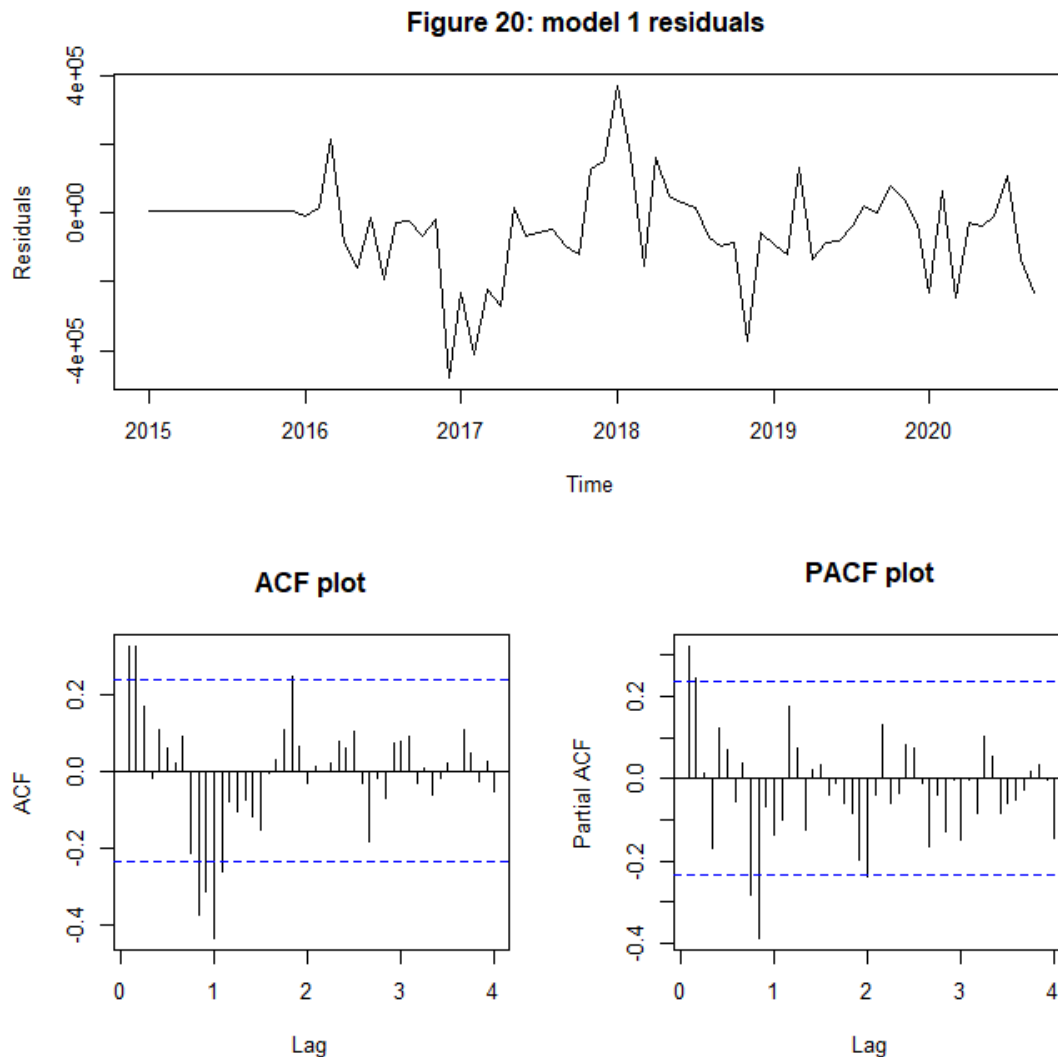
A more complex model that will be used to forecast this dataset will be a SARIMA model. This is because there is clear seasonality in the monthly aggregated time series with a period of 12 months. There is also relatively constant variance with no clusters of high variance and therefore a GARCH model is not suitable for this time series. This model is therefore more likely to notice the slight year deviations in values. To begin finding

possible PDQ and pdq parameters for the SARIMA model, first the seasonal component must be modelled by finding the P and Q values. The residuals analysis approach was used to find possible seasonal P and Q values.

## 6.1 Residual Approach

To deal with the seasonal trend effect, we will fit an initial plain model (model 1) with the first seasonal difference to see if the seasonal trend is affected and to what extent.

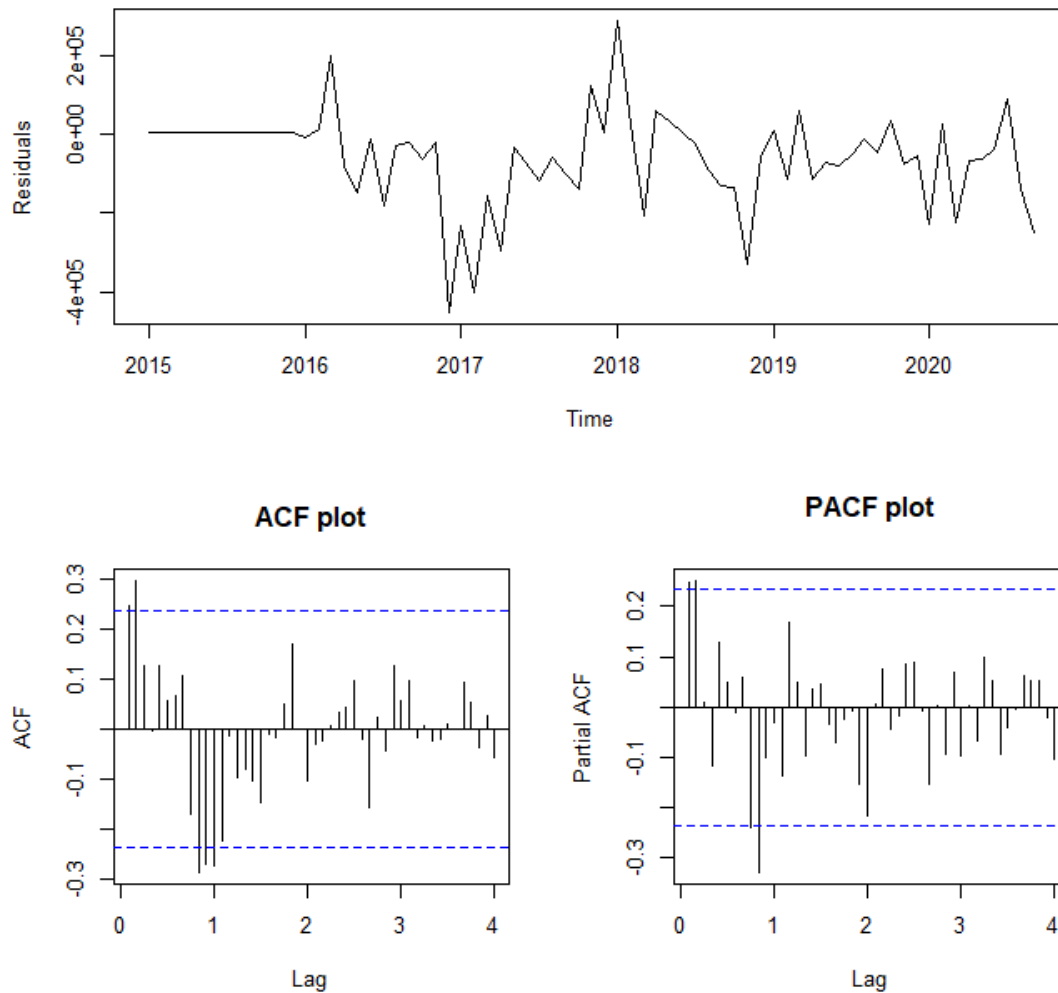
```
# Model 1 with first seasonal differencing
m1.ts<-Plot_Model(Data_ts_agg,1,"20",0,0,0,0,1,0)
```



It can be seen in the Time series plot of the residuals plotted above, there is no obvious seasonal components and therefore this model can likely account for the seasonality in the time series. It can be seen in the ACF plot in figure 20 above that there is a significant lag on lag 1. There are no significant lags shown on the PACF plot shown in figure 20 above. Therefore, the model will be updated with a seasonal MA order of 1.

```
# Model 2 SAMRA(0,1) with first seasonal differencing
m2.ts<-Plot_Model(Data_ts_agg,2,"21",0,0,0,0,1,1)
```

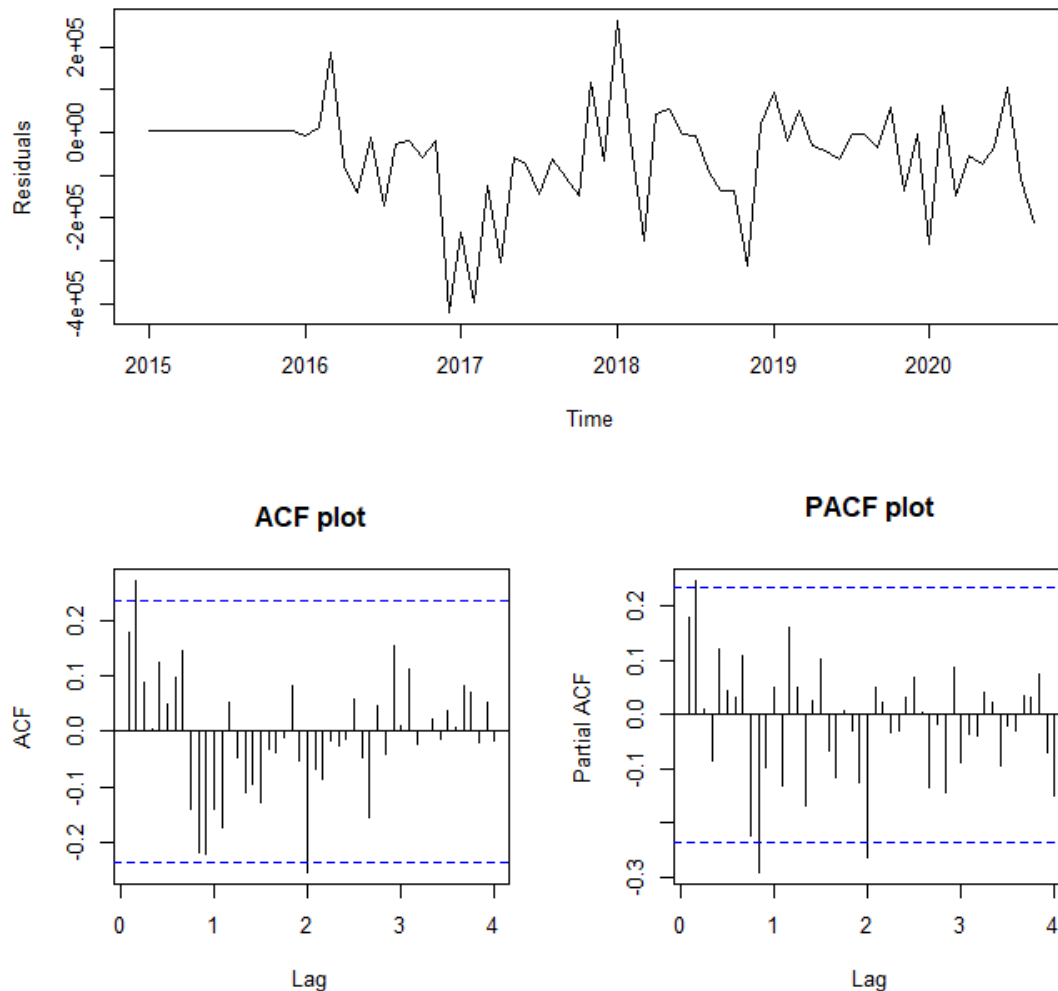
Figure 21: model 2 residuals



The residual plot of model two can be seen in figure 21 above, similarly to model one it can be seen there is no obvious seasonal component in the residuals. There is however still a significant lag on the first lag in the ACF shown in figure 21 above therefore for Model 3, Q will be updated to two and checked if there is still a significant lag.

```
# Model 3 SAMRA(0,2) with first seasonal differencing
m3.ts<-Plot_Model(Data_ts_agg,3,"22",0,0,0,0,1,2)
```

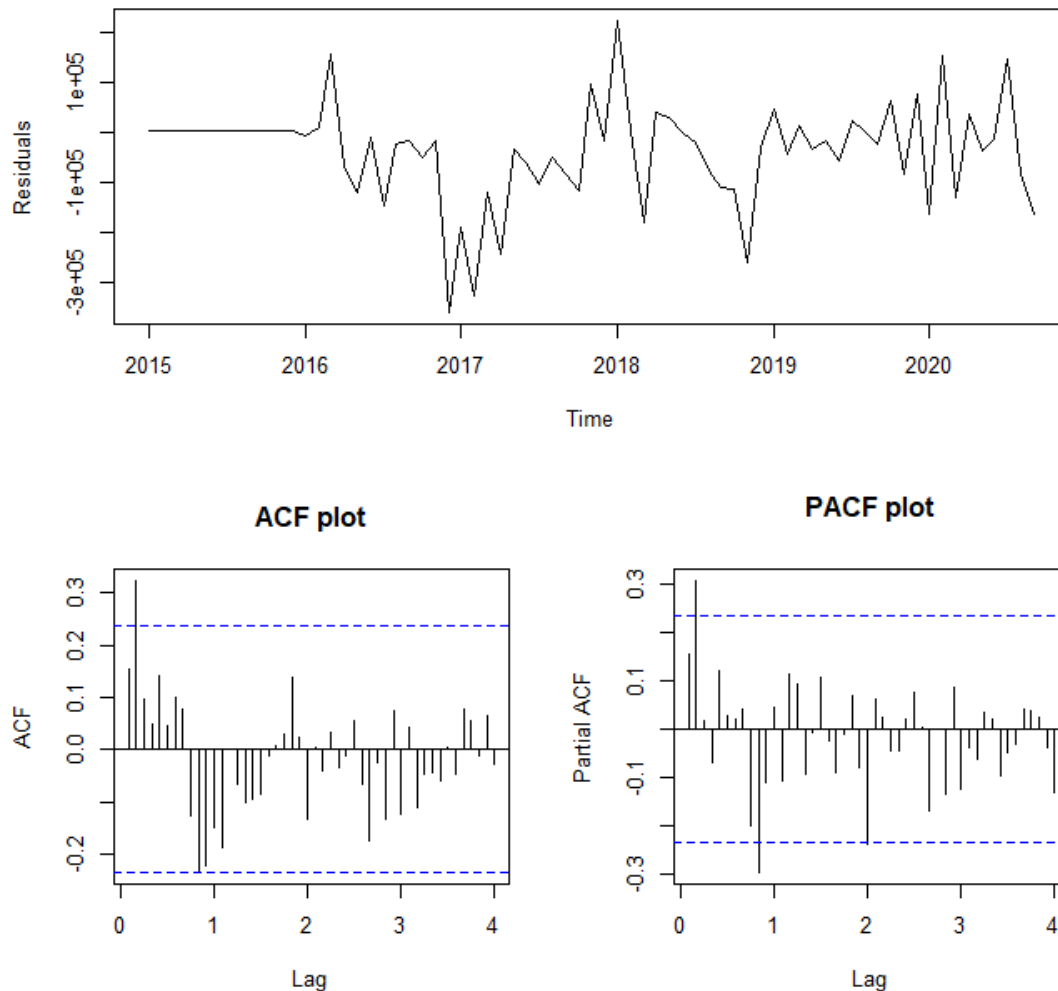
Figure 22: model 3 residuals



The residual plot of model two can be seen in figure 22 above, similarly to model two it can be seen there is no obvious seasonal component in the residuals. There is however still a significant lag on the second lag in the ACF shown in figure 22 above as well as the PACF shown in figure 22 above. Therefore, for Model 3, Q will be updated to three for model four and checked if there is still a significant lag.

```
# Model 4 SAMRA(0,3) with first seasonal differencing
m4.ts<-Plot_Model(Data_ts_agg,4,"23",0,0,0,0,1,3)
```

Figure 23: model 4 residuals



The residual plot of model four can be seen in figure 23 above, similarly to model three it can be seen there is no obvious seasonal component in the residuals. There are also no significant lags in the PACF or ACF plots shown above and therefore  $P = 0$  and  $Q = 3$  can be considered to model the seasonal component of the monthly aggregated time series.

```
Stationary_tests(residuals(m4.ts))
```

```
##
## Augmented Dickey-Fuller Test
##
## data: data
## Dickey-Fuller = -2.6091, Lag order = 4, p-value = 0.3276
## alternative hypothesis: stationary
##
##
## Phillips-Perron Unit Root Test
##
## data: data
## Dickey-Fuller Z(alpha) = -65.459, Truncation lag parameter = 3, p-value
## = 0.01
## alternative hypothesis: stationary
```

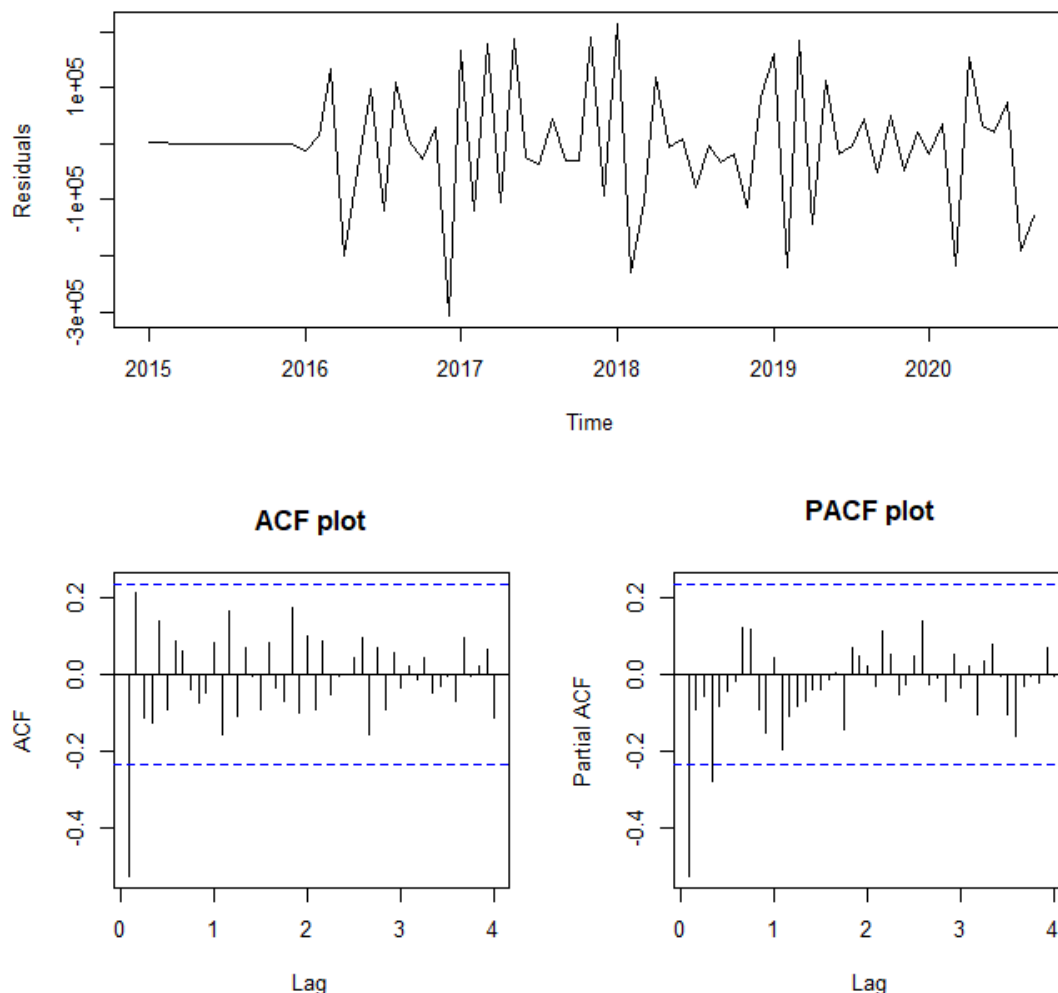
```
##
##
## KPSS Test for Level Stationarity
##
## data: data
## KPSS Level = 0.13522, Truncation lag parameter = 3, p-value = 0.1
```

For the Dickey-Fuller Test above, a p-value of 0.32 is greater than 0.05 indicates that under the 95% confidence interval the null hypothesis of non-stationarity can be rejected and there is statistically significant stationarity in the data. This also shown for the pp test value of 0.01 which is less than 0.05 and therefore also shows the time series is stationary.

The KPSS test, however, does not show stationarity as the p-value of 0.1 is greater than 0.05.

```
# Base model for BIC and eacf
base.ts<-Plot_Model(Data_ts_agg,1,"24",0,1,0,0,1,3)
```

**Figure 24: model 1 residuals**



```
adf.test(residuals(base.ts))
```

```
##
## Augmented Dickey-Fuller Test
##
```



```
## data: residuals(base.ts)
## Dickey-Fuller = -5.2585, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

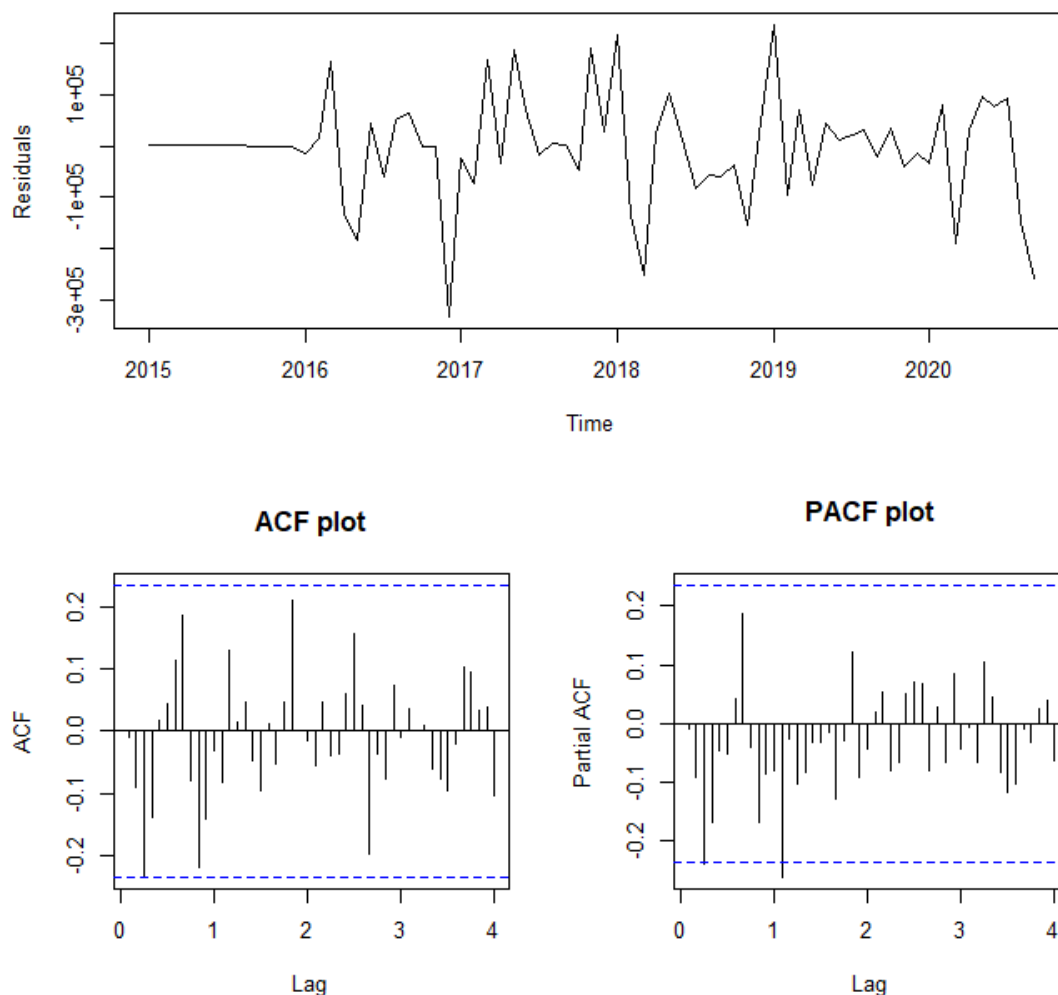
To find initial p and q values, the ACF and PACF plots will be analysed. A first difference will be used as per the ADF test above with a p-value of 0.01, the differenced series can be considered stationary. Looking at the lags in the ACF and PACF plots in figure 24 above, there is one significant lag in ACF and two significant lags in PACF. There could be a decreasing pattern.

Therefore, possible p and q values for the ARIMA section can be 2 and 1 respectively.

so, we will try to use an ARIMA (2,1) model.

```
# Model 5 ARIMA(2,1)+ SAMRA(0,3) with first seasonal differencing
m5.ts<-Plot_Model(Data_ts_agg,5,"25",2,1,1,0,1,3)
```

Figure 25: model 5 residuals



As can be seen in the ACF and PACF in figure 25 above,  $p = 2$  and  $q = 1$  reduced significant lag in both the ACF and PACF plot. There are also no obvious trends or seasonality in the plot of residuals.

```
# GET EACF
res.base = residuals(base.ts)
eacf(res.base)
```

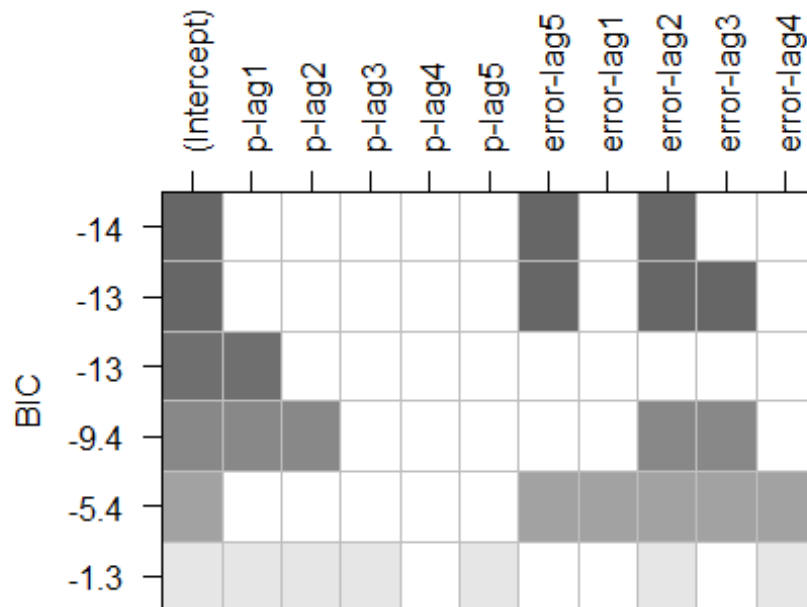
```
## AR/MA
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o o o o o o o o o o o
## 1 o o o o o o o o o o o o o o
## 2 x o x o o o o o o o o o o o
## 3 x x x x o o o o o o o o o
## 4 x o o o o o o o o o o o o
## 5 x o o x o o o o o o o o o
## 6 x o o o o o o o o o o o o
## 7 o x o o o o o o o o o x o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

The tentative models found using EACF table above, with the point (0,1) being the top leftmost 'o' are specified as:

- SARIMA(0,1,2) x(0,1,3)\_12
- SARIMA(0,1,1)x(0,1,3)\_12
- SARIMA(1,1,1)x(0,1,3)\_12
- SARIMA(1,1,2)x(0,1,3)\_12
- SARIMA(2,1,1)x(0,1,3)\_12

```
# BIC
par(mfrow=c(1,1))
bic_table = armasubsets(y=res.base, nar=5, nma=5, y.name='p', ar.method='ols')

## Reordering variables and trying again:
plot(bic_table)
```



The tentative models found using the BIC table are specified as:

- SARIMA(0,1,5)x(0,1,3)\_12
- SARIMA(0,1,2)x(0,1,3)\_12
- SARIMA(0,1,3)x(0,1,3)\_12

The TOTAL SET of identified possible models is:

- SARIMA(0,1,1)x(0,1,3)\_12
- SARIMA(1,1,1)x(0,1,3)\_12
- SARIMA(1,1,2)x(0,1,3)\_12
- SARIMA(2,1,1)x(0,1,3)\_12
- SARIMA(0,1,5)x(0,1,3)\_12
- SARIMA(0,1,2)x(0,1,3)\_12
- SARIMA(0,1,3)x(0,1,3)\_12

## 6.2 Model Fitting and Diagnostic Checking

### 6.2.1 Parameter significance

*SARIMA(0, 1, 1)(0, 1, 3)\_12*

```
# SARIMA(0,1,1)x(0,1,3)_12
```

```
# ML METHOD
```

```
m4_011_ML <- Parameter.estimation(Data_ts_agg,c(0,1,1),c(0,1,3),method="ML", '26')
```

```
##
```

```
## Shapiro-Wilk normality test
```

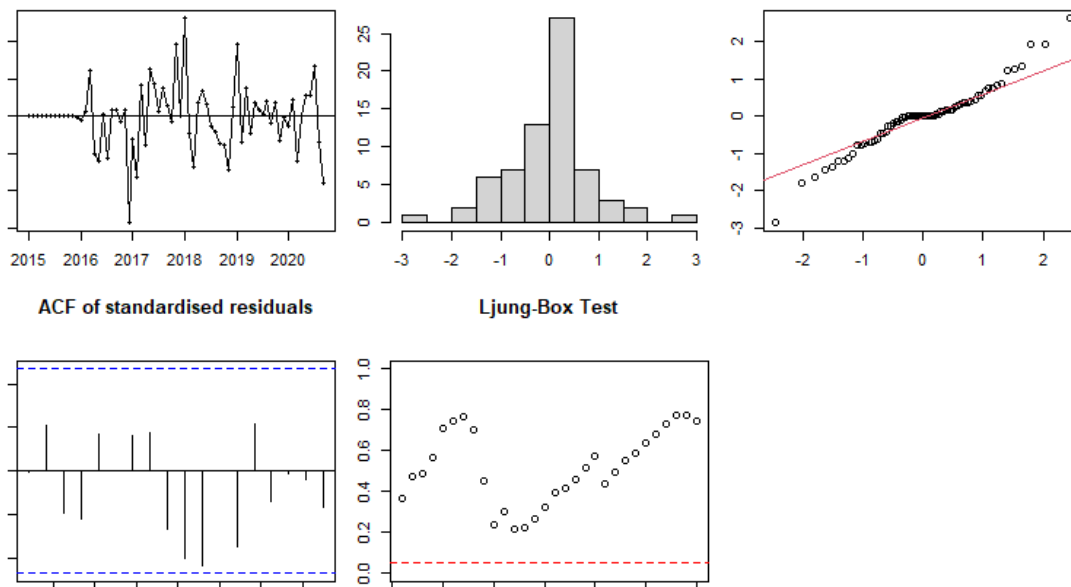
```
##
```

```
## data: res.model
```

```
## W = 0.95234, p-value = 0.01036
```

Figure 26: SARIMA(0,1,1)(0,1,3) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
```

```
## z test of coefficients:
```

```
##
```

```
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.792971   0.164764 -4.8128 1.489e-06 ***
## sma1  -1.000633   0.847262 -1.1810  0.2376
## sma2  -0.016741   0.272201 -0.0615  0.9510
## sma3   0.018542   0.249494  0.0743  0.9408
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model that the MA(1) value is significant, however not the seasonal P and Q values. Additionally in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.01 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. Because the residuals are not normal, CSS method will also be used to confirm the trends seen through the ML method.

```
# CSS METHOD
```

```
m4_011_CSS <- Parameter.estimation(Data_ts_agg,c(0,1,1),c(0,1,3),method="CSS", '27')
```

```
##
```

```
## Shapiro-Wilk normality test
```

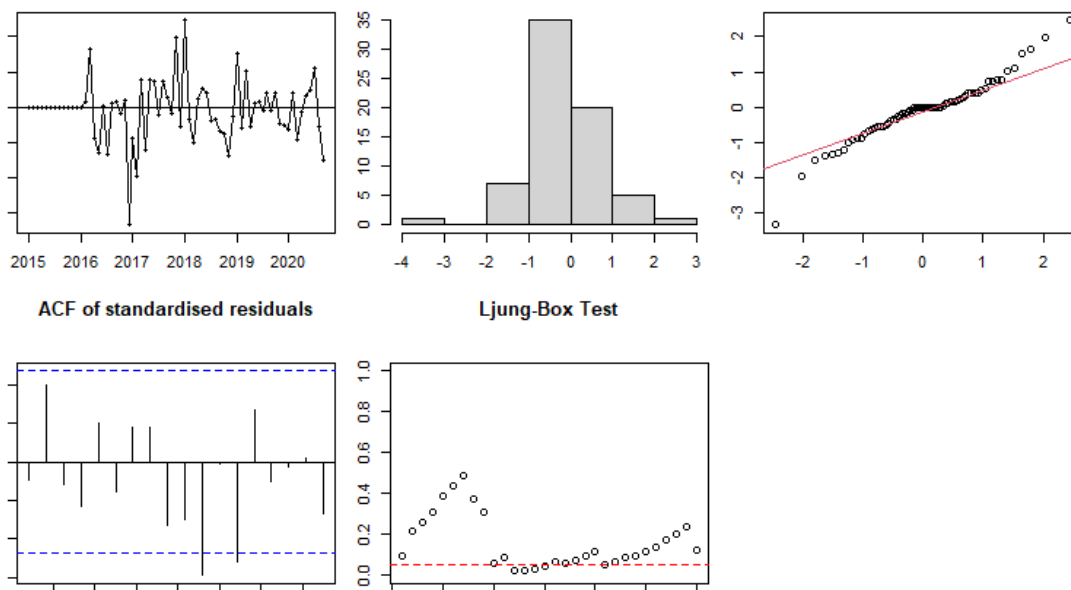
```
##
```

```
## data: res.model
```

```
## W = 0.94238, p-value = 0.00321
```

Figure 27: SARIMA(0,1,1)(0,1,3) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
```

```
## z test of coefficients:
```

```
##
```

```
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.830103   0.119658 -6.9373 3.996e-12 ***
## sma1  -0.676582   0.155618 -4.3477 1.376e-05 ***
## sma2  -0.013952   0.220682 -0.0632  0.9496
## sma3  -0.041265   0.209363 -0.1971  0.8438
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen from the above model that the MA (1) value is significant as well as the seasonal MA value. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be

further seen by the Shapiro Wilks t-test which returned a p-value of 0.01 which is less than 0.003. There are also significant lags in the ACF plot, and all p values are not above the confidence interval in the Ljung-Box test of 0.05 suggesting there is still a trend not covered by the model.

It is evident that the seasonal component is not being captured as none of the SIMRA parameters are significant. Therefore, the P and Q values will be altered using over fitting to see which values are optimal.

### Trail 1: reducing Q value by 1

# ML METHOD

```
m4_011_ML <- Parameter.estimation(Data_ts_agg,c(0,1,1),c(0,1,2),method="ML", '28')
```

##

## Shapiro-Wilk normality test

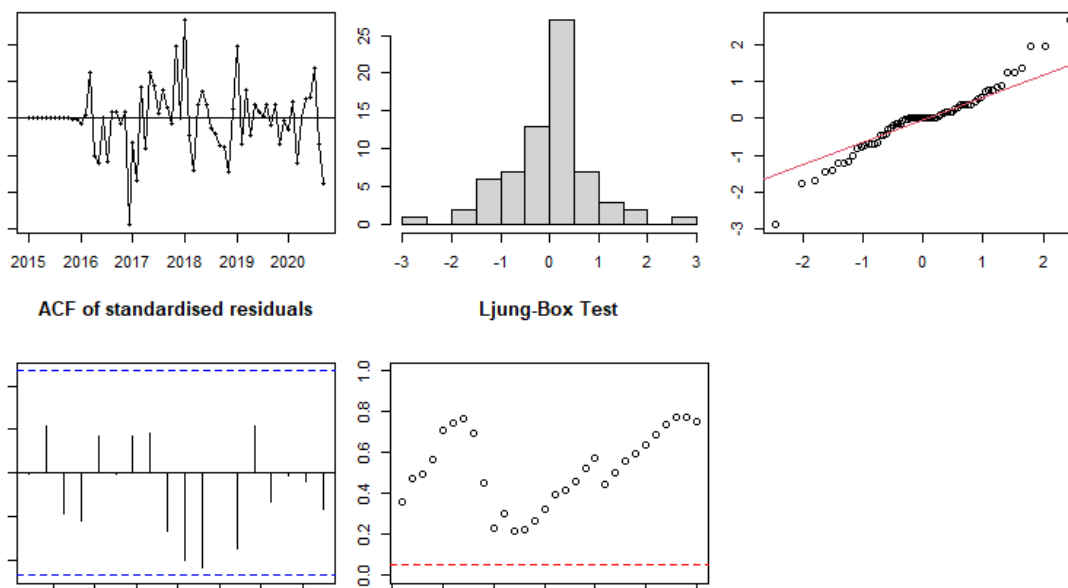
##

## data: res.model

## W = 0.9521, p-value = 0.01007

Figure 28: SARIMA(0,1,1)(0,1,2) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



##

## z test of coefficients:

##

	Estimate	Std. Error	z value	Pr(> z )
ma1	-0.7974117	0.1535891	-5.1919	2.082e-07 ***
sma1	-0.9976695	0.7282187	-1.3700	0.1707
sma2	-0.0022906	0.1900177	-0.0121	0.9904

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

It can be seen from the above model that the MA (1) value is significant, however not the seasonal P and Q values. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.01 which is less than 0.05. However,

there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test.

**Trail 2:** reducing  $Q$  value by two.

# ML METHOD

```
m4_011_ML <- Parameter. estimation(Data_ts_agg, c(0,1,1), c(0,1,1), method="ML", '29')
```

```
##
```

```
## Shapiro-Wilk normality test
```

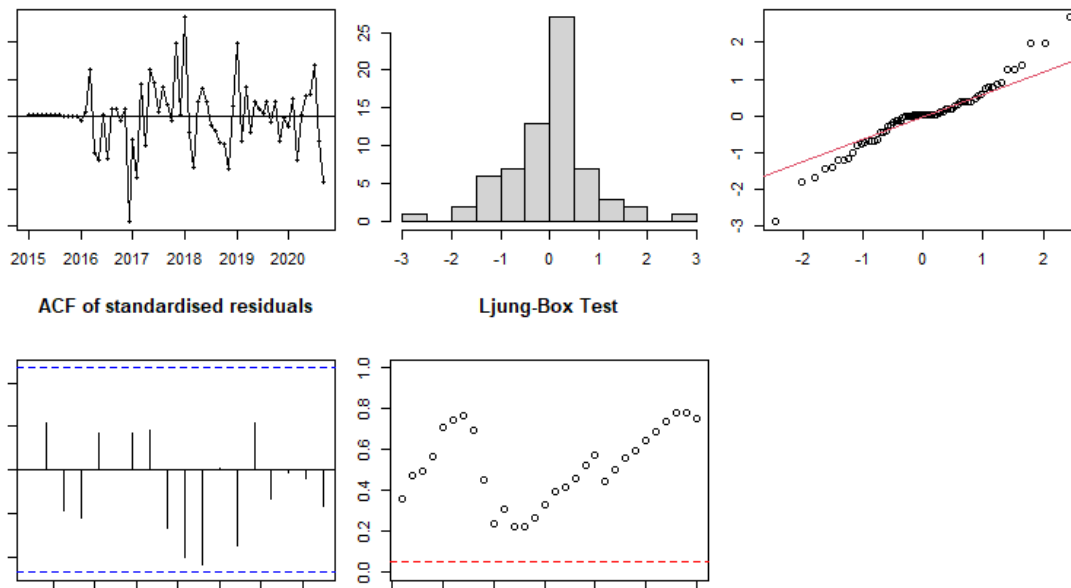
```
##
```

```
## data: res.model
```

```
## W = 0.95199, p-value = 0.009934
```

Figure 29: SARIMA(0,1,1)(0,1,1) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
```

```
## z test of coefficients:
```

```
##
```

```
## Estimate Std. Error z value Pr(>|z|)
```

```
## ma1 -0.79829 0.13531 -5.8997 3.643e-09 ***
```

```
## sma1 -0.99918 0.72796 -1.3726 0.1699
```

```
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model the MA(1) value is significant, however not the seasonal Q value. Due to this, continual testing will be conducted until seasonal significance is found.

**Trail 3:** reducing  $Q$  value by three and increasing  $P$ .

# SARIMA(0,1,1)x(0,1,3)\_12

# ML METHOD

```
m4_011_ML <- Parameter. estimation(Data_ts_agg, c(0,1,1), c(1,1,0), method="ML", '30')
```

```
##
```

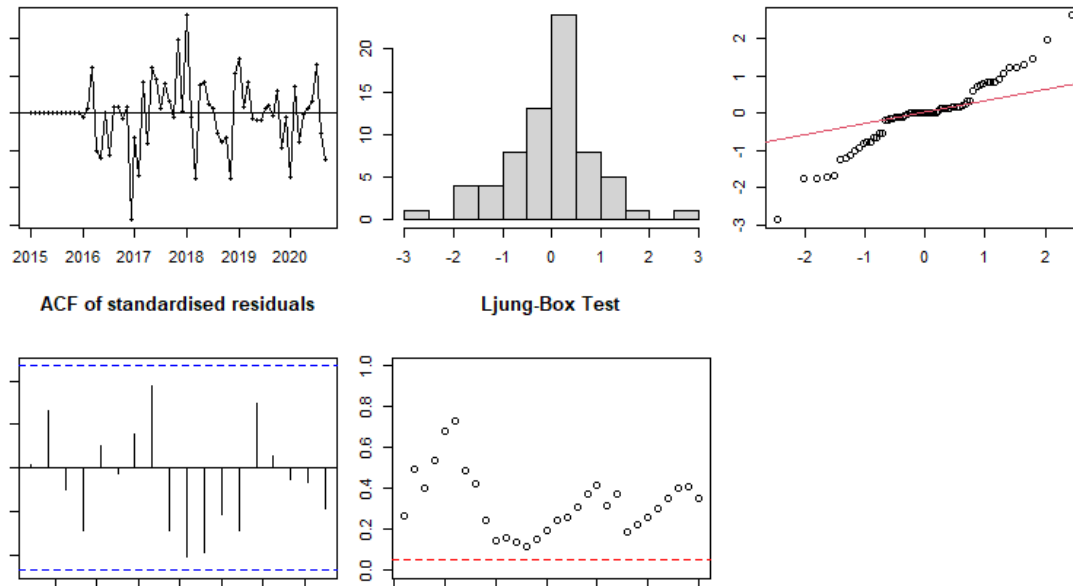
```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: res.model
## W = 0.94454, p-value = 0.004114
```

Figure 30: SARIMA(0,1,1)(1,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.81526    0.14323 -5.6920 1.255e-08 ***
## sar1  -0.50536    0.12883 -3.9227 8.757e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen from the above model the MA(1) value is significant and so is the seasonal AR(1) value. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.004 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test.

Now that the optimal Q value has been found, further testing will be conducted by adjusting P.

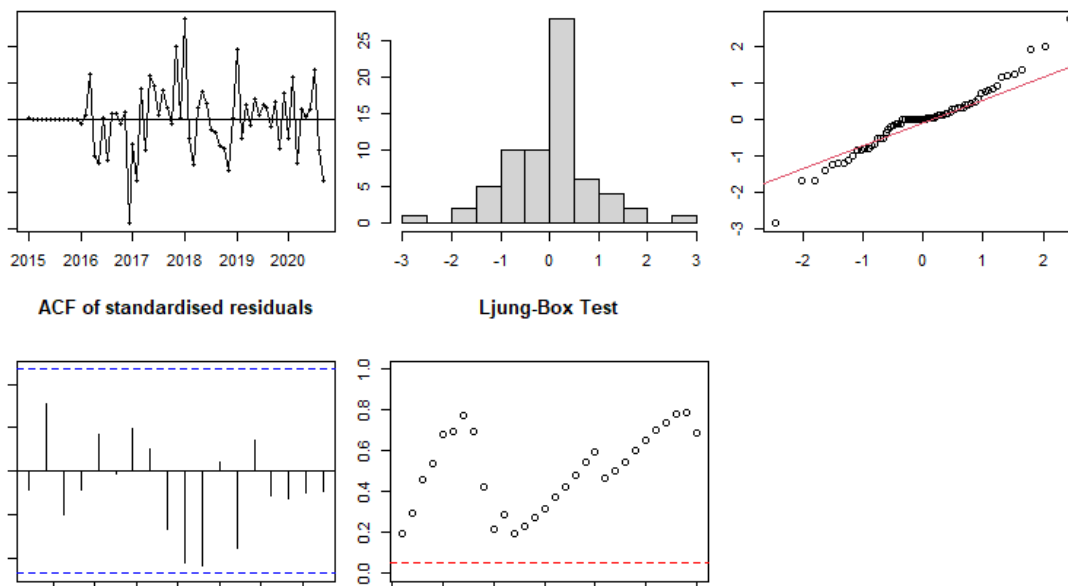
#### Trail 4: Increasing P by one

```
# SARIMA(0,1,1)x(0,1,3)_12
# ML METHOD
m4_011_ML <- Parameter.estimation(Data_ts_agg,c(0,1,1),c(2,1,0),method="ML", '31')

##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.95544, p-value = 0.01511
```

Figure 31: SARIMA(0,1,1)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.81680    0.15595 -5.2374 1.628e-07 ***
## sar1  -0.71999    0.15954 -4.5128 6.397e-06 ***
## sar2  -0.41048    0.14444 -2.8419 0.004485 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

base_011_ML <- m4_011_ML
```

It can be seen from the above model the MA(1) value is significant and so is the seasonal AR(1) and AR(2) values. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.015, which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. Due to the higher number of significant seasonal parameters in this model compared to previous, it is more likely to capture the seasonal trend to a greater degree. This is the most optimal model for increasing any further causes and reduction in the number of significant parameters. The updated seasonal parameters will be used now to check the ARIMA p and q values for previously identified possible models.

Final seasonal parameters:

- P = 2
- D = 1
- Q = 0

It can also be seen from the above tested model that the p and q values of 0 and 1 respectively, show allot of promise due to all parameters being significant.



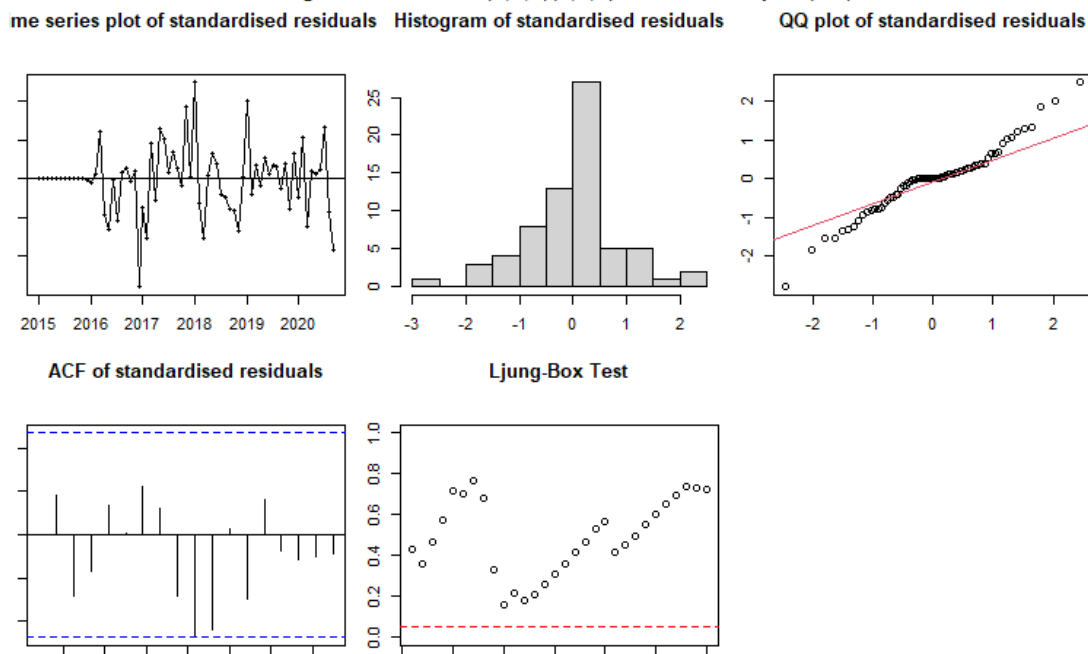
```

SARIMA(1, 1, 1)(2, 1, 0)_12
# SARIMA(1,1,1)x(2,1,0)_12
# ML METHOD
base_111_ML <- Parameter.estimation(Data_ts_agg,c(1,1,1),c(2,1,0),method="ML", '32')

##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.95747, p-value = 0.01942

```

Figure 32: SARIMA(1,1,1)(2,1,0) Residual Analysis (ML)



```

##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  -0.11184    0.30303  -0.3691  0.71209
## ma1  -0.72021    0.29987  -2.4017  0.01632 *
## sar1  -0.70556    0.17030  -4.1429 3.429e-05 ***
## sar2  -0.42200    0.15851  -2.6624  0.00776 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

It can be seen for the above model none of the AR or MA values for the ARIMA component are significant. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.02 which is less than 0.05. However, there is no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will also be used because the residuals are not normally distributed.

```

# CSS METHOD
base_111_CSS <- Parameter.estimation(Data_ts_agg,c(1,1,1),c(2,1,0),method="CSS", '33')

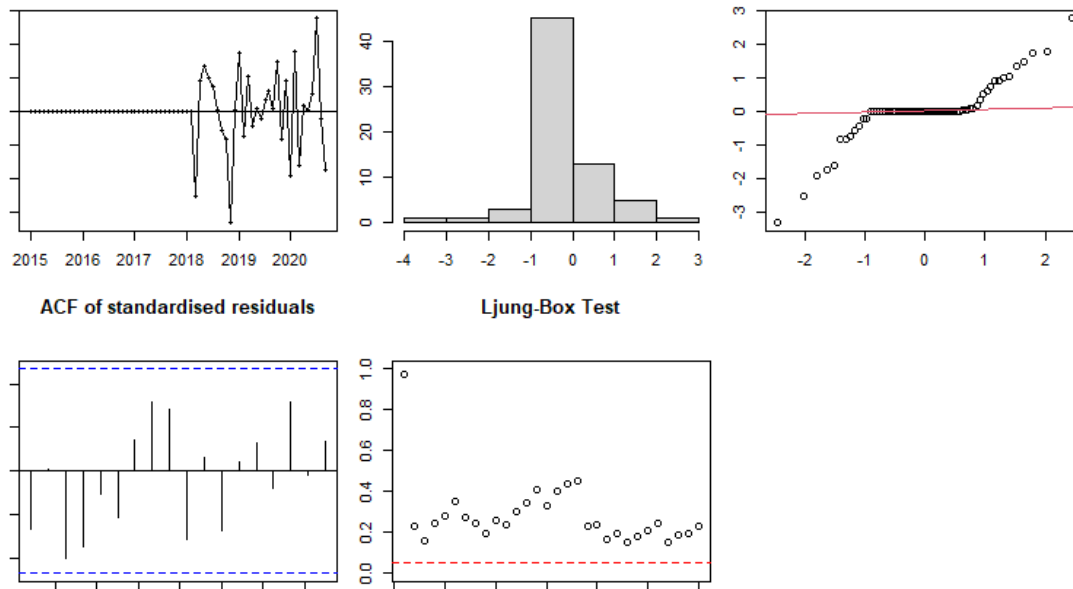
##
## Shapiro-Wilk normality test

```

```
##
## data: res.model
## W = 0.78084, p-value = 9.102e-09
```

Figure 33: SARIMA(1,1,1)(2,1,0) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.0875129  0.0023623   37.046 < 2.2e-16 ***
## ma1  -1.2745978  0.0117095  -108.852 < 2.2e-16 ***
## sar1 -0.5545065  0.0025092  -220.991 < 2.2e-16 ***
## sar2 -0.1613516  0.0087511   -18.438 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

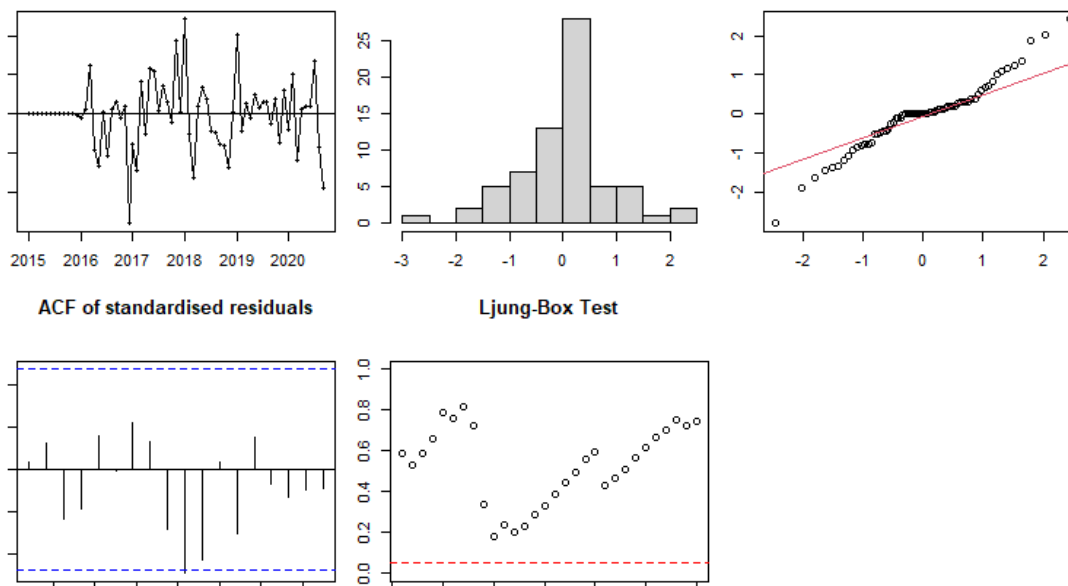
It can be seen for the above model using the CSS method, All the AR and MA values for the ARIMA component are significant. Additionally, in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.02 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. Therefore, this is likely to be a promising model.

SARIMA (1, 1, 2)(2, 1, 0)<sub>12</sub>

```
# ML METHOD
base_112_ML <- Parameter.estimation(Data_ts_agg, c(1, 1, 2), c(2, 1, 0), method="ML", '34')
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.95424, p-value = 0.01305
```

Figure 34: SARIMA(1,1,2)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



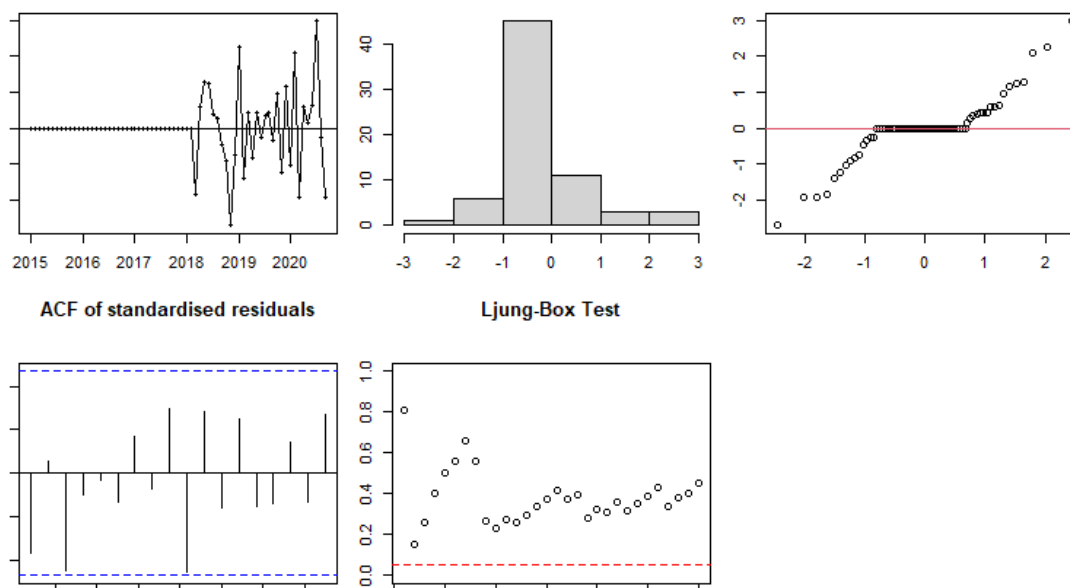
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1   -0.52569    0.50196  -1.0473  0.294977
## ma1   -0.32004    0.51600  -0.6202  0.535107
## ma2   -0.30295    0.44495  -0.6809  0.495960
## sar1  -0.70596    0.16101  -4.3847 1.162e-05 ***
## sar2  -0.42545    0.14098  -3.0179  0.002546 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model none of the AR or MA values for the ARIMA component are significant. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.013 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will also be used because the residuals are not normally distributed.

```
# CSS METHOD
base_112_CSS <- Parameter. estimation(Data_ts_agg, c(1, 1, 2), c(2, 1, 0), method = "CSS", '35')
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.82263, p-value = 1.18e-07
```

Figure 35: SARIMA(1,1,2)(2,1,0) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value Pr(>|z|)
## ar1   0.0035225  0.0203324   0.1732  0.8625
## ma1  -1.1662376  0.1202955  -9.6948 <2e-16 ***
## ma2  -0.1153427  0.1619569  -0.7122  0.4764
## sar1 -0.6806481  0.0209258 -32.5268 <2e-16 ***
## sar2 -0.3865264  0.0212518 -18.1879 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model using the CSS method, All of the AR and MA values for the ARIMA component are significant. Using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.02 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test.

```
SARIMA(2, 1, 1)(2, 1, 0)_I2
```

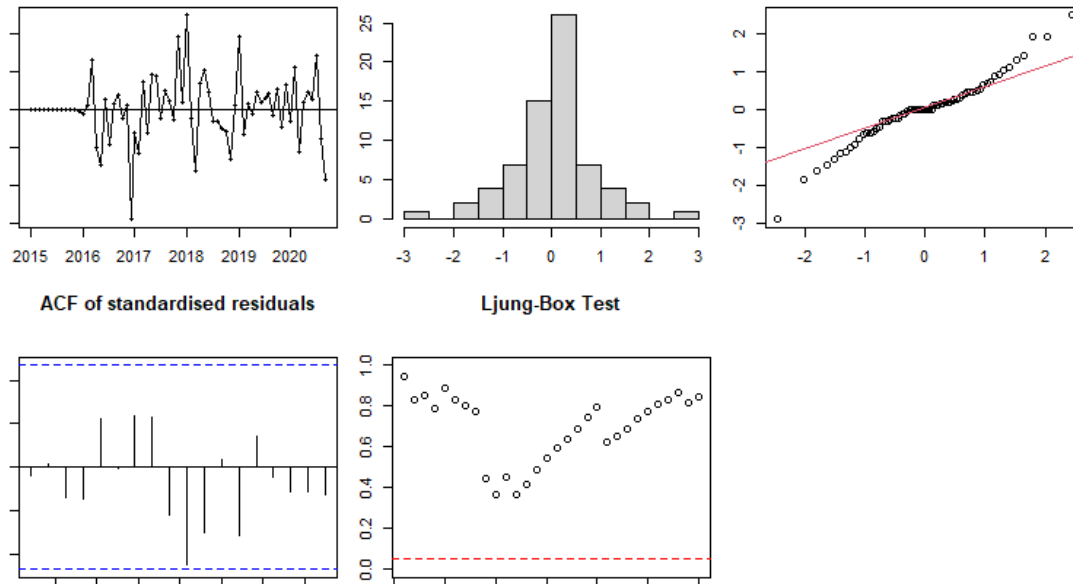
```
# ML METHOD
```

```
base_211_ML <- Parameter.estimation(Data_ts_agg,c(2,1,1),c(2,1,0),method="ML", '36')
```

```
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.95635, p-value = 0.01692
```

Figure 36: SARIMA(2,1,1)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



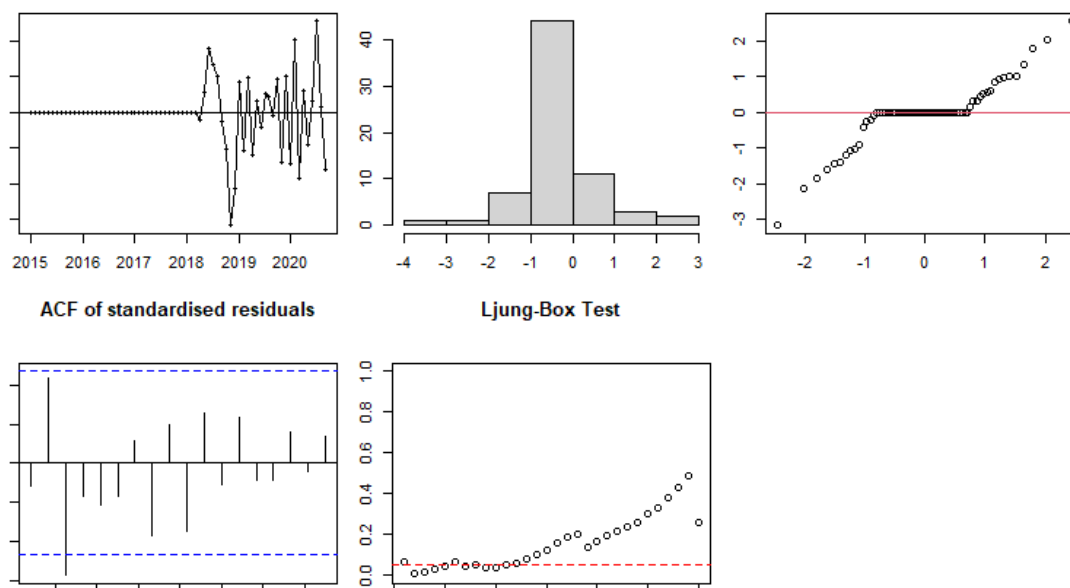
```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value  Pr(>|z|)
## ar1    0.116663   0.139368   0.8371   0.402544
## ar2    0.276976   0.137889   2.0087   0.044570 *
## ma1   -0.999867   0.083334  -11.9984 < 2.2e-16 ***
## sar1  -0.641745   0.150041   -4.2771  1.893e-05 ***
## sar2  -0.405200   0.139817   -2.8981   0.003755 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen from the above model that none of the AR values and the MA(1) value for the ARIMA component are significant. It is also shown in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.012 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will be used because the residuals are not normally distributed.

```
# CSS METHOD
base_211_CSS <- Parameter.estimation(Data_ts_agg,c(2,1,1),c(2,1,0),method="CSS", '37')
##
## Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.82392, p-value = 1.284e-07
```

Figure 37: SARIMA(2,1,1)(2,1,0) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.223200   0.077451  -2.8818  0.003954 **
## ar2  -0.466185   0.037515 -12.4266 < 2.2e-16 ***
## ma1  -1.250171   0.021429 -58.3394 < 2.2e-16 ***
## sar1 -0.728645   0.021386 -34.0707 < 2.2e-16 ***
## sar2 -0.405981   0.018689 -21.7233 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model using the CSS method, All of the AR and MA values for the ARIMA component are significant. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.004 which is less than 0.05. There is also a significant lag in the ACF plot and not all p values are above the confidence interval in the Ljung-Box test suggesting this model is not acceptably capturing all trends in the time series.

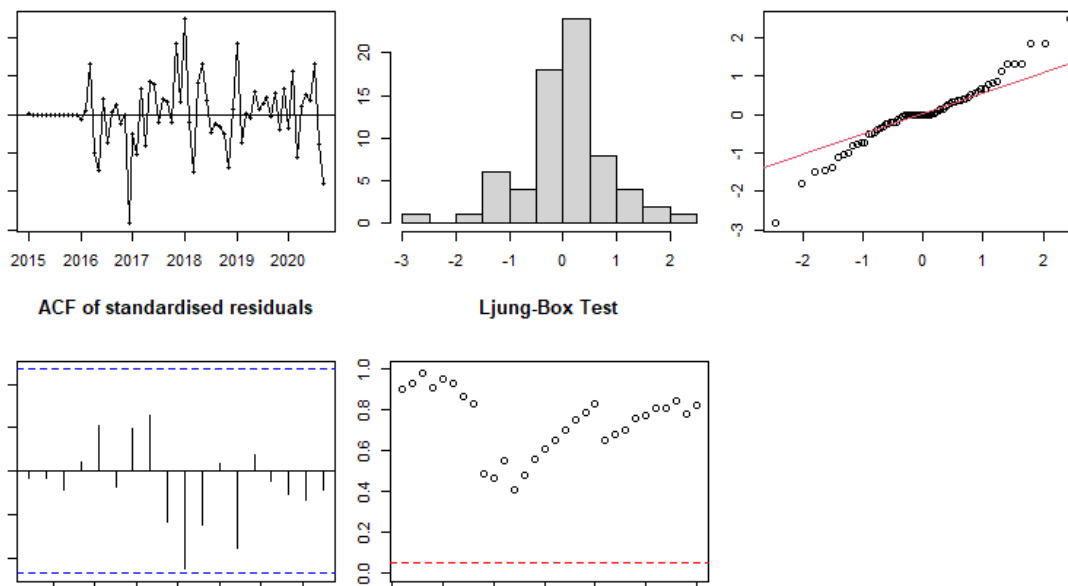
*SARIMA(0, 1, 5)(2, 1, 0)<sub>12</sub>*

```
base_015_ML <- Parameter. estimation(Data_ts_agg, c(0, 1, 5), c(2, 1, 0), method = "ML", '38')
```

```
##
## Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95543, p-value = 0.0151
```

Figure 38: SARIMA(0,1,5)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



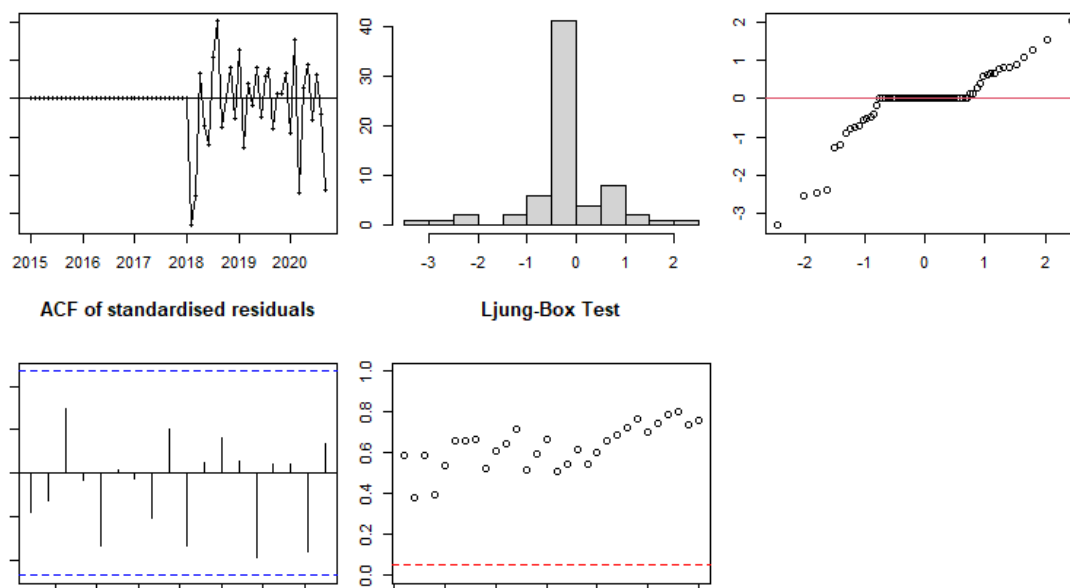
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.8730969  0.1743786 -5.0069 5.531e-07 ***
## ma2   0.1954657  0.1936328  1.0095 0.312751
## ma3  -0.2699945  0.2430327 -1.1109 0.266595
## ma4  -0.0582800  0.2360817 -0.2469 0.805014
## ma5   0.0060052  0.1250764  0.0480 0.961707
## sar1  -0.6187762  0.1588614 -3.8951 9.817e-05 ***
## sar2  -0.3873829  0.1445589 -2.6798 0.007368 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model the MA(1) value for the ARIMA component is significant and none of the others. Additionally, in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.015, which is less than 0.05. However, there is no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will also be used because the residuals are not normally distributed.

```
base_015_CSS <- Parameter. estimation(Data_ts_agg, c(0,1,5), c(2,1,0), method="CSS", '39')
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.7918, p-value = 1.726e-08
```

Figure 39: SARIMA(0,1,5)(2,1,0) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.50653    0.15146 -3.3443 0.0008249 ***
## ma2   0.25607    0.17834  1.4358 0.1510465
## ma3  -0.57897    0.15413 -3.7563 0.0001725 ***
## ma4   0.17350    0.16677  1.0403 0.2981929
## ma5   0.59534    0.14348  4.1493 3.335e-05 ***
## sar1 -0.63321    0.10866 -5.8275 5.628e-09 ***
## sar2 -0.61838    0.13624 -4.5391 5.651e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model using the CSS method, only the MA(1) and MA(3) parameters for the ARIMA component are significant. Additionally in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of ~0 which is less than 0.05. However, there is no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test.

SARIMA(0, 1, 2)(2, 1, 0)<sub>I2</sub>

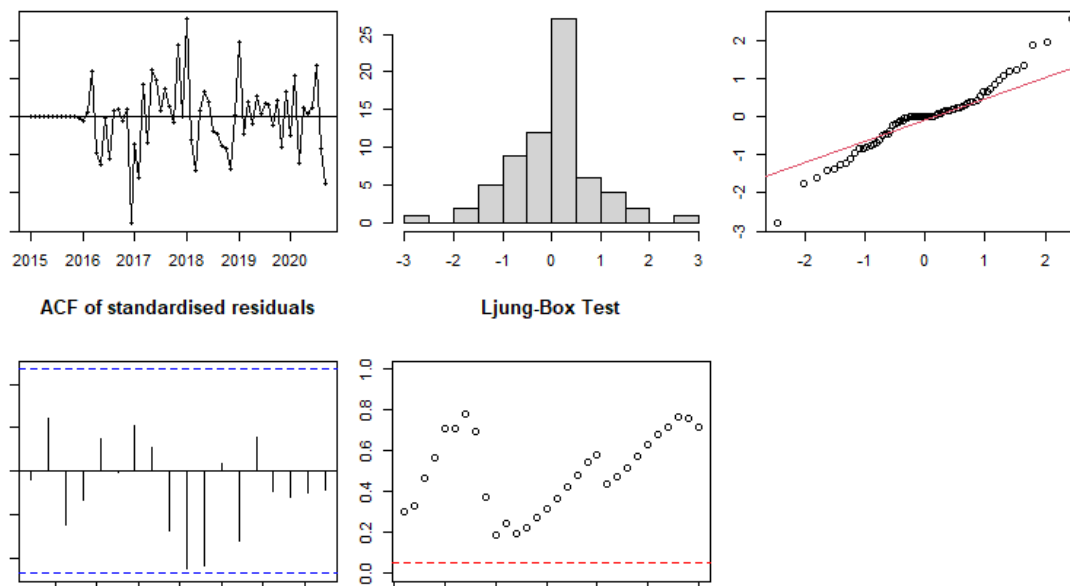
```
base_012_ML <- Parameter.estimation(Data_ts_agg, c(0, 1, 2), c(2, 1, 0), method="ML", '40')
```

```
##
## Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95775, p-value = 0.02012
```



Figure 40: SARIMA(0,1,2)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



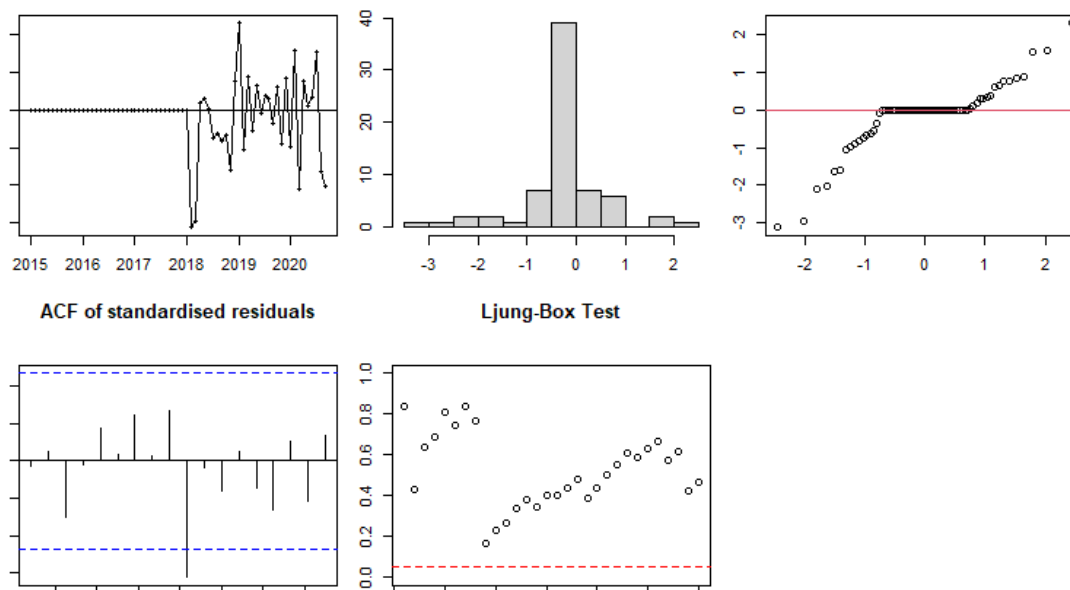
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.821572   0.149305 -5.5026 3.741e-08 ***
## ma2   0.048129   0.150831  0.3191 0.749658
## sar1 -0.712262   0.162707 -4.3776 1.200e-05 ***
## sar2 -0.416974   0.144295 -2.8897 0.003856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model that the MA(1) value for the ARIMA component is significant and none of the others. Additionally, in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.02 which is less than 0.05. However, there is no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will be used because the residuals are not normally distributed.

```
base_012_CSS <- Parameter.estimation(Data_ts_agg, c(0,1,2),c(2,1,0),method="CSS", '41')
##
## Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.80252, p-value = 3.298e-08
```

Figure 41: SARIMA(0,1,2)(2,1,0) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.516560   0.136887 -3.7736 0.0001609 ***
## ma2  -0.024451   0.139904 -0.1748 0.8612620
## sar1  -0.597292   0.136534 -4.3747 1.216e-05 ***
## sar2  -0.379309   0.118047 -3.2132 0.0013126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model using the CSS method, only the MA(1) parameter for the ARIMA component are significant. Additionally in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of ~0 which is less than 0.05. There is also a significant lag in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. This is likely a less promising model.

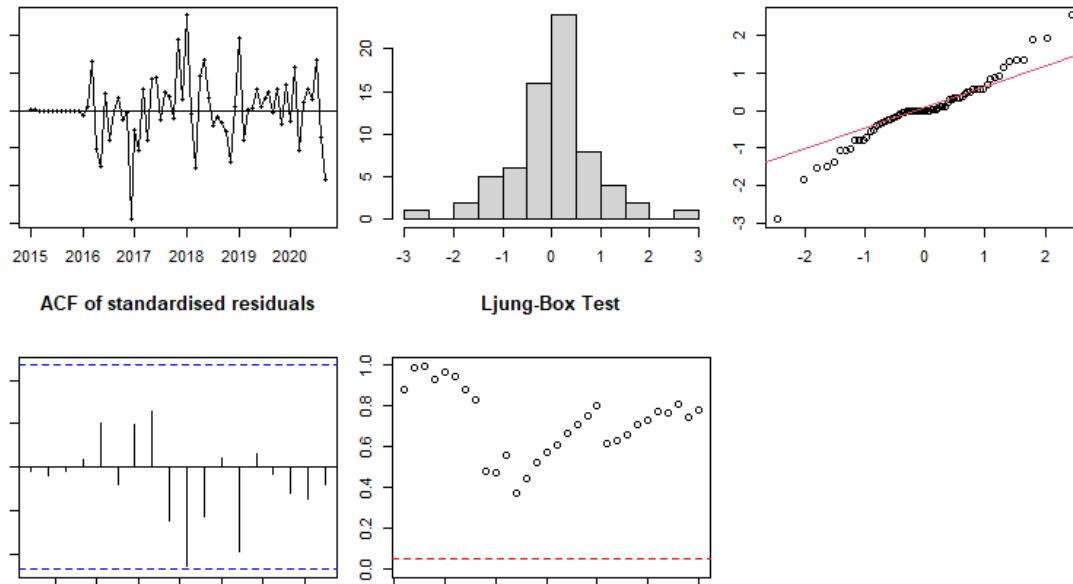
```
SARIMA(0, 1, 3)(2, 1, 0)_I2
```

```
base_013_ML <- Parameter. estimation(Data_ts_agg, c(0, 1, 3), c(2, 1, 0), method = "ML", '42')
```

```
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.9533, p-value = 0.01164
```

Figure 42: SARIMA(0,1,3)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



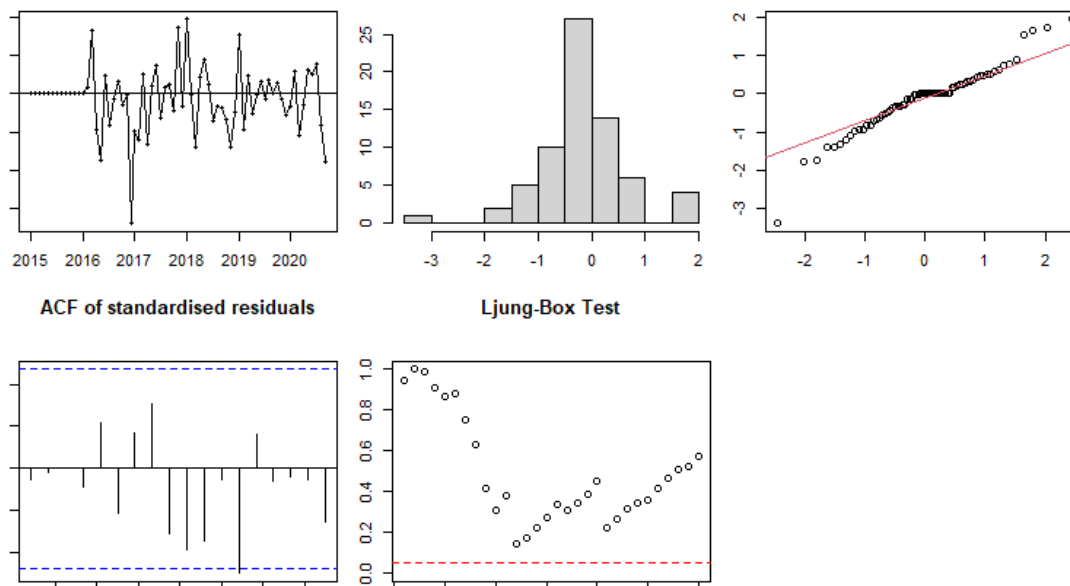
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.88581    0.17577 -5.0397 4.663e-07 ***
## ma2   0.20971    0.17038  1.2308 0.218392
## ma3  -0.32389    0.13919 -2.3270 0.019965 *
## sar1 -0.63225    0.15045 -4.2023 2.642e-05 ***
## sar2 -0.39644    0.13844 -2.8637 0.004188 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model that the MA(1) value for the ARIMA component is significant and none of the others. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.01 which is less than 0.05. However, there are no significant lags in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test. The CSS method will be used because the residuals are not normally distributed.

```
base_013_CSS <- Parameter.estimation(Data_ts_agg,c(0,1,3),c(0,1,3),method="CSS", '43')
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.93896, p-value = 0.002181
```

Figure 43: SARIMA(0,1,3)(0,1,3) Residual Analysis (CSS)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.83184    0.14559 -5.7136 1.106e-08 ***
## ma2   0.26912    0.16276  1.6534 0.0982429 .
## ma3  -0.32952    0.13260 -2.4851 0.0129501 *
## sma1 -0.61233    0.16467 -3.7186 0.0002004 ***
## sma2 -0.16972    0.23594 -0.7193 0.4719408
## sma3  0.04330    0.20938  0.2068 0.8361630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen for the above model using the CSS method, only the MA(1) parameter for the ARIMA component are significant. It can be seen in the residual analysis, there is no obvious trend in the residuals plot. However, using the histogram of residuals and the QQ plot, it does not seem the residuals are normally distributed. This can be further seen by the Shapiro Wilks t-test which returned a p-value of 0.02 which is less than 0.05. There is also a significant lag in the ACF plot, and all p values are above the confidence interval in the Ljung-Box test.

Overall, from the cumulative residual analysis performed the most promising models are SARIMA(1, 1, 1)(2, 1, 0)<sub>12</sub> and SARIMA(0, 1, 1)(2, 1, 0)<sub>12</sub>. It is important to note however that the AR(1) parameter was not significant using CSS for the model SARIMA(1, 1, 1)(2, 1, 0)<sub>12</sub>.

## 6.2.2 BIC/AIC values

An alternative method to evaluate and compare the suitability of various SIMRA models is by utilising AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores. In both instances, lower values of AIC and BIC indicate better-fitting models.

AIC and BIC evaluate models based on their goodness of fitness and complexity. Both criteria consider the likelihood of the data fitting the model and penalise models for their complexity, effectively avoiding

overfitting. By striking a balance between fit and complexity, AIC and BIC provide a comprehensive assessment of model quality.

```
sc.AIC =
AIC(base_011_ML, base_111_ML, base_112_ML, base_211_ML, base_015_ML, base_012_ML, base_013_ML)

sc.BIC = BIC(base_011_ML, base_111_ML, base_112_ML, base_211_ML, base_015_ML, base_012_ML,
base_013_ML)

kable(sort.score(sc.AIC, score = "aic"))
```

	df	AIC
base_011_ML	4	1483.568
base_013_ML	6	1483.708
base_211_ML	6	1484.183
base_111_ML	5	1485.380
base_012_ML	5	1485.475
base_112_ML	6	1486.979
base_015_ML	8	1487.630

```
kable(sort.score(sc.BIC, score = "bic"))
```

	df	BIC
base_011_ML	4	1491.669
base_111_ML	5	1495.507
base_012_ML	5	1495.602
base_013_ML	6	1495.860
base_211_ML	6	1496.335
base_112_ML	6	1499.131
base_015_ML	8	1503.833

Based on the AIC and BIC tables mentioned, the 'base\_011\_ML' model attained the lowest AIC and BIC values, indicating that the SARIMA(0,1,1)x(0,1,3)<sub>12</sub> model is the best fitting model according to these scores, with the SARIMA(1,1,1)x(0,1,3)<sub>12</sub> model achieving the second best score. On the other hand, the SARIMA(0,1,5)x(0,1,3)<sub>12</sub> model exhibited the highest values, suggesting a poorer fit.

### 6.2.3 Accuracy Scores

In addition to AIC and BIC, another approach to comparing the fit or accuracy of models is by using various accuracy scores. These scores include metrics such as ME (Mean Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error), MASE (Mean Absolute Scaled Error), and ACF1 (Autocorrelation Function at lag 1).

Lower values for most of these accuracy metrics indicate a better fit in terms of the accuracy of the model. By assessing these metrics, one can gauge how well the model predicts or approximates the observed data. The selection of the model with the lowest values for these accuracy scores suggests its superior performance in accurately capturing the underlying patterns and trends in the data.

```

Sm4_011_ML <- accuracy(m4_011_ML)[1:7]
Sbase_111_ML <- accuracy(base_111_ML)[1:7]
Sbase_112_ML <- accuracy(base_112_ML)[1:7]
Sbase_211_ML <- accuracy(base_211_ML)[1:7]
Sbase_015_ML <- accuracy(base_015_ML)[1:7]
Sbase_012_ML <- accuracy(base_012_ML)[1:7]
Sbase_013_ML <- accuracy(base_013_ML)[1:7]

df.Smodels <- data.frame(
  rbind(Sm4_011_ML, Sbase_111_ML, Sbase_112_ML, Sbase_211_ML,
        Sbase_015_ML, Sbase_012_ML, Sbase_013_ML)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(0,1,1)x(2,1,0)_12", "SARIMA(1,1,1)x(2,1,0)_12",
                          "SARIMA(1,1,2)x(2,1,0)_12", "SARIMA(2,1,1)x(2,1,0)_12",
                          "SARIMA(0,1,5)x(2,1,0)_12", "SARIMA(0,1,2)x(2,1,0)_12",
                          "SARIMA(0,1,3)x(2,1,0)_12")
kable(round(df.Smodels, digits = 3))

```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
SARIMA(0,1,1)x(2,1,0)_12	-1558.413	105731.8	72165.75	-0.129	2.019	0.614	-0.042
SARIMA(1,1,1)x(2,1,0)_12	-2938.686	105795.6	72354.91	-0.166	2.020	0.616	0.001
SARIMA(1,1,2)x(2,1,0)_12	-2804.154	105329.9	71654.10	-0.162	1.999	0.610	0.016
SARIMA(2,1,1)x(2,1,0)_12	2249.273	100606.6	69026.11	-0.014	1.924	0.587	-0.021
SARIMA(0,1,5)x(2,1,0)_12	3228.639	100428.4	68468.06	0.015	1.911	0.583	-0.016
SARIMA(0,1,2)x(2,1,0)_12	-2543.015	105825.4	72425.96	-0.156	2.024	0.616	-0.020
SARIMA(0,1,3)x(2,1,0)_12	3870.899	100153.2	67915.61	0.033	1.895	0.578	-0.009

Based on the table provided, it is observed that the SARIMA(1,1,1)x(2,1,0)\_12 model achieved the best ME (Mean Error) and MPE (Mean Percentage Error) scores, indicating a better accuracy in terms of these metrics. On the other hand, the SARIMA(0,1,3)x(2,1,0)\_12 model attained the best scores in terms of RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and MASE (Mean Absolute Scaled Error).

In contrast, the SARIMA(0,1,1)x(2,1,0)\_12 model did not achieve the best accuracy values and was ranked fourth or fifth in most categories.

### 6.3 Summary

After conducting parameter significance and residuals analysis, it was observed that the SARIMA(0,1,1)x(2,1,0)\_12 model had the most significant parameters, indicating its suitability as an optimal fit. Additionally, the model achieved the lowest AIC and BIC scores, further supporting its strong performance.

Although the SARIMA(0,1,1)x(2,1,0)\_12 model did not achieve the best accuracy scores, it is crucial to consider multiple model diagnostic tools for a comprehensive evaluation. By considering the consensus among these diagnostic tools, the SARIMA(0,1,1)x(2,1,0)\_12 model emerges as the best choice for forecasting.

## 6.4 Model Diagnostics

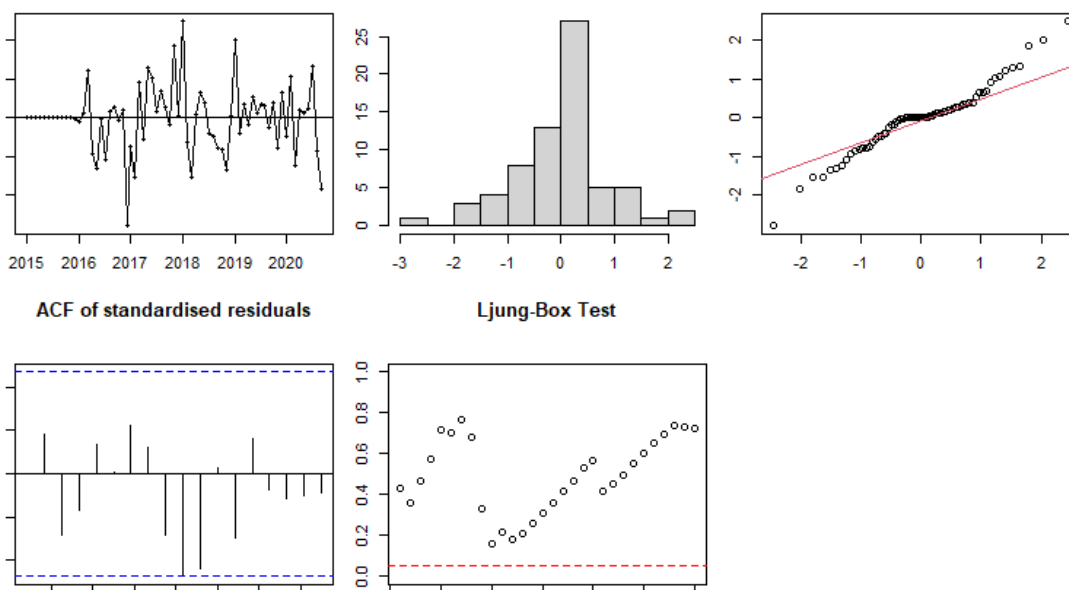
To assess whether the SARIMA(0,1,1)x(2,1,0)<sub>12</sub> model is not underfit, we can employ model overfitting as an additional diagnostic tool. This approach involves increasing the values of both  $p$  (autoregressive order) and  $q$  (moving average order) in the SARIMA model. By doing so, we can examine if the number of significant parameters and the quality of fit improve, indicating potential underfitting in the initial SARIMA(0,1,1)x(2,1,0)<sub>12</sub> model.

```
# Increasing the value of P
m4_111 <- Parameter.estimated(Data_ts_agg,c(1,1,1),c(2,1,0),method="ML", '44')

##
## Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95747, p-value = 0.01942
```

Figure 44: SARIMA(1,1,1)(2,1,0) Residual Analysis (ML)

me series plot of standardised residuals   Histogram of standardised residuals   QQ plot of standardised residuals



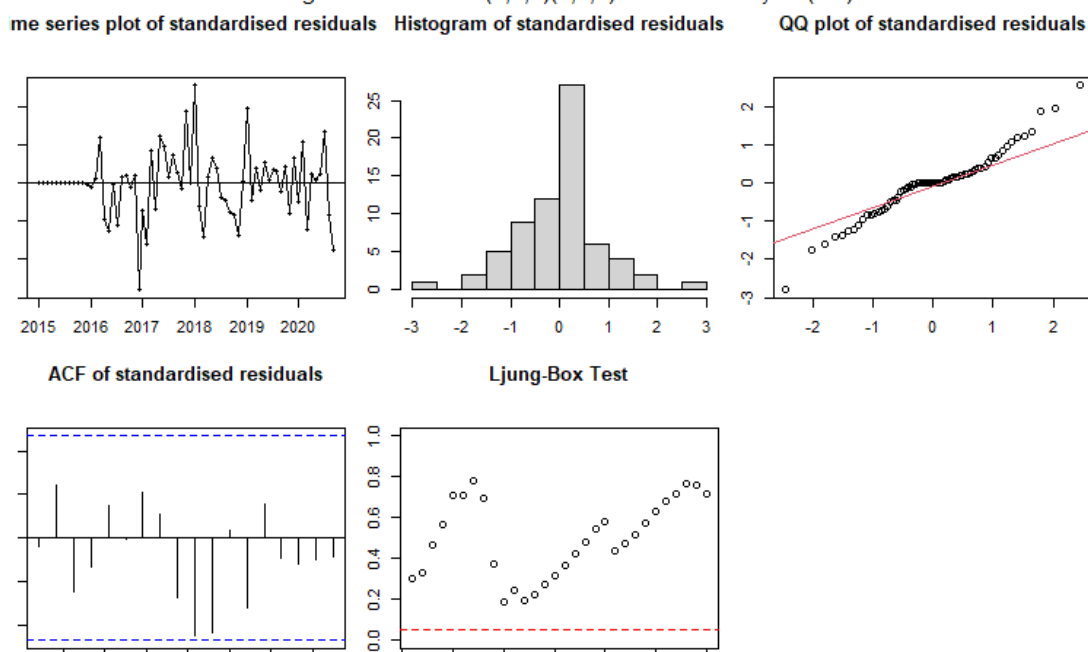
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  -0.11184    0.30303  -0.3691   0.71209
## ma1  -0.72021    0.29987  -2.4017   0.01632 *
## sar1 -0.70556    0.17030  -4.1429 3.429e-05 ***
## sar2 -0.42200    0.15851  -2.6624   0.00776 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By increasing the value of  $p$ , the AR(1) is not significant, meaning no improvement of fit has occurred by overfitting the model in this direction.

```
m4_012 <- Parameter.estimated(Data_ts_agg,c(0,1,2),c(2,1,0),method="ML", '45')
```

```
##
## Shapiro-Wilk normality test
##
## data: res.model
## W = 0.95775, p-value = 0.02012
```

Figure 45: SARIMA(0,1,2)(2,1,0) Residual Analysis (ML)



```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  -0.821572   0.149305 -5.5026 3.741e-08 ***
## ma2   0.048129   0.150831  0.3191 0.749658
## sar1  -0.712262   0.162707 -4.3776 1.200e-05 ***
## sar2  -0.416974   0.144295 -2.8897 0.003856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By increasing the value of  $p$ , the MA(1) is not significant, meaning no improvement of fit has occurred by overfitting the model in this direction.

Upon increasing both the  $p$  and  $q$  values and evaluating the resulting models, no significant improvement was observed in the overall model fit. This suggests that the initial SARIMA(0,1,1)x(2,1,0)<sub>12</sub> model is not underfit. Consequently, there is no indication of an insufficient capturing of the underlying patterns and dynamics in the data.

Since increasing the  $p$  and  $q$  values did not lead to an improved fit, it can be concluded that the SARIMA(0,1,1)x(2,1,0)<sub>12</sub> model is already well-suited for the dataset at hand. This confirmation allows for greater confidence in proceeding with forecasting tasks using the established model.

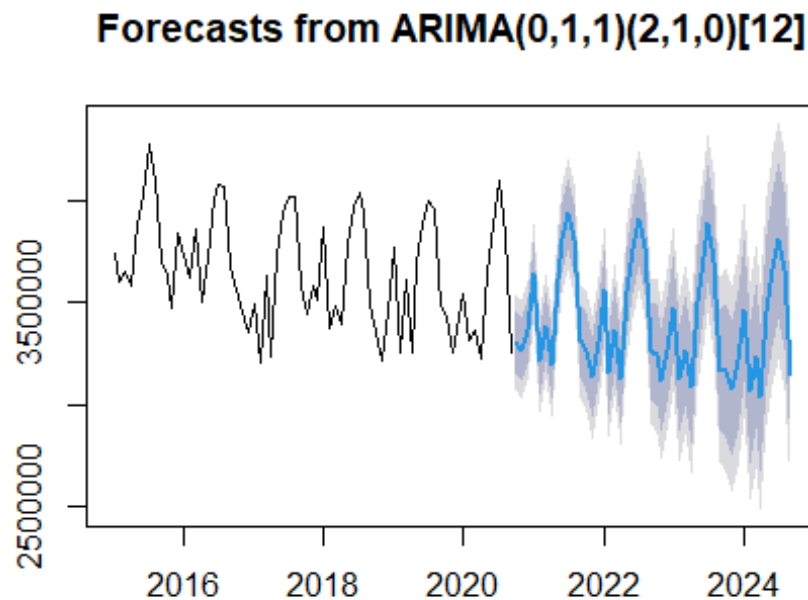
## 6.5 Forecasting

```
m5_1012.LandingA = Arima(Data_ts_agg, order=c(0,1,1), seasonal=List(order=c(2,1,0),
period=12),
method = "ML")
```



```
# Notice that I use lambda = 0 and send NMFS_Landings.ts instead of log(NMFS_Landings.ts)
# to get Arima() function to do the transformation.
# This way, I will get the forecasts in the original scale.
preds1 = forecast(m5_1012.landingA, h = 48)

plot(preds1)
```



From the forecast provided, it is evident that the SARIMA model SARIMA(0,1,1)x(2,1,0)\_12 successfully captures the seasonal trends present in the data as well as the slowly decreasing trend. This indicates that the model is effectively accounting for the patterns and dynamics in the dataset.

However, it is important to note that as the forecasting horizon extends further into the future, the confidence interval for the predictions widens significantly. This widening of the confidence interval is expected, primarily due to insufficient data caused by data loss during monthly aggregation. With less historical data available for modelling and forecasting, there is increased uncertainty and variability in the predictions.

## 7. Conclusion

After conducting various model specification and parameter estimation techniques to analyse the Victoria energy demand time series, the SARIMA(0,1,1)x(2,1,0) model emerged as the most preferable model based on several supporting factors, including the significance of its parameters and the favourable AIC/BIC scores.

On the other hand, a simpler deterministic model that combined linear and seasonal components also produced satisfactory forecasting results. However, this model suffers from overfitting, as indicated by the large R-squared value recorded. Additionally, it lacks the capability to capture changing variances across different years. Consequently, the SARIMA model is slightly more preferable.

With the inclusion of more historical data, the SARIMA model can be used effectively in future renewable energy applications. It has the potential to accurately predict energy requirements on a monthly basis, enabling the estimation of necessary battery and grid capacities to support the transition towards renewables.

## 8. References

- [1] “Victorian renewable energy and storage targets.” <https://www.energy.vic.gov.au/renewable-energy/victorian-renewable-energy-and-storage-targets> (accessed Jun. 09, 2023).
- [2] “Daily Electricity Price and Demand Data | Kaggle.” <https://www.kaggle.com/datasets/aramacus/electricity-demand-in-victoria-australia> (accessed Jun. 07, 2023).
- [3] Haydar Demirhan (2023) ‘Module 2: Analysis of Trends’[PowerPoint slides, MATH1318], RMIT University, Melbourne.
- [4] Haydar Demirhan (2023) ‘Module 8: Seasonal Models’[PowerPoint slides, MATH1318], RMIT University, Melbourne.