An Analysis of Tea Price Index

Time Series Project

Group 61 : Data Masters

Table of Contents

1. Introduction

The report aims to analyse the price index of the tea commodity between the years 1850 and 2015.The main objective of this study involves fitting the best model for the time series data among a set of possible models and predict the price index values for the next 10 years using the model proposed.

2. Methods

The data analysed in this report represents long-term commodity price index of tea since 1850, relative to real prices in the year 1900 (i.e. prices in the year 1900 equivalent to 100). The dataset has been transformed before performing any analysis on the dataset. For deterministic trend models, model specification simply consisted of identifying possible trend models to fit the data (e.g., linear and quadratic trend models). The identified models' parameters were estimated, and the model residuals were further analysed.In order to fit a stochastic trend model for the data, various model specification procedures were applied. The time series plot was investigated to identify any trend, changing variance, seasonality and the behaviour of the series. A Dickey Fuller Unit Root Test was implemented to check whether the transformed and detrended series followed a stationary process. A plausible set of ARMA models were then identified using ACF, PACF,EACF and the BIC plot. Model selection was done based on the residual analysis of the models, parameter estimation coefficient significance test and the AIC/BIC values based on the maximum likelihood(ML) estimation. The most appropriate model was then chosen amongst a set of candidate models based on the strategies mentioned and was used to forecast the price index for the next 10 years. The statistical analyses were implemented at 0.05 level of significance.

3. Results

3.1. Descriptive Statistics

From the time series plot(Figure 1), we can observe that the series exhibits a downward trend indicating a significant decrease in the price index of tea over the years with little variation in the series. There does not appear to be any seasonality however there seems to be a large noticeable decrease in the year 1861 indicating a possible intervention point.

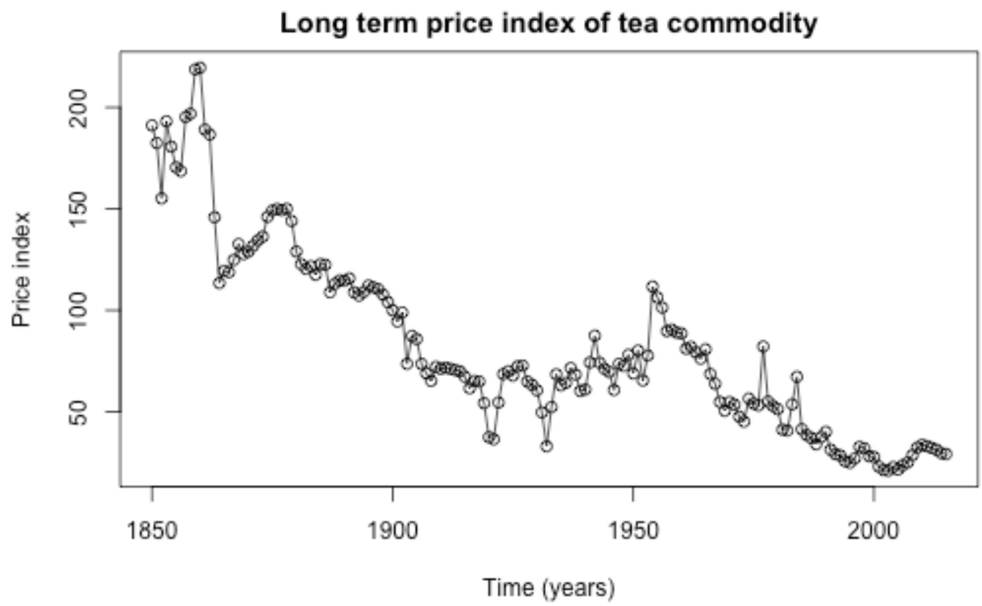

Figure 1 – Time series plot of the price index of tea from 1850 to 2015

It can be noticed that for the consecutive years ,the reported values tend to be closely related to each other. The scatterplot between the neighbouring datapoints displays a relatively strong positive correlation(Figure 2). The correlation value of 0.9737 verifies that the previous year datapoints have an effect on subsequent years and there is an autocorrelation explaining the autoregressive behaviour of the series. This result suggests that the price index values of tea between consecutive years have a tendency to be similarly valued.

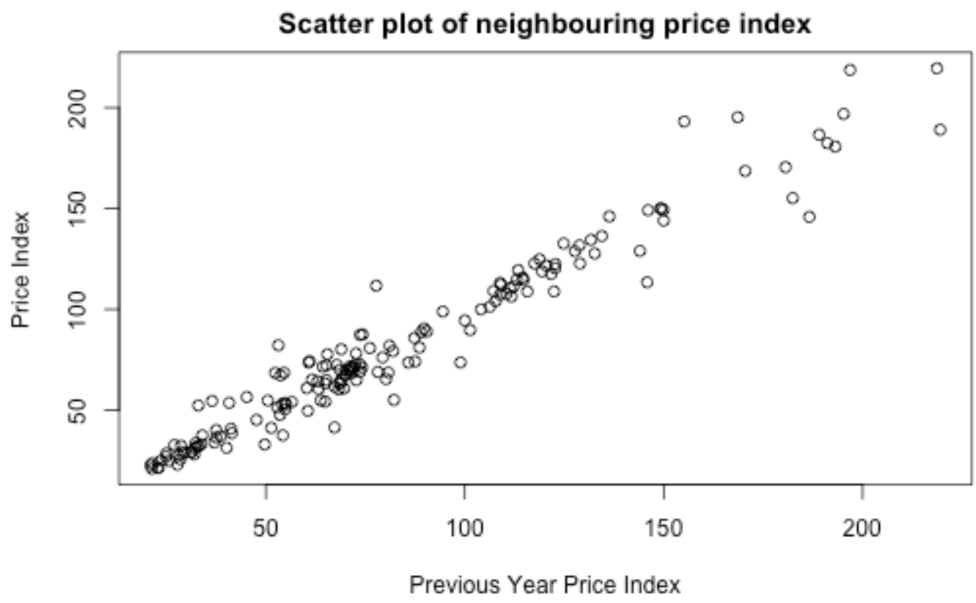

Figure 2 –Scatter plot of the consecutive yearly price values of tea

The normality distribution of the data was analysed using the Quantile-Quantile(QQ) plot and the Shapiro Wilk test. From the QQ plot(Figure 3) it can be inferred that the data distribution is not normal as the datapoints appear to be away from the straight line. The Shapiro-Wilk normality test rejects the null hypothesis with a p-value significantly less than 0.05 stating that the data is not normally distributed. (See Appendix for results)

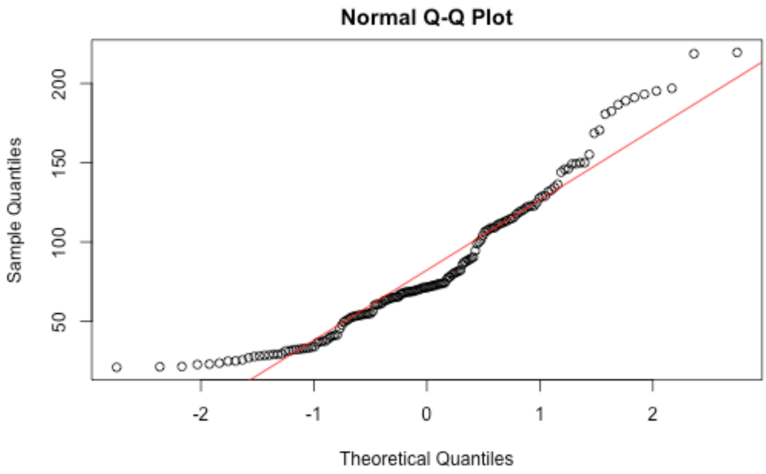

Figure 3 – Normal Quantile-Quantile Plot

3.2. Deterministic Trend Model Building Strategy

3.2.1. Linear Trend Model

A linear trend model of the form $\mu t = \beta o + \beta 1 t$ was fitted to the data where βo and $\beta 1$ stands for the intercept and the slope of the model respectively. The linear trend model assumes that the mean level of the price index changes linearly through time.

The R output below displays the least squares estimates of the linear trend model:

```
Call:
lm(formula = tea_ts ~ time(tea_ts))

Residuals:
    Min      1Q  Median      3Q     Max 
-56.186 -11.820  -1.664  12.417  76.169 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1688.98670   70.07981   24.10   <2e-16 ***
time(tea_ts)   -0.83099    0.03625  -22.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.38 on 164 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7607 
F-statistic: 525.4 on 1 and 164 DF,  p-value: < 2.2e-16
```

The estimates of slopes and intercepts were found to be -0.83099 and 1688.98670 respectively. Here the slope $\beta 1$ can be interpreted as a decrease in the mean price index of tea by 0.83 units for every additional year since negative value represents a decrease. As the p-value is less than the significance level of 0.05 for the model , linear trend model is found to be statistically significant. The adjusted R-squared value determines the variation observed in the series which is about 76%.

Figure 4 displays the linear trend line plotted over the time series plot:

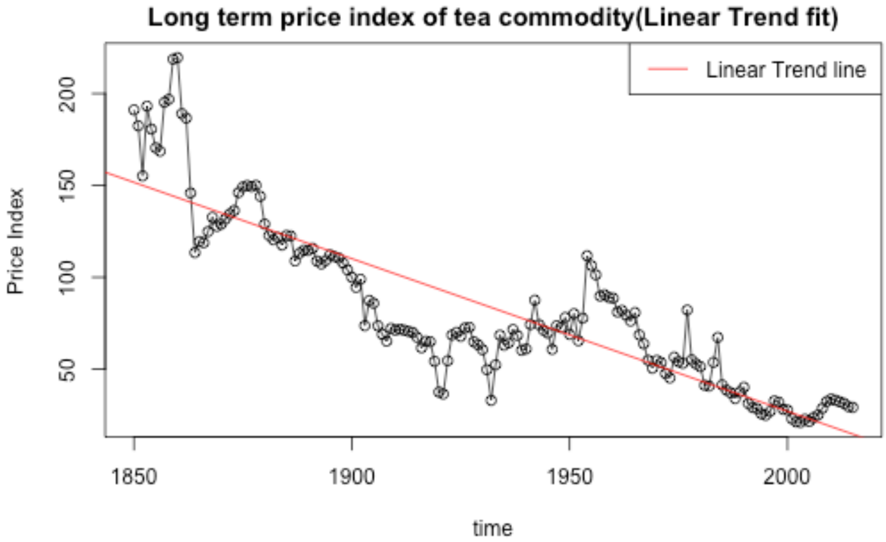

Figure 4 – Time series plot of the price index of tea from 1850 to 2015 (Linear trend fit line)

From the plot it appears that the linear trend line plotted over the time series in the plot slightly underfits the time series.

3.2.2. Residual Analysis of Linear Trend Model

We analyse the residuals of the linear trend model to help understand the stochastic component of the model for better estimation . In order to find if the linear model fits correctly, the residuals should behave roughly like the true stochastic component. Figure 5 displays the residual analysis of the linear trend model.

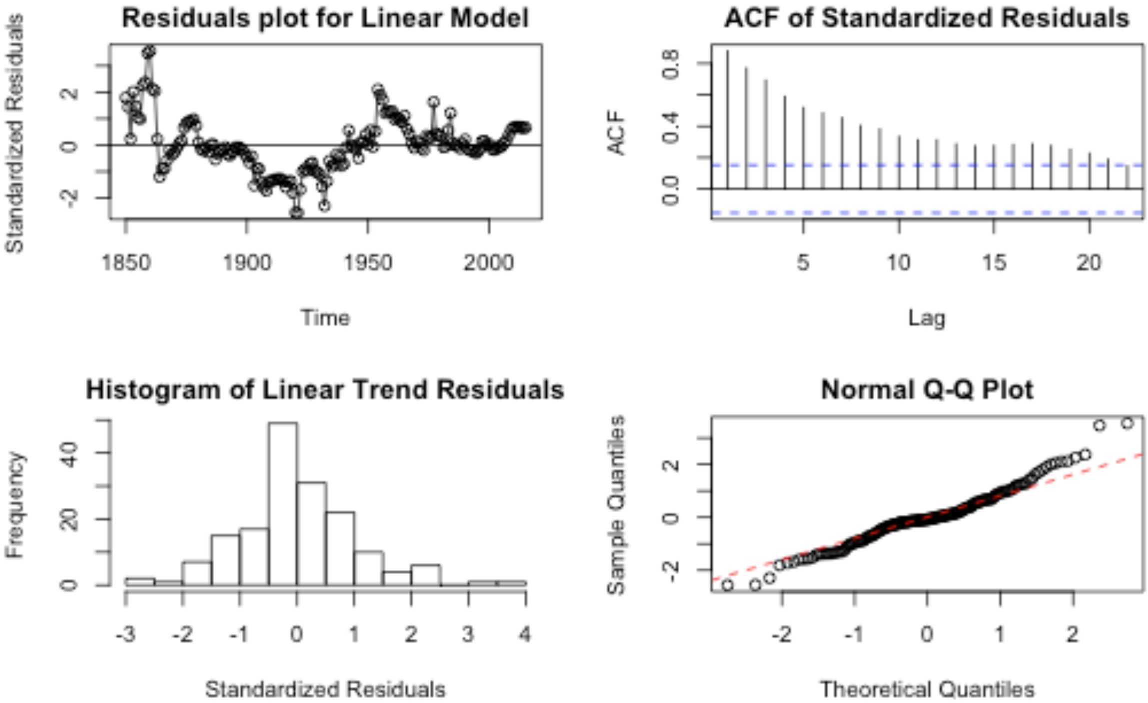

Figure 5 – Residual analysis of the linear trend model

- The time series plot of the residuals was inspected for any trend in the series and fluctuating variance. There appears to be a significant trend in the residuals as the values are not around the mean level of 0 and the plot shows considerable variation in the series.
- The sample autocorrelation function was considered to check the independence of the noise terms in the model and detect anomalies in terms of the independence of residuals. The ACF plot of the residuals show significant lags and a slowly decaying pattern thus implying non-stationarity.
- The histogram of the standardised residuals of the linear trend does not seem to follow a normal distribution . The normal Q-Q plot displays an S-shape pattern and the points seem to move away from the straight line thus indicating that the residuals are not normally distributed. The Shapiro-Wilk normality test applied to the residuals shows a p-value less than the significance level of 0.05 thus rejecting the null hypothesis stating that the residuals are not normally distributed.(see appendix for results)

3.2.3. Quadratic Trend Model

A quadratic trend model of the form $\mu t = \beta o + \beta 1t + \beta 2t2$ was fitted to the data. βo, $\beta 1$ and $\beta 2$ stands for the intercept, the slope and quadratic trend of the model, respectively. The quadratic trend model assumes that the mean level of the observed series follows a quadratic trend throughout time.

The R output below displays the least squares estimates of the quadratic trend model:

```
Call:
lm(formula = tea_ts ~ t + t2)

Residuals:
    Min      1Q  Median      3Q     Max 
-44.701 -11.581  -3.189   9.837  60.460 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.150e+04  2.761e+03   7.785 7.62e-13 ***
t           -2.134e+01  2.859e+00  -7.465 4.73e-12 ***
t2           5.307e-03  7.397e-04   7.175 2.41e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.57 on 163 degrees of freedom
Multiple R-squared:  0.8192,	Adjusted R-squared:  0.817 
F-statistic: 369.3 on 2 and 163 DF,  p-value: < 2.2e-16
```

From the summary statistics we observe that the p-value is less than the significance level of 0.05 for the model , quadratic trend term is found to be statistically significant. The model explains approximately 81% of the variation observed in the series . With a higher R-squared value than the linear trend model, the quadratic trend model appears to be a better fitting model.

Figure 6 displays the quadratic trend line plotted over the time series plot:

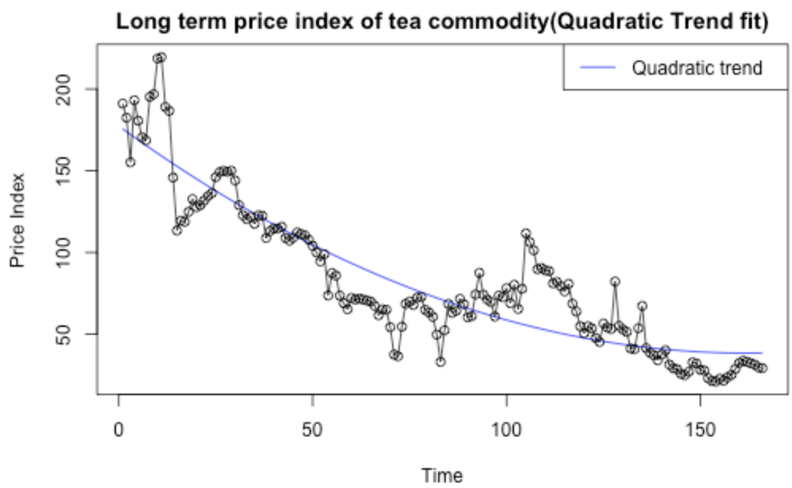

Figure 6 – Time series plot of the price index of tea from 1850 to 2015 (Quadratic trend fit line)

The quadratic trend line in the data plot appears to capture the series considerably well.

3.2.4. Residual Analysis of Quadratic Trend Model

We analyse the residuals of the quadratic trend model to help understand the stochastic component of the model. In order to find if the quadratic model is a good fit for the series, the residuals should behave roughly like the true stochastic component. Figure 7 displays the residual analysis of the quadratic trend model.

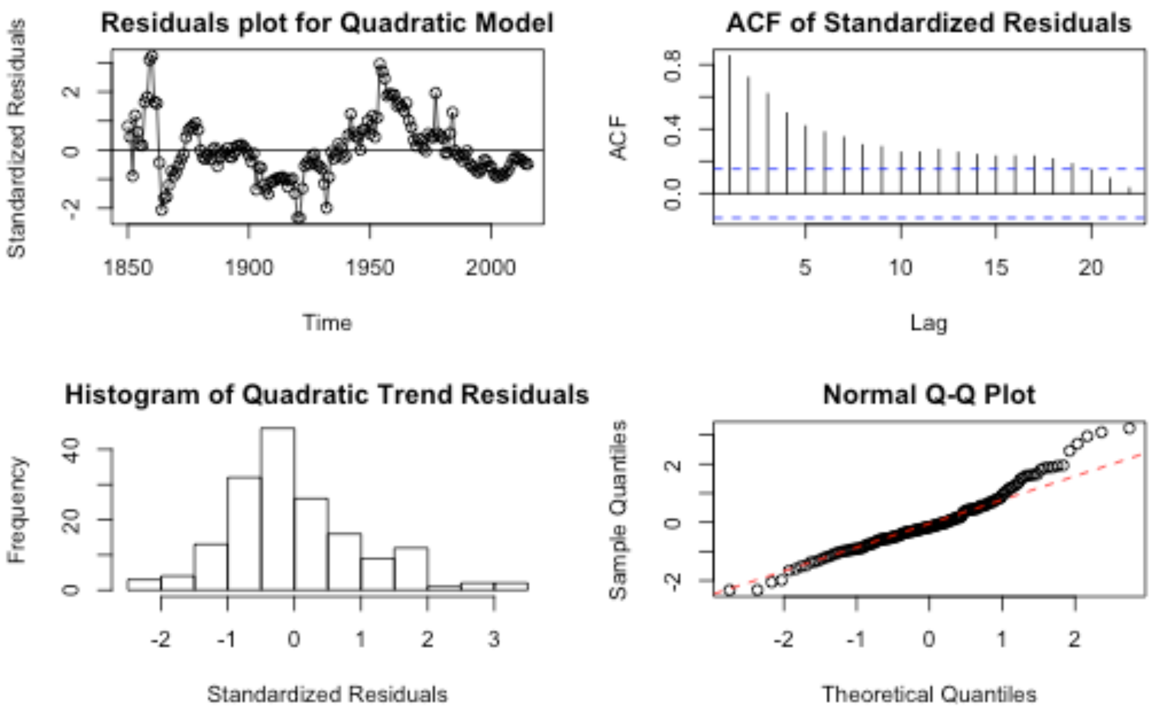

Figure 7 – Residual analysis of the quadratic trend model

- The time series plot of the residuals shows that the observations seem to follow each other and the plot appears to be too smooth to be white noise. The data does not appear to be random and the residuals seem to have variation around the mean.
- The ACF plot of the residuals show significant lags and a slowly decaying pattern thus implying the residuals do not contain any white noise .
- The histogram of the standardised residuals of the quadratic trend model again does not appear to follow a normal distribution . In the normal Q-Q plot the points seem to move away from the straight line thus indicating that the residuals are not normally distributed. The Shapiro-Wilk normality test confirms non-normality of the residuals with a p-value significantly less than 0.05 thus rejecting the null hypothesis.(see appendix for results)

3.2.5. Deterministic Trend Model Discussion

Overall, both deterministic trend models fitted were found to be significant (p-values<0.05). Between these two models, the quadratic trend model fits the series better than the linear trend model with a R squared value of 0.81 . The residuals seem to show significant autocorrelations across different lags for both models thus indicating the presence of a trend. Shapiro Wilk's Test for normality indicates that both models' residuals are not normally distributed (p-value<0.05). This is possibly due to the extreme values observed in the series as well as the small number of observations. It can be concluded that the deterministic trend model residuals are not independent ,show non-normality and cannot be said to follow a white noise process thereby violating the conditions of a true deterministic trend model. Stochastic trend modelling performed in the next section is expected to perceive the trend in a better way. We proceed with non-stationary ARIMA modelling for the series.

3.3. ARIMA Model Building Strategy

From the time series plot(Figure 8), we already know that there is a declining trend with the time points gradually decreasing. The series exhibits no seasonality however there seems to be a possible intervention point in the year 1861. The successive points appear to follow each other closely which indicate that the series displays a autoregressive behaviour and there is little change in the variation.

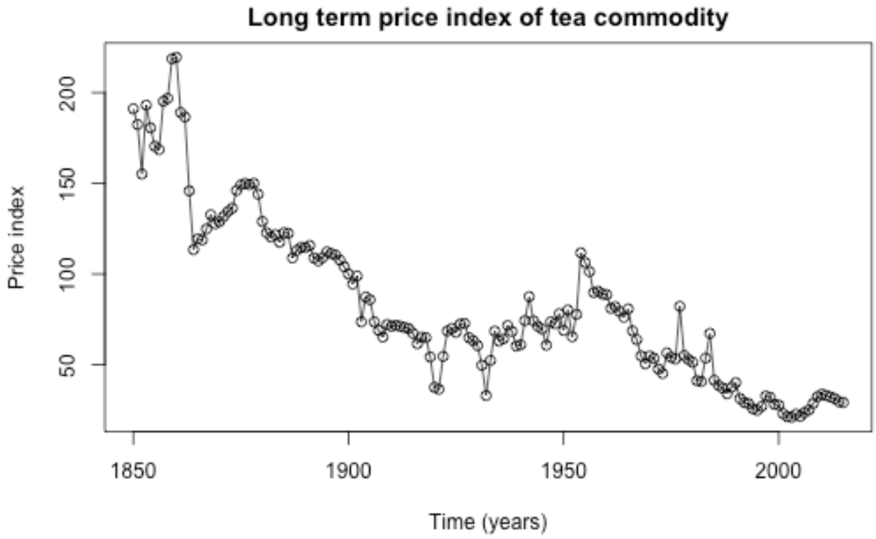

Figure 8 – Time series plot of the price index of tea from 1850 to 2015

The preliminary verification of the presence of a trend in the series was done using the autocorrelation(ACF) & partial autocorrelation(PACF) plots(Figure 9). A slow decay in the ACF plot can be seen possibly due to a strong correlation between the successive datapoints in the ACF indicating that the series is non-stationary. The PACF plot also displays a high significant autocorrelation at lag 1 and then appears to cut off.

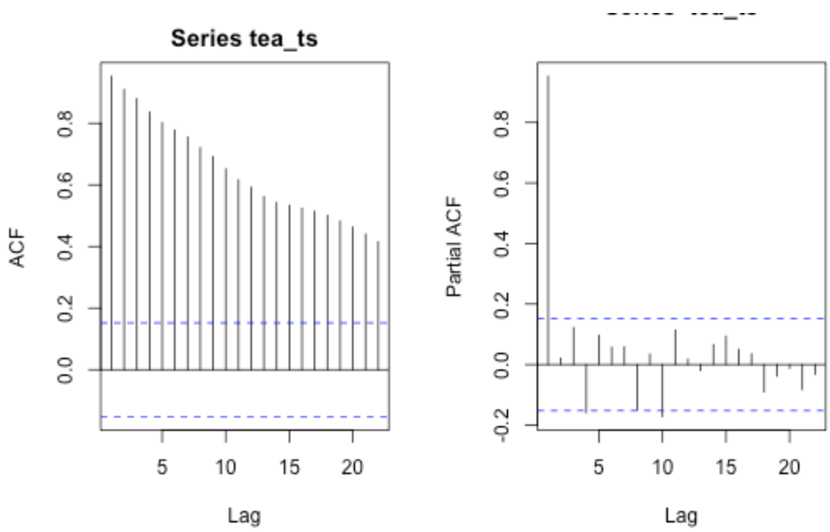

Figure 9 – ACF & PACF Plot of the time series

An Augmented Dickey-Fuller test(ADF) test was further used to find evidence and confirm whether the series is stationary or non-stationary. The p-value was found to be 0.3728 significantly larger than the stated significance level of 0.05 thus failing to reject the null hypothesis of stationarity.

```
	Augmented Dickey-Fuller Test

data:  tea_ts
Dickey-Fuller = -2.4875, Lag order = 5, p-value = 0.3728
alternative hypothesis: stationary
```

3.3.1. Series Transformation

The series needs to be stationary differenced before specifying ARIMA models for the series. However since the series has a non-normal distribution and non-constant variance , the series needs to be transformed using box-cox transformation. The lambda confidence interval was estimated for the box-cox transformation using the method of moments(Figure 10).

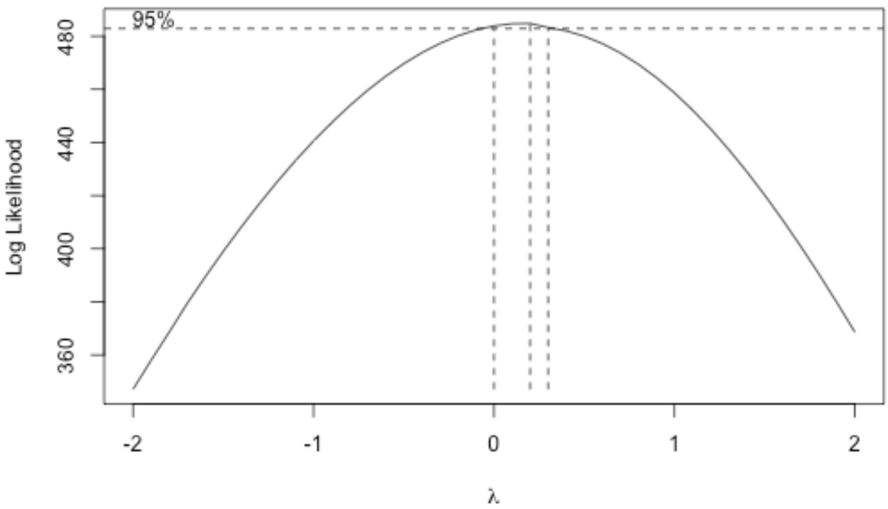

Figure 10 – Confidence Interval for lambda

The lambda value of 0 was considered and the series was transformed by applying the log transformation to normalize the data and stabilize the variance. However ,the log transformed series does not seem to stabilize the variance for the series as compared to the original time series. Figure 11 displays the log transformed series.

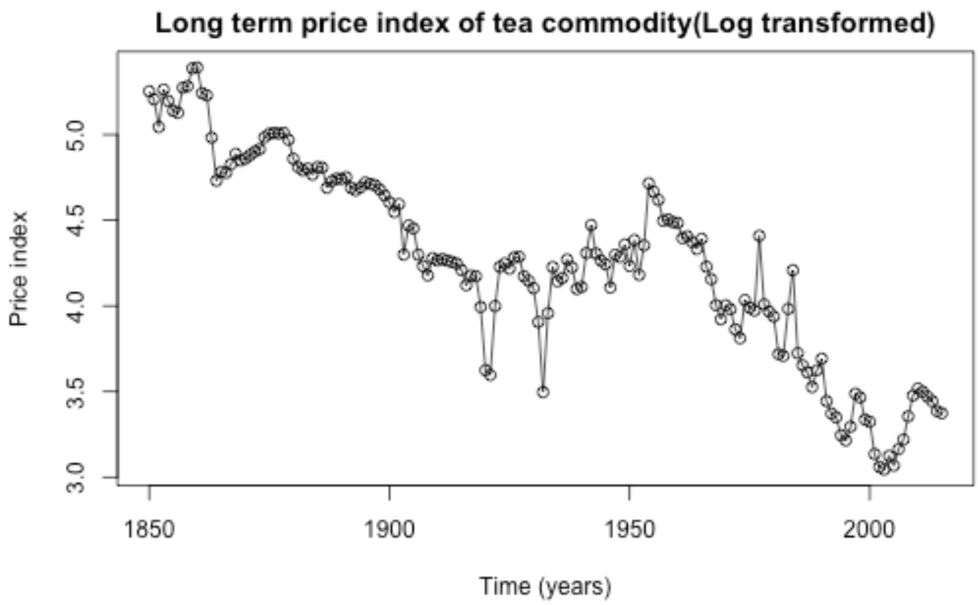

Figure 11- Time series plot of the log transformed series

The normality distribution of the series was again assessed for the log transformed series using the QQ plot and the Shapiro Wilk test. The log transformation of the series did not seem to have improved the normality of the data as it was found that the Shapiro-Wilk normality test displayed a p-value of 0.003 significantly less than the stated significance level of 0.05 and the datapoints does not appear on the straight line in the QQ-plot.(Fig.12)

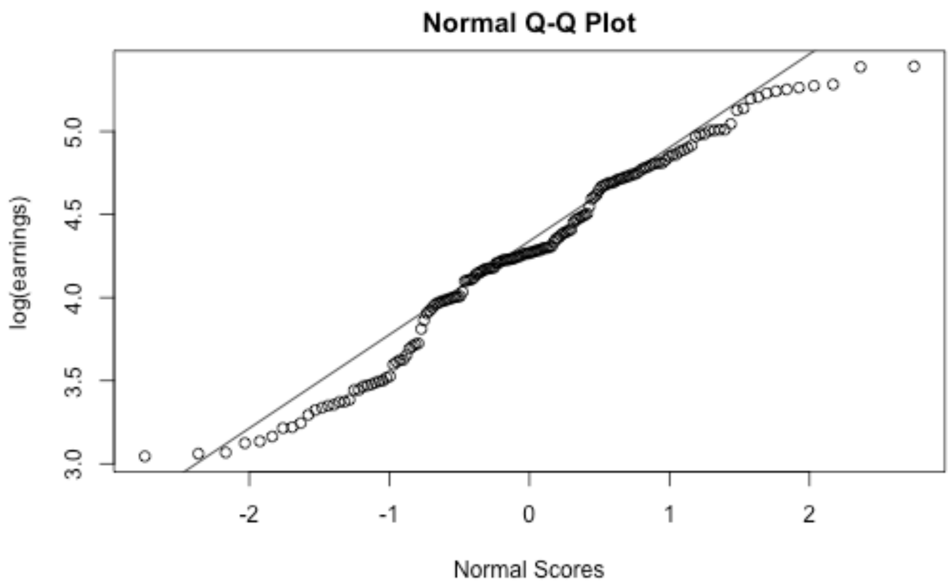

Figure 12 – Normal Quantile-Quantile Plot

Applying the ADF test to verify whether the transformed series is stationary or non-stationary. With a p-value of 0.4655 significantly larger than significance level of 0.05,the transformed series was still found to be non-stationary. The log transformed series can now be differenced in order to eliminate the trend observed.

```
        Augmented Dickey-Fuller Test

data:  tea_ts.log
Dickey-Fuller = -2.2655, Lag order = 5, p-value = 0.4655
alternative hypothesis: stationary
```

3.3.2. Series Differencing

Figure 13 displays the time series plot after taking the first difference. The first differenced series plot appears to be relatively stationary and the Augmented Dickey Fuller test was again applied to verify whether the differenced series is stationary or not. With a resulting p-value of 0.01 it can be inferred that the differenced series follows a stationary process. Hence stationary models can be used to model the log transformed and the first differenced series.

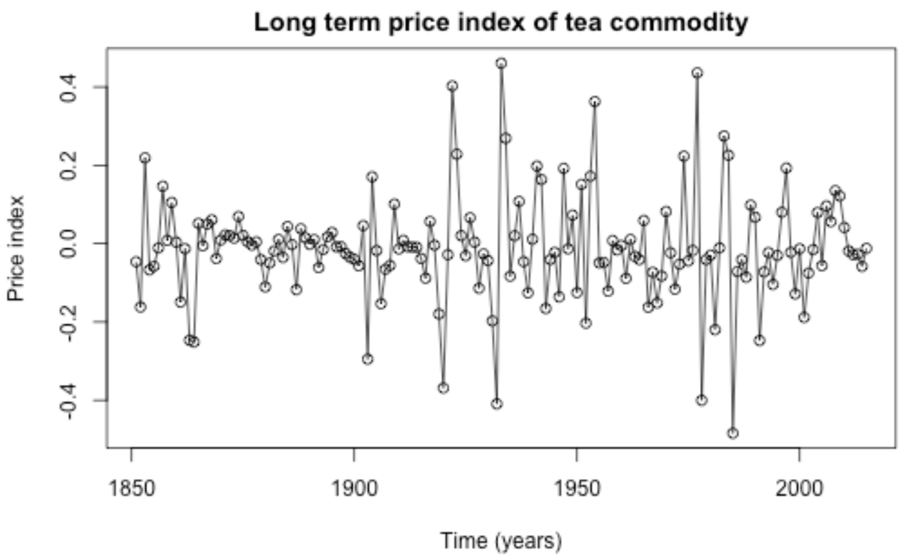

Figure 13 – Time series plot of the first differenced series

ADF test results:

```
        Augmented Dickey-Fuller Test

data:  diff.tea_ts
Dickey-Fuller = -6.806, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

3.3.3. Model Specification

The ARIMA model specification was done using the ACF, PACF,EACF and the BIC plot.

3.3.3.1. Sample Autocorrelation & Partial Autocorrelation (ACF & PACF)

The ACF and PACF plots were investigated for patterns and significant lags. From the plots(Fig 14) , we cannot clearly infer a possible model as there is no clear pattern visible in the PACF however there seems to be significant autocorrelations at lags 2 so we need to explore the extended autocorrelation function(EACF).

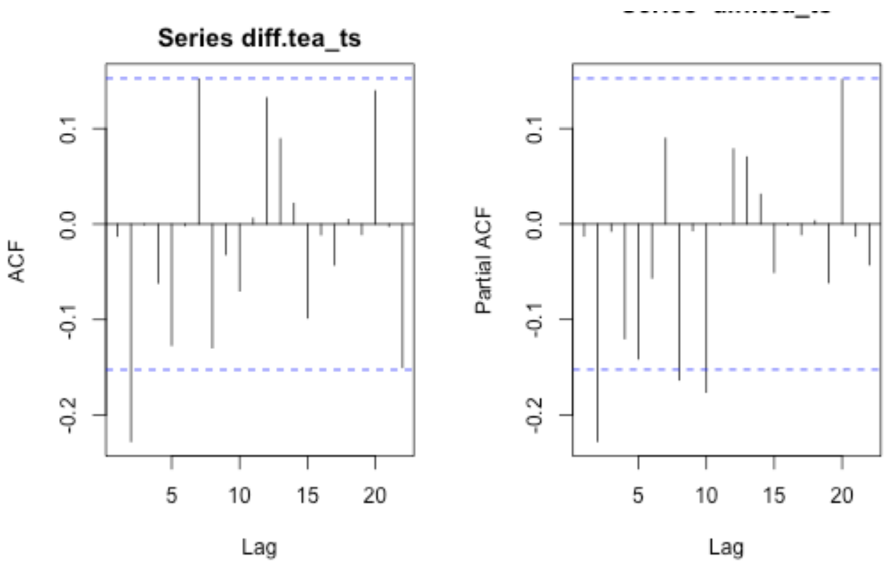

Figure 14 – ACF & PACF plot of the first differenced series

3.3.3.2. Extended Autocorrelation function(EACF)

Extended autocorrelation function identifies the orders p and q of an ARMA(p,q) model. The AR & MA component for the EACF function is set to a minimum value of 5 in order to obtain simple and parsimonious models.

AR/MA	0	1	2	3	4	5
0	o	x	o	o	o	o
1	o	x	o	o	o	o
2	o	x	o	o	o	o
3	o	x	x	o	o	o
4	x	x	o	x	o	o
5	x	x	o	x	o	o

Figure 15 – EACF table

The EACF table(Figure 15) suggests ARIMA(0,1,2),ARIMA(1,1,2),ARIMA(2,1,2) models in the set of possible models.

3.3.3.3. Bayesian Information Criterion(BIC)

The BIC plot based on the maximum likelihood estimation (MLE) shows the orders of the ARIMA models ranking from the lowest to highest.

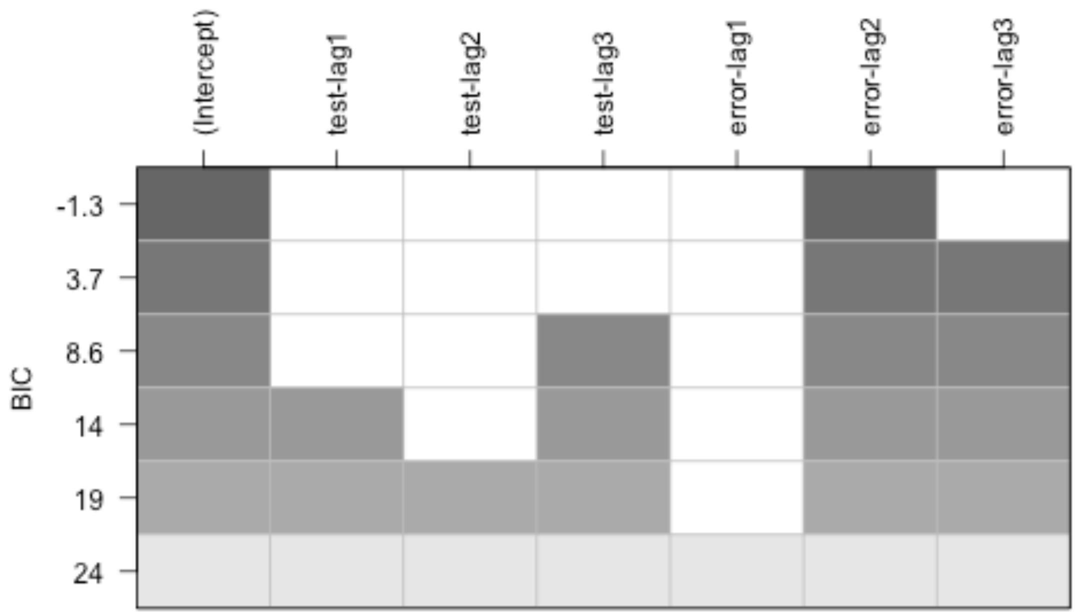

Figure 16 – BIC plot of the first differenced series

Figure 16 displays the BIC plot where the shaded regions correspond to MA(2) & MA(3) and we include ARIMA(0,1,3) also as a plausible model. ARIMA(0,1,2) is already considered based on the EACF table.

The final set of possible models proposed based on the model specification tools for the series are {ARIMA(0,1,2), ARIMA(1,1,2), ARIMA(2,1,2),ARIMA(0,1,3)}

3.3.4. Model Fitting

3.3.4.1. Parameter Estimation

The parameter estimates were tested for significance for the candidate models. The MA(2) parameter turned out to be the only parameter significant across all the specified models using the maximum likelihood estimation(ML) and conditional least square(CSS) estimation. This suggests the presence of MA(2) component to be more likely present in the final model.

Parameter estimation for ARIMA(0,1,2) based on maximum likelihood(ML) & conditional least squares(CSS) estimators:

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ma1 -0.023611   0.077264 -0.3056 0.759918
ma2 -0.278547   0.084702 -3.2885 0.001007 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ma1 -0.022105   0.077298 -0.2860 0.774903
ma2 -0.278709   0.085079 -3.2759 0.001053 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Parameter estimation for ARIMA(1,1,2) based on maximum likelihood(ML) & conditional least squares(CSS) estimators:

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1  0.272024   0.289985  0.9381  0.34821
ma1 -0.290167   0.287087 -1.0107  0.31215
ma2 -0.256558   0.091145 -2.8148  0.00488 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1  0.231208   0.332067  0.6963 0.486259
ma1 -0.247002   0.328511 -0.7519 0.452122
ma2 -0.264350   0.091403 -2.8921 0.003826 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Parameter estimation for ARIMA(2,1,2) based on maximum likelihood(ML) & conditional least squares(CSS) estimators:

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.17018    0.19325 -0.8806 0.3785357
ar2  0.33173    0.17765  1.8673 0.0618660 .
ma1  0.14487    0.17245  0.8401 0.4008661
ma2 -0.59744    0.15913 -3.7544 0.0001738 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1 -0.088549   0.183760 -0.4819   0.62989
ar2  0.359578   0.171795  2.0931   0.03634 *
ma1  0.068162   0.151340  0.4504   0.65243
ma2 -0.621688   0.139418 -4.4592 8.227e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Parameter estimation for ARIMA(0,1,3) based on maximum likelihood(ML) & conditional least squares(CSS) estimators:

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1 -0.020434   0.077151 -0.2649 0.7911189
ma2 -0.277528   0.082367 -3.3694 0.0007533 ***
ma3 -0.037017   0.083806 -0.4417 0.6587104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1 -0.019749   0.077148 -0.2560 0.7979553
ma2 -0.276981   0.083015 -3.3365 0.0008484 ***
ma3 -0.032713   0.084698 -0.3862 0.6993289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3.4.2. Model Selection

Sorting the AIC and BIC scores of all the possible models to identify the best model. The sort.score function was used to sort the models ordered from lowest to highest AIC and BIC scores. Below table shows the AIC & BIC scores of the candidate ARMA models sorted from lowest to highest.

	df <dbl>	AIC <dbl>
model_012_ml	3	-195.2331
model_112_ml	4	-193.5817
model_212_ml	5	-193.4079
model_013_ml	4	-193.3812

Table 1 – Sorting models based on AIC scores

	df <dbl>	BIC <dbl>
model_012_ml	3	-185.9153
model_112_ml	4	-181.1579
model_013_ml	4	-180.9574
model_212_ml	5	-177.8782

Table 2 – Sorting models based on BIC scores

As per the AIC & BIC table, ARIMA(0,1,2) seems to be the best fit for the series.

3.3.4.3. Model Overfitting

The ARIMA(1,1,2) & ARIMA(0,1,3) model was found to be an overfitting model for ARIMA(0,1,2).We already performed parameter estimation for the models and the AR(1) & MA(3) coefficient estimates were found to be insignificant inferring that ARIMA(1,1,2) & ARIMA(0,1,3) is an overfitting model. The ARIMA(0,1,2) model proposed seems to be the best fitting model and residual analysis was performed further for the checking the model diagnostics.

3.3.5. Model Diagnostics

Residual analysis of the selected ARIMA(0,1,2) model was conducted to determine the goodness of fit for the series. We check whether the residuals have nearly the properties of white noise by conducting diagnostic tests on the residuals. This residual analysis involves analysing the standardised residuals time series plot , histogram, QQ Plot ,ACF & PACF, Shapiro-Wilk test of the standardised residuals and the Ljung-Box test. Figure 17 displays the plot for the residuals of the ARIMA(0,1,2) model.

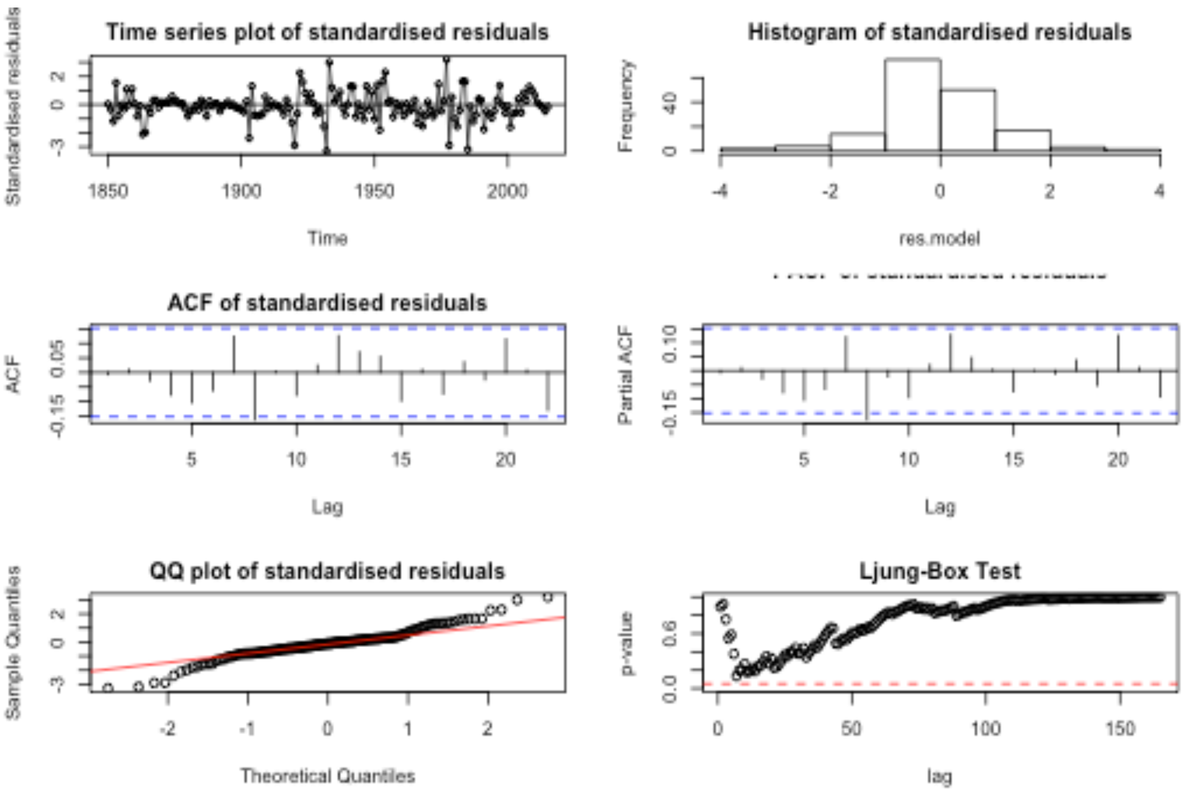

Figure 17 – Residual analysis of the ARIMA(0,1,2) model

- The time series plot of the residuals was inspected for any trend in the series and fluctuating variance. There appears to be no trend in the series however the plot showed an increased variation in the middle of the series.
- The QQ plot and the histogram was used for the analysis of the normality of the residuals. Most data points seem to follow along the straight line in the QQ plot hence the normality of the error terms is not rejected in the model. The histogram of the standardised residuals also appears to follow a normal distribution. However, the Shapiro-Wilk normality test applied to the residuals shows a p-value less than 0.05 thus rejecting the null hypothesis inferring that the residuals are not normal. This could be due to the presence of potential outlier values in the series. (see appendix for shapiro-wilk test results)
- The sample autocorrelation & partial autocorrelation function was considered to check on the independence of the noise terms in the model and to detect anomalies in terms of the independence of residuals. The ACF & PACF plot returns a slightly significant correlation at lag 8. However, since the significance is at a later lag it is considered to be insignificant thus concluding that there is no statistically significant evidence of nonzero autocorrelation in the residuals.
- The Ljung-Box test was then performed to test the lack of fit of the time series model and analyse the residual correlations at a whole with a range of p-values plotted for the Ljung-Boxtest statistic. All the p-values tend to be well above the horizontal red dashed line at 5% confidence level.

Based on the results of the performed analysis, ARIMA(0,1,2) or IMA(0,1,2) was found to be the best fit model for the time series data as the model had the lowest AIC and BIC value, the models' residuals contain white noise and the model contained the least number of parameters that needed to be estimated among the set of models.

3.4. Forecasting

Figure 18 displays the time series plot of the observed series with the forecast values of the price index for the years 2016-2025 from the ARIMA(0,1,2) model. The model predicts that the price index will remain steady for tea over the next 10 years using the CSS estimate and the ML estimate.

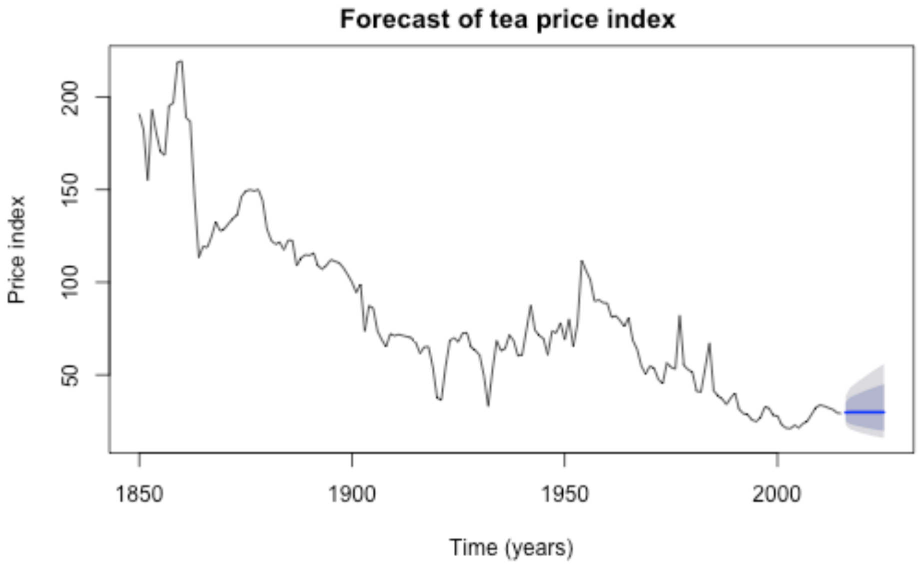

Figure 18 – Forecasting the price index values using the ARIMA(0,1,2) model

The forecast values predicted for the next 10 years using CSS estimate:

```
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1] 29.67442 29.83563 29.83563 29.83563 29.83563 29.83563 29.83563 29.83563 29.83563 29.83563
```

The forecast values predicted for the next 10 years using CSS estimate:

```
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1] 29.67514 29.83686 29.83686 29.83686 29.83686 29.83686 29.83686 29.83686 29.83686 29.83686
```

4. Conclusion

The report analysis dealt with the price index of the tea commodity to find the best fitting model and predict the yearly changes for the next 10 years. Deterministic trend models such as the linear trend and the quadratic trend models were used to fit the series however the due to their inability to fit the observed sudden increase and decrease of values in the observed series ,the stochastic models were considered. A set of possible ARIMA models were proposed using the suitable model specification tools (ACF,PACF,EACF,BIC) after using various model specification methods. From the possible ARIMA models the best possible model was identified by performing the parameter estimation and the residual analysis. ARIMA(0,1,2) was shown to rightly fit the inherent stochastic trend present in the series and selected as the most appropriate model. The model was used to forecast the price index of the tea commodity and suggested no trend in the upcoming years.

5. References

- Data Source: Jacks, D.S. (2019), "From Boom to Bust: A Typology of Real Commodity Prices in the Long Run." *Cliometrica* 13(2), 202-220. Accessed on 23 May 2020
- Demirhan ,H, Time Series Analysis ,Module notes ,RMIT University ,Melbourne , viewed 25 May 2020
- Cryer, J. and Chan, K., 2011. Time Series Analysis. New York: Springer.

6. Appendix(R Codes)

```
library(TSA)

library(tseries)

library(fUnitRoots)

library(lmtest)

library(FitAR)

library(forecast)

source('sort.score.R')

source('residual.analysis.R')


tea <- read.csv("longtermpriceindex.csv",header = TRUE)    # Reading the dataset using read.csv
function.

head(tea)

class(tea)

tea_ts <- ts(tea$Tea, start = 1850)  # Converting the data frame to time series object

class(tea_ts)   # rechecking if the  conversion was successful
```

plot(tea_ts,xlab='Time (years)',type="o",ylab = "Price index", main = "Long term price index of tea commodity") # plot time series

#ScatterPlot

plot(y=tea_ts,x=zlag(tea_ts), xlab = "Previous Year Price Index",ylab="Price Index",main = "Scatter plot of neighbouring price index")

#Correlation

y=tea_ts   #Read the abundance data into y

x=zlag(tea_ts)   #Generate first lag of the abundance series

index=2:length(x)   # Create an index to get rid of the first NA values

cor(y[index],x[index]) #Calculate correlation between x and y.

#QQ plot & Shapiro wilk test

qqnorm(tea_ts)

qqline(tea_ts, col = 2)

shapiro.test(tea_ts)

```
Shapiro-Wilk normality test

data:  tea_ts
W = 0.92684, p-value = 1.896e-07
```

Linear model

linear_model=lm(formula = tea_ts ~ time(tea_ts))

summary(linear_model)

#Linear Model with Trend Line

plot(tea_ts, main="Long term price index of tea commodity(Linear Trend fit)", xlab='Time', ylab='Price Index', type='o')

abline(temp, col="red")

legend ("topright", lty = 1, col = c("red"), text.width = 38,

c("Linear Trend line"))

```
# Residual analysis of Linear Trend

res.model = rstudent(linear_model)

par(mfrow=c(2,2))

plot(y = res.model, x = as.vector(time(tea_ts)),xlab = 'Time', ylab='Standardized Residuals',type='o',
main = "Residuals plot for Linear Model")

abline(h=0)

acf(res.model,main="ACF of Standardized Residuals")

hist(res.model,xlab='Standardized Residuals',main='Histogram of Linear Trend Residuals')

qqnorm(res.model)

qqline(res.model, col = 2, lwd = 1, lty = 2)

shapiro.test(res.model)

Shapiro-Wilk normality test

data:  res.model
W = 0.97042, p-value = 0.001282


# Quadratic model with curve

t = time(tea_ts)

t2 = t^2

quadratic_model = lm(tea_ts~ t + t2)

summary(quadratic_model)

plot(ts(fitted(quadratic_model)), ylim=c(min(c(fitted(quadratic_model),

                          as.vector(tea_ts))),max(c(fitted(quadratic_model),as.vector(tea_ts)))),
main="Long term price index of tea commodity(Quadratic Trend fit)", xlab='Time', ylab='Price
Index',col='blue')

lines(as.vector(tea_ts),type="o")

legend ("topright", lty = 1, col = c("blue"), text.width = 38,

c("Quadratic trend"))

# Residual analysis of Quadratic trend

res.model2 = rstudent(quadratic_model)

par(mfrow=c(2,2))

plot(y = res.model2, x = as.vector(time(tea_ts)),xlab = 'Time', ylab='Standardized Residuals',type='o',
main = "Residuals plot for Quadratic Model")

abline(h=0)
```

```
acf(res.model2,main="ACF of Standardized Residuals")

hist(res.model2,xlab='Standardized Residuals',main='Histogram of Quadratic Trend Residuals')

qqnorm(res.model2)

qqline(res.model2, col = 2, lwd = 1, lty = 2)

shapiro.test(res.model2)

# ACF & PACF

par(mfrow=c(1,2))

acf(tea_ts)

pacf(tea_ts)


# adf test

adf.test(tea_ts)


# Log transformation of series

tea_ts.transform = BoxCox.ar(tea_ts,method="yule-walker")

tea_ts.transform$ci

tea_ts.log = log(tea_ts) # 0 is in the interval or you can go for mid point of the interval as well


# QQ-plot & shapiro wilk test

qqnorm(tea_ts.log, ylab="log(earnings)", xlab="Normal Scores")

qqline(tea_ts.log)

shapiro.test(tea_ts.log)

Shapiro-Wilk normality test

data:  tea_ts.log
W = 0.97455, p-value = 0.003762

#Log transformed series plot

plot(tea_ts.log,xlab='Time (years)',type="o",ylab = "Price index", main = "Long term price index of
tea commodity(Log transformed)")

# adf test to check whether series stationary or not

adf.test(tea_ts.log)
```

```
# Time series plot of differenced series

diff.tea_ts= diff(tea_ts.log,differences=1)

plot(diff.tea_ts,xlab='Time (years)',type="o",ylab = "Price index", main = "Long term price index of tea commodity")

adf.test(diff.tea_ts) # Adf test

# ADF & PACF plot for differenced series

par(mfrow=c(1,2))

acf(diff.tea_ts)

pacf(diff.tea_ts)

# EACF plot

eacf(diff.tea_ts,ar.max = 5, ma.max = 5)

#BIC plot

bic = armasubsets(y=diff.tea_ts,nar=3,nma=3,y.name='test',ar.method='ols')

plot(bic)

# Parameter Estimation

#ARIMA(0,1,2)

model_012_css = arima(tea_ts.log,order=c(0,1,2),method='CSS')

coeftest(model_012_css)

model_012_ml = arima(tea_ts.log,order=c(0,1,2),method='ML')

coeftest(model_012_ml)

# ARIMA(1,1,2)

model_112_css = arima(tea_ts.log,order=c(1,1,2),method='CSS')

coeftest(model_112_css)

model_112_ml = arima(tea_ts.log,order=c(1,1,2),method='ML')

coeftest(model_112_ml)

# ARIMA(0,1,3)

model_013_css = arima(tea_ts.log,order=c(0,1,3),method='CSS')

coeftest(model_013_css)

model_013_ml = arima(tea_ts.log,order=c(0,1,3),method='ML')

coeftest(model_013_ml)
```

Sort AIC & BIC

sort.score(AIC(model_012_ml,model_013_ml,model_112_ml,model_212_ml), score = "aic")

sort.score(BIC(model_012_ml,model_013_ml,model_112_ml,model_212_ml), score = "bic")

Residual analysis

residual.analysis(model = model_012_ml)

```
Shapiro-Wilk normality test

data:  res.model
W = 0.95388, p-value = 2.818e-05
```

Forecasting

fit = Arima(tea_ts,c(0,1,2),lambda = 0) #Specifying the lambda value of the box cox transformation

plot(forecast(fit,h=10),xlab='Time (years)',ylab = "Price index", main = "Forecast of tea price index ")

Forecast values for ML & CSS estimate

arima012.mle <- Arima(tea_ts, c(0, 1, 2), method = "ML", include.mean = FALSE,lambda = 0) # mle estimate

arima012.mle.pred <- exp(predict(arima012.mle, n.ahead = 10)$pred) # mle estimate prediction

arima012.mle.pred

arima012.css <- Arima(tea_ts, c(0, 1, 2), method = "CSS", include.mean = FALSE,lambda = 0) # css estimate

arima012.css.pred <- exp(predict(arima012.css, n.ahead = 10)$pred) # css estimate prediction

arima012.css.pred