Trend & Seasonality Elimination by Differencing Method

Pranav Gopalkrishna, 1940223

2022-01-26

Table of Contents

[INTRODUCTION & OBJECTIVES 1](#_Toc94114863)

[Need for elimination of trend & seasonality 1](#_Toc94114864)

[Objectives 1](#_Toc94114865)

[DATA FOR BOTH QUESTIONS 1](#_Toc94114866)

[Importing necessary data 1](#_Toc94114867)

[Defining and formatting the ‘Month’ variable 2](#_Toc94114868)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Removing trend from time series data on softlog prices

[DATA 2](#_Toc94114869)

[Defining and reformatting the ‘Softlog.Price’ variable 2](#_Toc94114870)

[Summarizing the data 2](#_Toc94114871)

[Creating a time series for softlog prices 3](#_Toc94114872)

[TIME PLOT AND POTENTIAL TIME SERIES COMPONENTS 3](#_Toc94114873)

[ELIMINATING TREND THROUGH METHOD OF DIFFERENCING 4](#_Toc94114874)

[TESTING STATIONARITY OF TIME SERIES 5](#_Toc94114875)

[ADF test 5](#_Toc94114876)

[Performing the test using the ‘tseries’ library 5](#_Toc94114877)

[Original time series stationarity 5](#_Toc94114878)

[Trend-removed time series stationarity 6](#_Toc94114879)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. Removing trend & seasonality from data on rubber prices

[DATA 7](#_Toc94114880)

[Defining and reformatting the ‘Rubber.Price’ variable 7](#_Toc94114881)

[Summarizing the data 8](#_Toc94114882)

[Creating a time series for softlog prices 8](#_Toc94114883)

[TIME PLOT AND POTENTIAL TIME SERIES COMPONENTS 8](#_Toc94114884)

[ELIMINATING TREND THROUGH METHOD OF DIFFERENCING 9](#_Toc94114885)

[TESTING STATIONARITY OF TIME SERIES 10](#_Toc94114886)

[Original time series stationarity 10](#_Toc94114887)

[Seasonality-removed time series stationarity 11](#_Toc94114888)

[Trend and seasonality removed time series stationarity 12](#_Toc94114889)

# INTRODUCTION & OBJECTIVES

## Need for elimination of trend & seasonality

The elimination of trend and seasonality in the data is done in order to make the time series stationary. This makes it easier to analyze and model.

## Objectives

In this assignment, I will aim to eliminate seasonality and trend in a given time series using the method of differencing. I will also be evaluating the stationarity of the resulting time series using time plots, ACF plots and the Dickey-Fuller test for stationarity.

# DATA FOR BOTH QUESTIONS

This dataset records data about the prices of various agricultural raw materials, including coarse wool, cotton, fine wool, sawn wood, etc., from the year 1990 up to 2017. The data points are separated by a monthly time interval i.e. data was observed for every month. In this assignment, we will only look at softlog prices and rubber prices.

## Importing necessary data

setwd("~/Documents/Study/computerScience/programming/r/data/")  
data = read.csv("agriculturalRawMaterial.csv")[c(1, 18, 20)]  
head(data)

**Month | Rubber.Price | Softlog.Price**  
Apr-90 | 0.84 | 120.66  
May-90 | 0.85 | 124.28  
Jun-90 | 0.85 | 129.45  
Jul-90 | 0.86 | 124.23  
Aug-90 | 0.88 | 129.70  
Sep-90 | 0.90 | 129.78

## Defining and formatting the ‘Month’ variable

t = c()  
for(x in data$Month)  
{  
 x = paste("01", x, sep = "-")  
 t = c(t, x)  
}  
t = as.Date(t, format = "%d-%b-%y")  
# %b => abbreviated month  
# %y => 2 digit year  
# It can recognize century on its own.  
# So for example, '93' will be interpreted as '1993'.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Removing trend from data on softlog prices

# DATA

## Defining and reformatting the ‘Softlog.Price’ variable

z = c()  
for(p in data$Softlog.Price)  
{  
 p = gsub(',', '', p)  
 # Removing commas in the numbers  
 # 1st argument => what to replace  
 # 2nd argument => what to put instead  
 # 3rd argument => full string  
 z = c(z, p)  
}  
z = as.numeric(z)

## Summarizing the data

df = data.frame(t, z)  
summary(df)

**t | z**   
Min. :1990-04-01 | Min. :119.3   
1st Qu.:1997-10-01 | 1st Qu.:146.0   
Median :2005-04-01 | Median :160.4   
Mean :2005-04-01 | Mean :164.5   
3rd Qu.:2012-10-01 | 3rd Qu.:180.2   
Max. :2020-04-01 | Max. :260.0   
 | NA's :34

We can see there are 34 missing values in the price column. Checking in the dataset itself, values for softlog prices are available until a certain point, after which we have these missing values. Hence, we can simply remove the tail end of the dataset where these missing values are concentrated.

t\_new = t[c(1:(length(t) - 34))]  
z\_new = z[c(1:(length(z) - 34))]  
df = data.frame(t\_new, z\_new)  
summary(df)

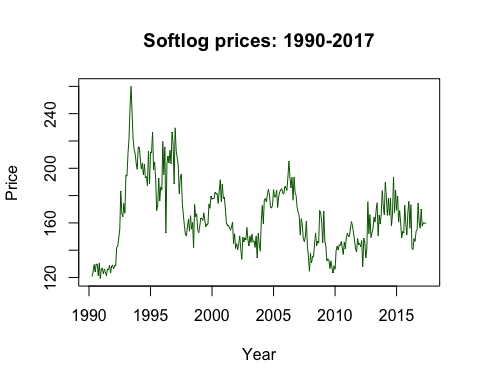
**t\_new | z\_new**   
Min. :1990-04-01 | Min. :119.3   
1st Qu.:1997-01-16 | 1st Qu.:146.0   
Median :2003-11-01 | Median :160.4   
Mean :2003-10-31 | Mean :164.5   
3rd Qu.:2010-08-16 | 3rd Qu.:180.2   
Max. :2017-06-01 | Max. :260.0

## Creating a time series for softlog prices

Z = ts(z\_new, start = c(1990, 4, 1), end = c(2017, 6, 1), frequency = 12)  
# frequency = 12 -> monthly frequency

# TIME PLOT AND POTENTIAL TIME SERIES COMPONENTS

ts.plot(Z,  
 main = "Softlog prices: 1990-2017",  
 ylab = "Price",  
 xlab = "Year",  
 col = "darkgreen")



Based on the time plot, we may conclude that there is no clear seasonality (i.e. no periodic fluctuations annually). There seems to be an overall downward trend over time, with a seemingly high level of irregular fluctuations. We also see no discernible long-term fluctuation pattern; hence we may conclude that there is no cyclical fluctuation. Hence, the major components of this time series seem to be trend and irregular fluctuations.

# ELIMINATING TREND THROUGH METHOD OF DIFFERENCING

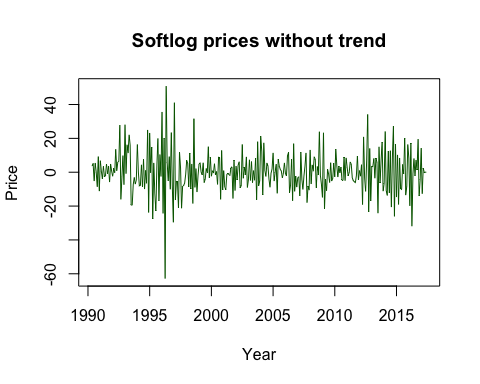
We will use the model

*zt = mt + et*

where zt is the softlog price at time t, mt is the trend level at time t, et is the error at time t (irregular fluctuations). We assume there to be a constant downward trend component.

To use the method of differencing to remove trend, we use the ‘diff’ function. Here, since we want to remove trend, and since we assume constant trend, we use lag = 1 i.e. we find take differences between each unit of time (the unit of time in our case in one month).

Z1 = diff(Z, lag = 1)  
ts.plot(Z1,  
 main = "Softlog prices without trend",  
 ylab = "Price",  
 xlab = "Year",  
 col = "darkgreen")



We can observe that the overall downward trend previously observed in this time series is now absent. Even now, we cannot observe any clear seasonal fluctuation, further confirming the validity of our model in this case.

# TESTING STATIONARITY OF TIME SERIES

## ADF test

ADF i.e. augmented Dickey-Fuller test is used to check the stationarity of a time series using an autoregression model based on the time series. Its null hypothesis is that there is the data is not stationary, and its alternative hypothesis is that the data is stationary. The ADF statistic is a negative number. The more negative it is, the stronger the rejection of the hypothesis. As with all statistical tests, the p-value indicates whether we may or may not reject the null hypothesis. Only if it is lower than the level of significance (in our case, this is 5% or 0.05) may we reject the null hypothesis with the given level of confidence.

## Performing the test using the ‘tseries’ library

# install.packages("tseries")  
library(tseries)

**Warning:** package 'tseries' was built under R version 3.6.2

Registered S3 method overwritten by 'quantmod':  
**method | from**  
as.zoo.data.frame | zoo

### Original time series stationarity

adf.test(Z)

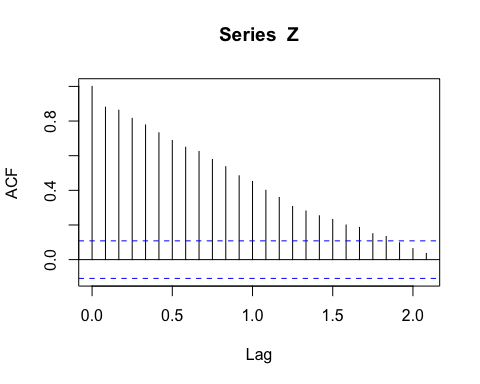
**Augmented Dickey-Fuller Test**

**data:** Z  
***Dickey-Fuller*** = -3.3339,***Lag order*** = 6, ***p-value*** = 0.06568  
**alternative hypothesis:** stationary

The p-value for lag order 6 is 0.06568, which is above 0.05 i.e. the significance level. Hence, we may accept the null hypothesis and conclude that the time series is not stationary.

**Observing ACF plot**

acf(Z)



As can be observed, the autocorrelation in the time series is significant from lags 1 to 1.5. Hence, we can say that there is significant autocorrelation between observations separated by one month. Such an ACF plot suggests that the data is not stationary, because the autocorrelation between observations is significant and change with in a relatively defined manner with respect to lag, which indicates that the data may show certain clear movements over time, such as trend.

### Trend-removed time series stationarity

adf.test(Z1)

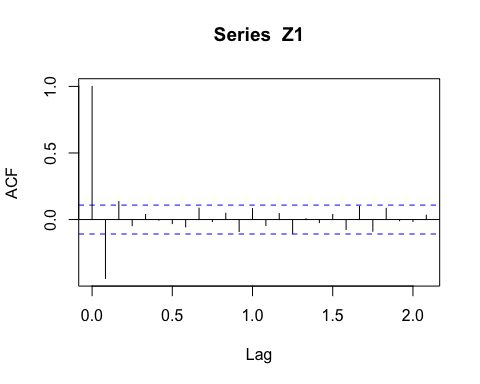
**Warning in adf.test(Z1):** p-value smaller than printed p-value

**Augmented Dickey-Fuller Test**  
**data:** Z1  
***Dickey-Fuller*** = -7.8913, ***Lag order*** = 6, ***p-value*** = 0.01  
**alternative hypothesis:** stationary

The p-value for lag order 6 is 0.01, which is below 0.05 i.e. the significance level. Hence, we may accept the reject hypothesis and conclude that the time series is stationary.

**Observing ACF plot**

acf(Z1)



As can be observed, the autocorrelation in the time series is insignificant for all lags from 1 to 2. Hence, we can say that there is no significant autocorrelation between observations separated by one or more months. Such an ACF plot suggests that the data is stationary, because the autocorrelation between observations is insignificant and change seemingly randomly with respect to lag, which indicates that the data does show certain clear movements over time, such as trend, indicating a white-noise quality and hence stationarity.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. Removing trend & seasonality from data on rubber prices

# DATA

## Defining and reformatting the ‘Rubber.Price’ variable

z = c()  
for(p in data$Rubber.Price)  
{  
 p = gsub(',', '', p)  
 # Removing commas in the numbers  
 # 1st argument => what to replace  
 # 2nd argument => what to put instead  
 # 3rd argument => full string  
 z = c(z, p)  
}  
z = as.numeric(z)

## Summarizing the data

df = data.frame(t, z)  
summary(df)

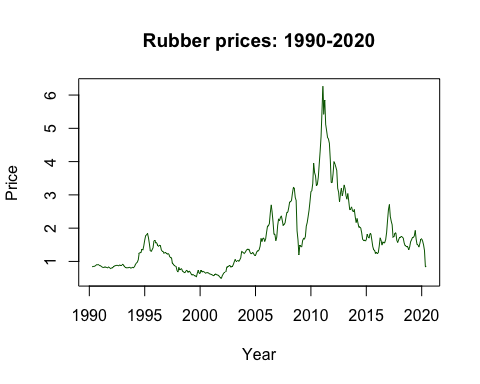
**t | z**   
Min. :1990-04-01 | Min. :0.490   
1st Qu.:1997-10-01 | 1st Qu.:0.860   
Median :2005-04-01 | Median :1.440   
Mean :2005-04-01 | Mean :1.656   
3rd Qu.:2012-10-01 | 3rd Qu.:2.060   
Max. :2020-04-01 | Max. :6.260

## Creating a time series for softlog prices

Z = ts(z, start = c(1990, 4, 1), end = c(2020, 6, 1), frequency = 12)  
# frequency = 12 -> monthly frequency

# TIME PLOT AND POTENTIAL TIME SERIES COMPONENTS

ts.plot(Z,  
 main = "Rubber prices: 1990-2020",  
 ylab = "Price",  
 xlab = "Year",  
 col = "darkgreen")



Based on the time plot, we may conclude that there may be some seasonality (i.e. some periodic fluctuations annually), as we see how at the start of most years, the prices shoot upwards, and tend to fall before rising again the next year. There is no constant trend over time, but we may interpret the graph’s movement from 2000 to 2011 as a roughly upward trend. Since we cannot confirm whether this trend flattens over time, we may consider the existence of a trend component. There is a noticeable of irregular fluctuations. We also see no discernible long-term fluctuation pattern. Hence we may conclude that there is no cyclical fluctuation. Hence, the major components of this time series seem to be trend and irregular fluctuations.

# ELIMINATING TREND THROUGH METHOD OF DIFFERENCING

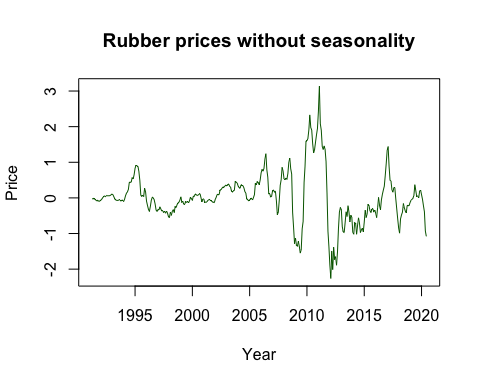
We will use the model

*zt = mt + + st + et*

where *zt* is the softlog price at time t, *mt* is the trend level at time t*, st* is the seasonal fluctuation at time t (seasonality), and *et* is the error at time t (irregular fluctuations). We assume there to be a constant upward trend component, although this is not always the case in the data, as far as we can see.

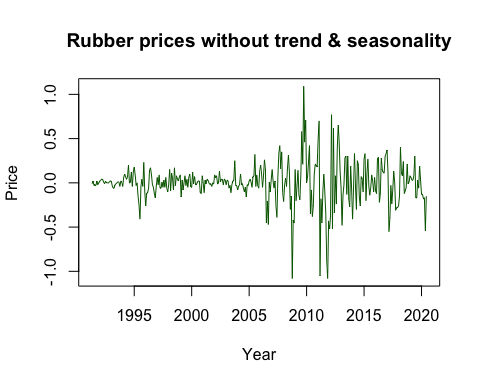
To use the method of differencing to remove trend, we use the ‘diff’ function. Seasonality is the existence of periodic patterns that span over a period of 12 months. Hence, we will use lag = 12 to ***eliminate seasonality***.

Z1 = diff(Z, lag = 12)  
ts.plot(Z1,  
 main = "Rubber prices without seasonality",  
 ylab = "Price",  
 xlab = "Year",  
 col = "darkgreen")



To use the method of differencing to remove trend, we use the ‘diff’ function. Here, since we want to ***remove trend***, and since we assume constant trend, we use lag = 1 i.e. we find take differences between each unit of time (the unit of time in our case in one month).

Z2 = diff(Z1, lag = 1)  
ts.plot(Z2,  
 main = "Rubber prices without trend & seasonality",  
 ylab = "Price",  
 xlab = "Year",  
 col = "darkgreen")



# TESTING STATIONARITY OF TIME SERIES

## Original time series stationarity

adf.test(Z)

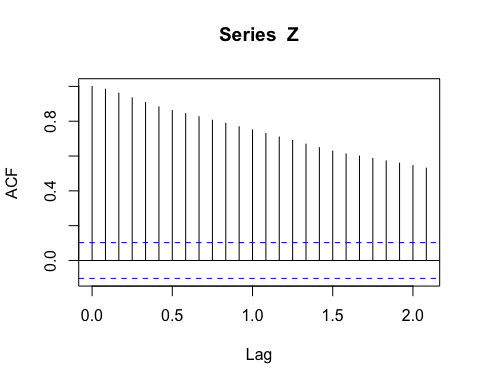
**Augmented Dickey-Fuller Test**

**data:** Z  
***Dickey-Fuller*** = -1.8319, ***Lag order*** = 7, ***p-value*** = 0.6479  
**alternative hypothesis:** stationary

The p-value for lag order 7 is 0.6479, which is well above 0.05 i.e. the significance level. Hence, we may accept the accept hypothesis and conclude that the time series is not stationary.

**Observing ACF plot**

acf(Z)



As can be observed, the autocorrelation in the time series is significant from lags 1 to 2. Hence, we can say that there is significant autocorrelation between observations separated by one to two months. Such an ACF plot suggests that the data is not stationary, because the autocorrelation between observations, even those two months apart, is significant and change with in a relatively defined manner with respect to lag, which indicates that the data may show certain clear movements over time, such as trend.

## Seasonality-removed time series stationarity

adf.test(Z1)

**Warning in adf.test(Z2):** p-value smaller than printed p-value

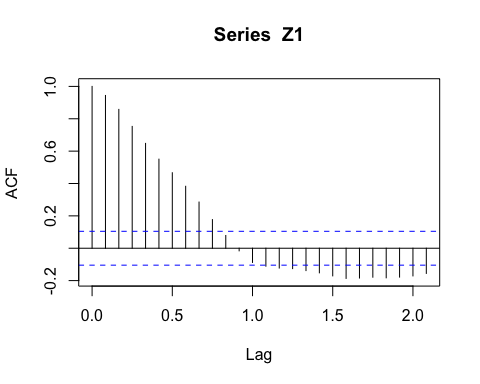
**Augmented Dickey-Fuller Test**

**data:** Z1  
***Dickey-Fuller*** = -5.7538, ***Lag order*** = 7, ***p-value*** = 0.01  
**alternative hypothesis:** stationary

The p-value for lag order 7 is 0.01, which is below 0.05 i.e. the significance level. Hence, we may accept the reject hypothesis and conclude that the time series is stationary.

**Observing ACF plot**

acf(Z1)



As can be observed, the autocorrelation in the time series is insignificant to weakly significant for all lags from 1 to 2. Hence, we can say that there is no significant autocorrelation between observations separated by one or more months. Such an ACF plot suggests that the data is stationary, because the autocorrelation between observations is insignificant and change seemingly randomly with respect to lag, which indicates that the data does show certain clear movements over time, such as trend, indicating a white-noise quality and hence stationarity.

## Trend and seasonality removed time series stationarity

adf.test(Z2)

**Warning in adf.test(Z2):** p-value smaller than printed p-value

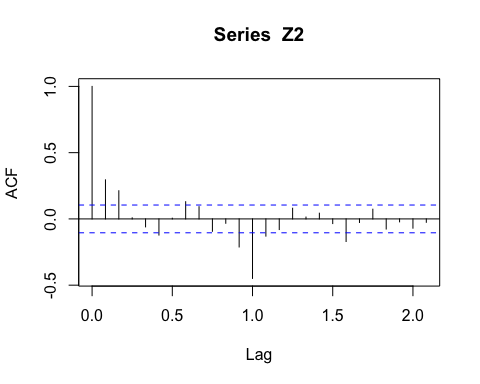
**Augmented Dickey-Fuller Test**

**data:** Z2  
***Dickey-Fuller*** = -5.3015, ***Lag order*** = 7, ***p-value*** = 0.01  
**alternative hypothesis:** stationary

The p-value for lag order 7 is 0.01, which is below 0.05 i.e. the significance level. Hence, we may accept the reject hypothesis and conclude that the time series is stationary. Furthermore, we see that the Dickey-Fuller statistic here is -5.3015, less negative than the Dickey-Fuller statistic of the time series where only seasonality was removed, whose value was -5.7538. This may suggest that trend does not play a key role in the time series, and hence, considering it to be constantly upwards does not improve the stationarity of our time series.

**Observing ACF plot**

acf(Z2)



As can be observed, the autocorrelation in the time series is insignificant to weakly significant for all lags from 1 to 2m with the clear exception at lag = 1. Hence, we can say that there is no significant autocorrelation between observations separated by more than one month, but there may be some autocorrelation between observations separated by a monthly interval. Such an ACF plot suggests that the data is stationary, but not as much as Z1, because the autocorrelation between observations is insignificant in most cases except lag = 1. However, autocorrelation changes seemingly randomly with respect to lag, which indicates that the data does show certain clear movements over time, such as trend, indicating a white-noise quality and hence stationarity.