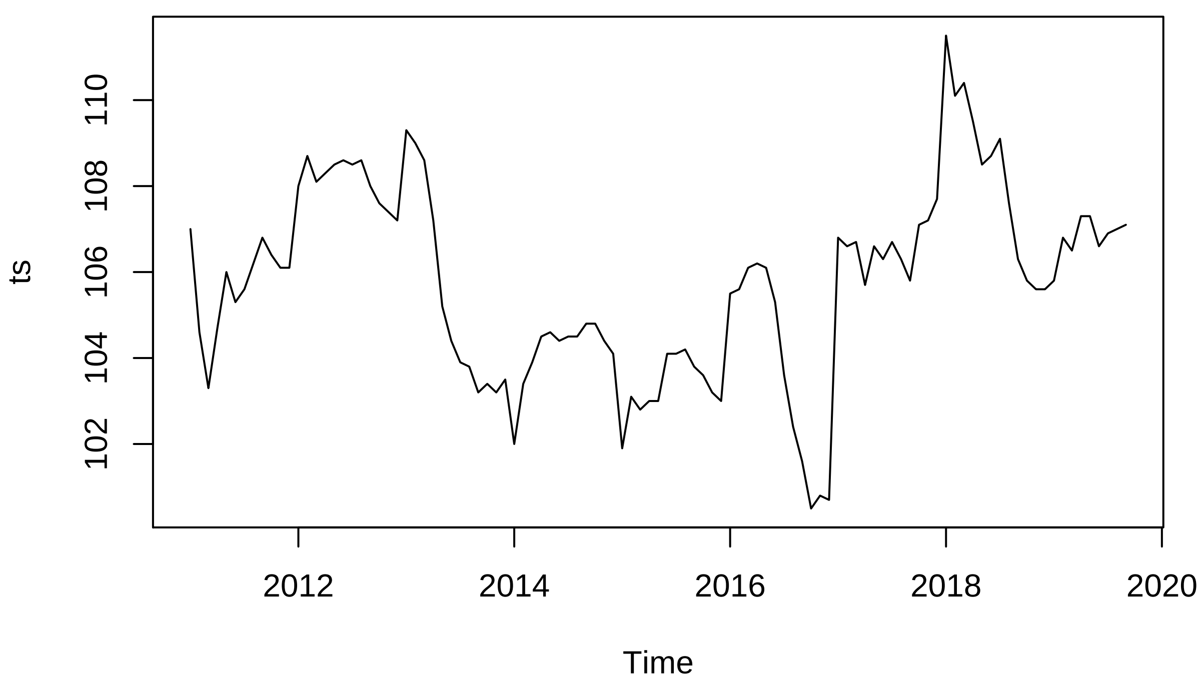
tsf third one

Kevork Sulahian

November 1, 2019

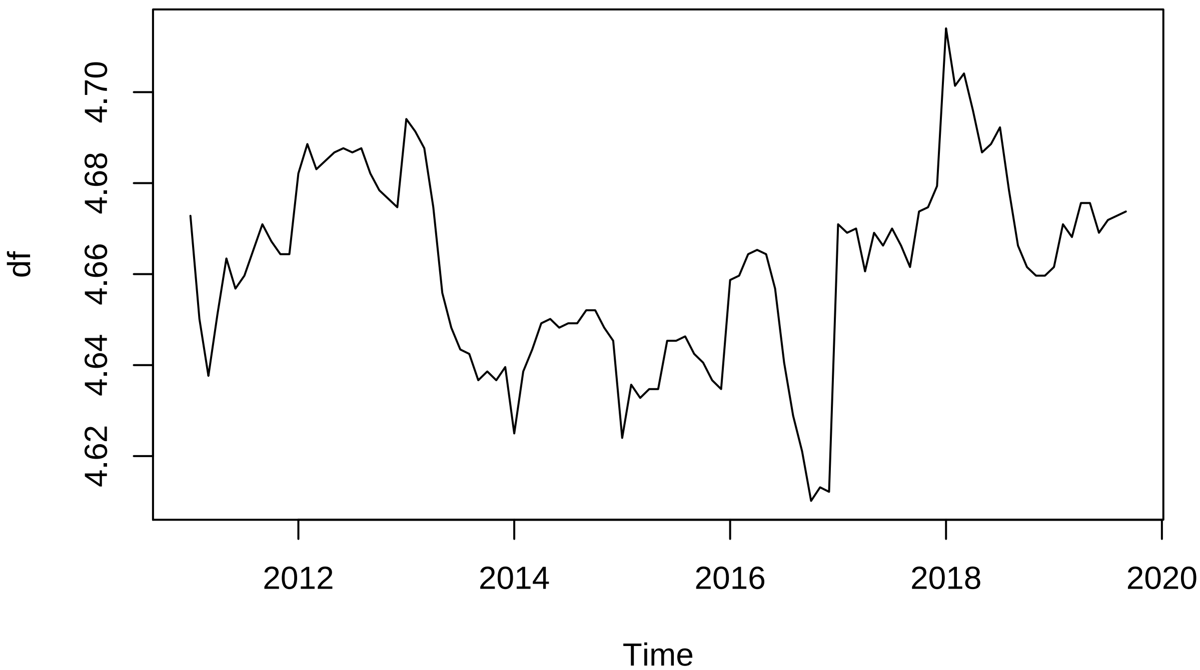
```r
library(readxl)
library(forecast)
```

```r
library(readxl)

df <- read_xlsx("ts-3.xlsx")

ts = ts(df,start=c(2011,1), frequency = c(12))
```



In this case, it appears that an additive model is not appropriate for describing this time series, since the size of the seasonal fluctuations and random fluctuations seem to increase with the level of the time series. Thus, we may need to transform the time series in order to get a transformed time series that can be described using an additive model. For example, we can transform the time series by calculating the natural log of the original data:

```r
log_ts <- log(ts)
plot.ts(log_ts)
```



Decomposing Time Series

Decomposing a time series means separating it into its constituent components, which are usually a trend component and an irregular component, and if it is a seasonal time series, a seasonal component.

Decomposing Seasonal Data

A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components: that is, estimating these three components.

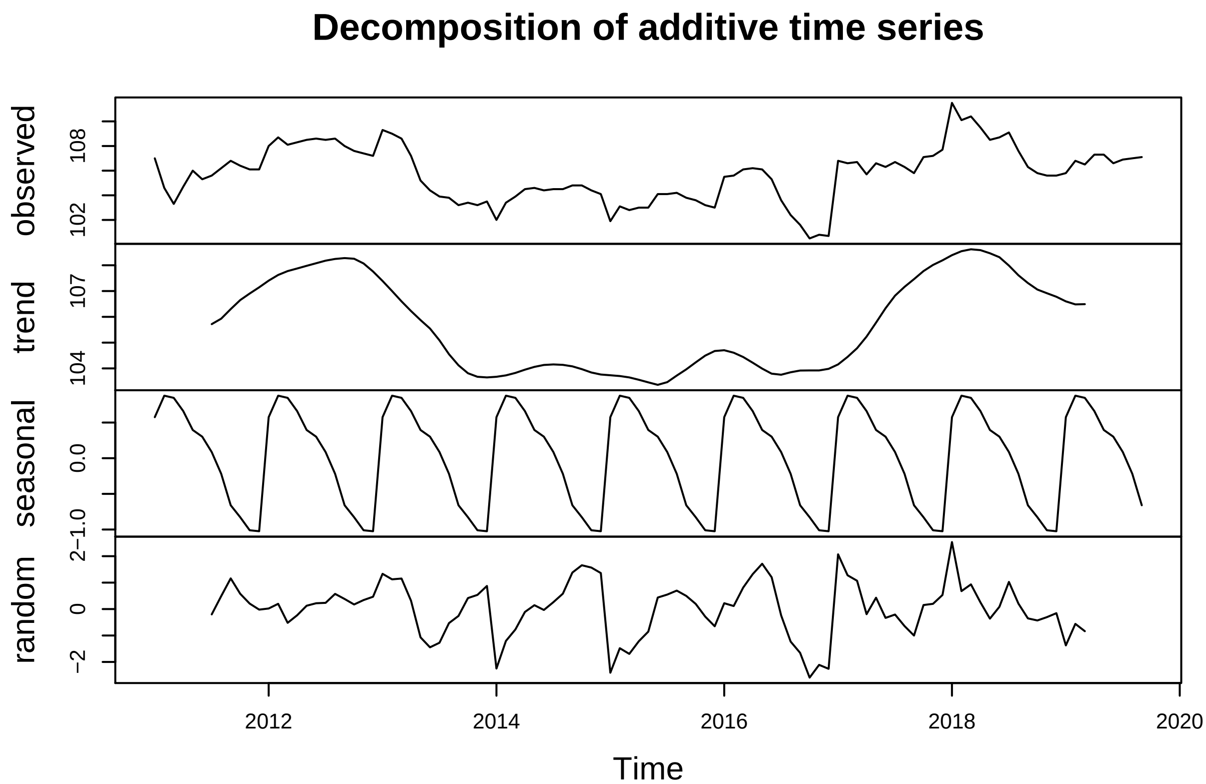
```
ts_components <- decompose(ts)
```

we can print out the estimated values of the seasonal component

```
ts_components$seasonal
```

```
##            Jan       Feb       Mar       Apr       May       Jun
## 2011  0.5750496 0.8766121 0.8458829 0.6603919 0.3955109 0.3014633
## 2012  0.5750496 0.8766121 0.8458829 0.6603919 0.3955109 0.3014633
## 2013  0.5750496 0.8766121 0.8458829 0.6603919 0.3955109 0.3014633
## 2014  0.5750496 0.8766121 0.8458829 0.6603919 0.3955109 0.3014633
## 2015  0.5750496 0.8766121 0.8458829 0.6603919 0.3955109 0.3014633
```

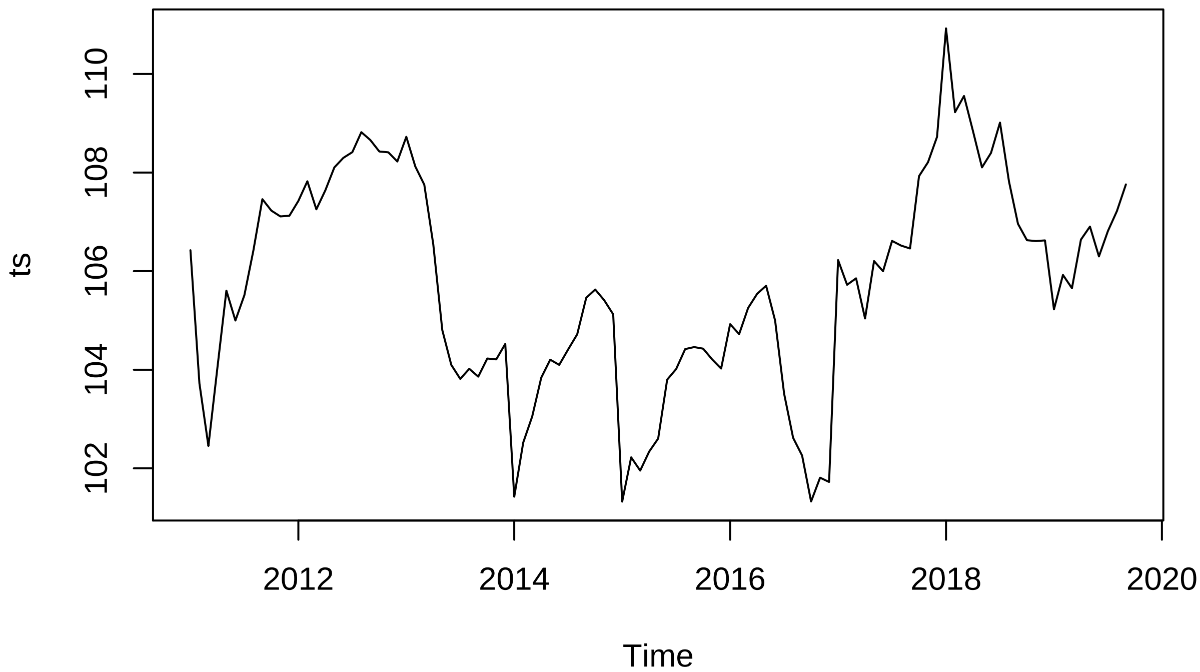
```
## 2016  0.5750496  0.8766121  0.8458829  0.6603919  0.3955109  0.3014633
## 2017  0.5750496  0.8766121  0.8458829  0.6603919  0.3955109  0.3014633
## 2018  0.5750496  0.8766121  0.8458829  0.6603919  0.3955109  0.3014633
## 2019  0.5750496  0.8766121  0.8458829  0.6603919  0.3955109  0.3014633
##              Jul        Aug        Sep        Oct        Nov        Dec
## 2011  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2012  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2013  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2014  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2015  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2016  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2017  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2018  0.0859871 -0.2192212 -0.6598462 -0.8275546 -1.0103671 -1.0239087
## 2019  0.0859871 -0.2192212 -0.6598462
```



The plot above shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom)

Seasonally Adjusting

```
ts_seasonall <- ts - ts_components$seasonal
```



Holt-Winters Exponential Smoothing

```r
ts_forcaste <- HoltWinters(ts)
ts_forcaste
```

```
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = ts)
##
## Smoothing parameters:
##  alpha: 0.9411093
##  beta : 0.01283223
##  gamma: 0.897372
##
## Coefficients:
##             [,1]
## a   107.12241937
## b     0.08333205
## s1   -0.30721959
## s2   -0.85329682
## s3   -1.15373938
## s4    0.46560241
```

```
## s5    0.77946235
## s6    0.35089672
## s7    0.15409626
## s8    0.16395033
## s9    0.19235885
## s10  -0.17674411
## s11  -0.20105992
## s12  -0.02136745
```

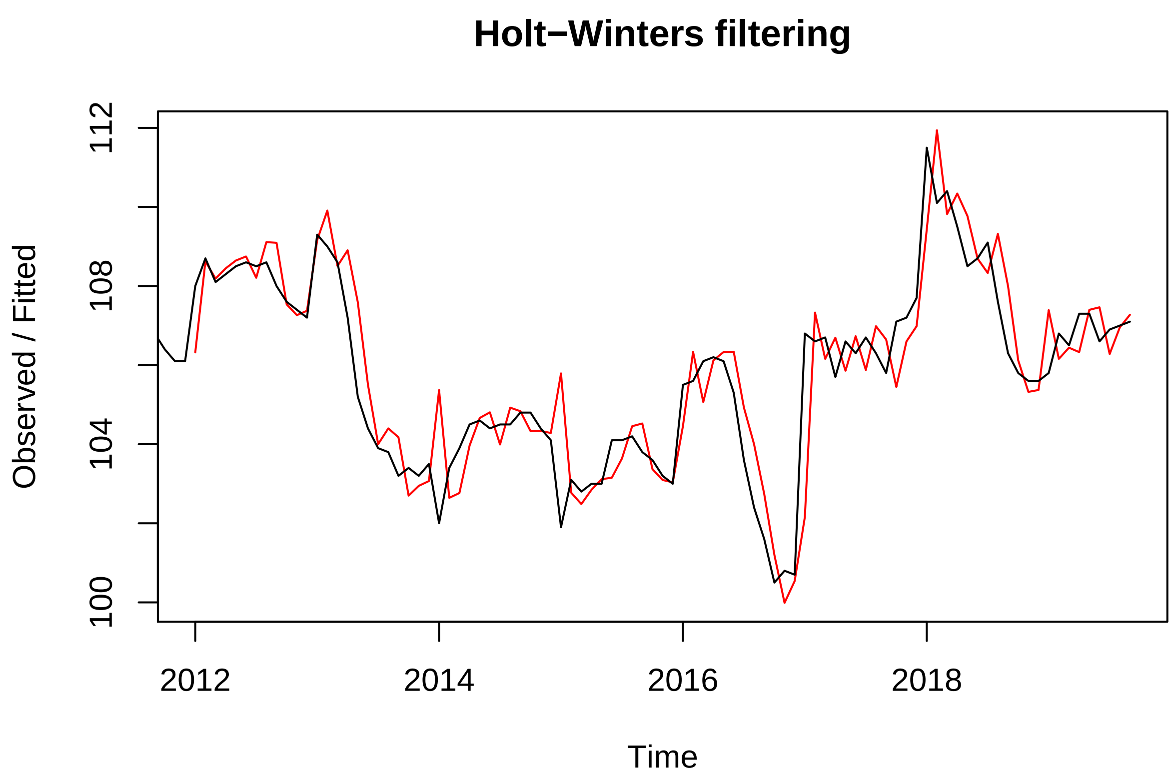
```
#
```

this is not true i need to change

The value of alpha (0.35) is relatively low, indicating that the estimate of the level at the current time point is based upon both recent observations and some observations in the more distant past. The value of beta is 0.01, indicating that the estimate of the slope b of the trend component is updated but doesn't have much effect over the time series, and instead is set equal to its initial value. This makes good intuitive sense, as the level changes quite a bit over the time series, but the slope b of the trend component remains roughly the same. In contrast, the value of gamma (0.38) is high, indicating that the estimate of the seasonal component at the current time point is not just based upon very recent observations

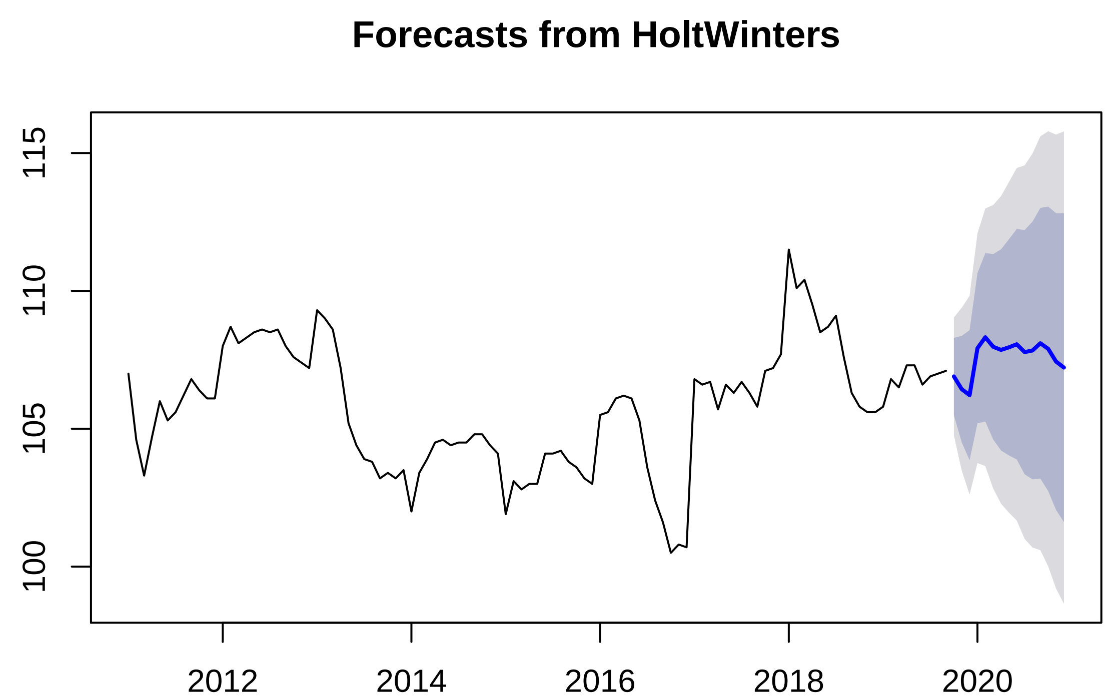
```
ts_forcaste$SSE
```

```
## [1] 110.9689
```



```
ts_forcaste2 = forecast:::forecast.HoltWinters(ts_forcaste, h= 15)
(as.data.frame(ts_forcaste2))[1]
```

```
##          Point Forecast
## Oct 2019       106.8985
## Nov 2019       106.4358
## Dec 2019       106.2187
## Jan 2020       107.9213
## Feb 2020       108.3185
## Mar 2020       107.9733
## Apr 2020       107.8598
## May 2020       107.9530
## Jun 2020       108.0648
## Jul 2020       107.7790
## Aug 2020       107.8380
## Sep 2020       108.1010
## Oct 2020       107.8985
## Nov 2020       107.4358
## Dec 2020       107.2187
```



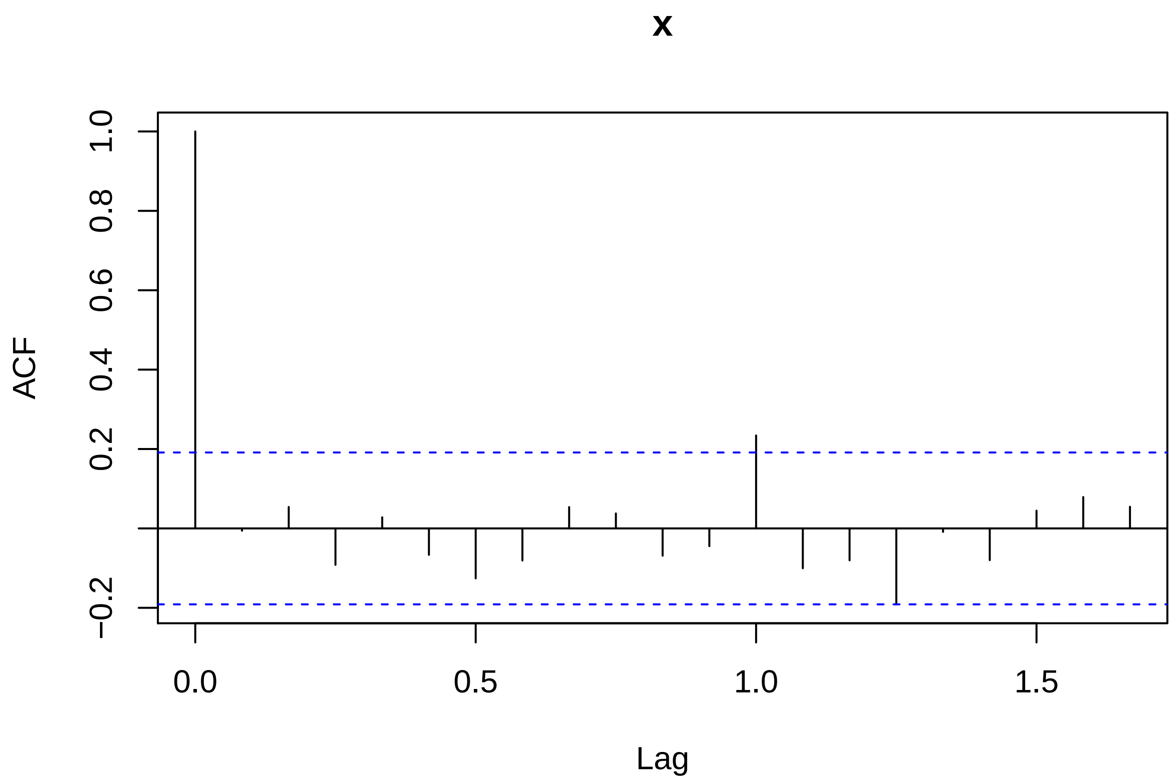
Growth

```
year_2019 <- window(ts, 2019)
year_2019_predict_HW <- (as.data.frame(ts_forcaste2))[1][c(1:3),]
sum_year_2019 = sum(c(year_2019,year_2019_predict_HW))
year_2020 = (as.data.frame(ts_forcaste2))[1][c(4:15),]
```

```r
growth_HW <- growth(sum(year_2020),sum_year_2019)
growth_HW
```

```
## [1] 0.01054675
```

We can investigate whether the predictive model can be improved upon by checking whether the in-sample forecast errors show non-zero autocorrelations at lags 1-20, by making a correlogram and carrying out the Ljung-Box test:



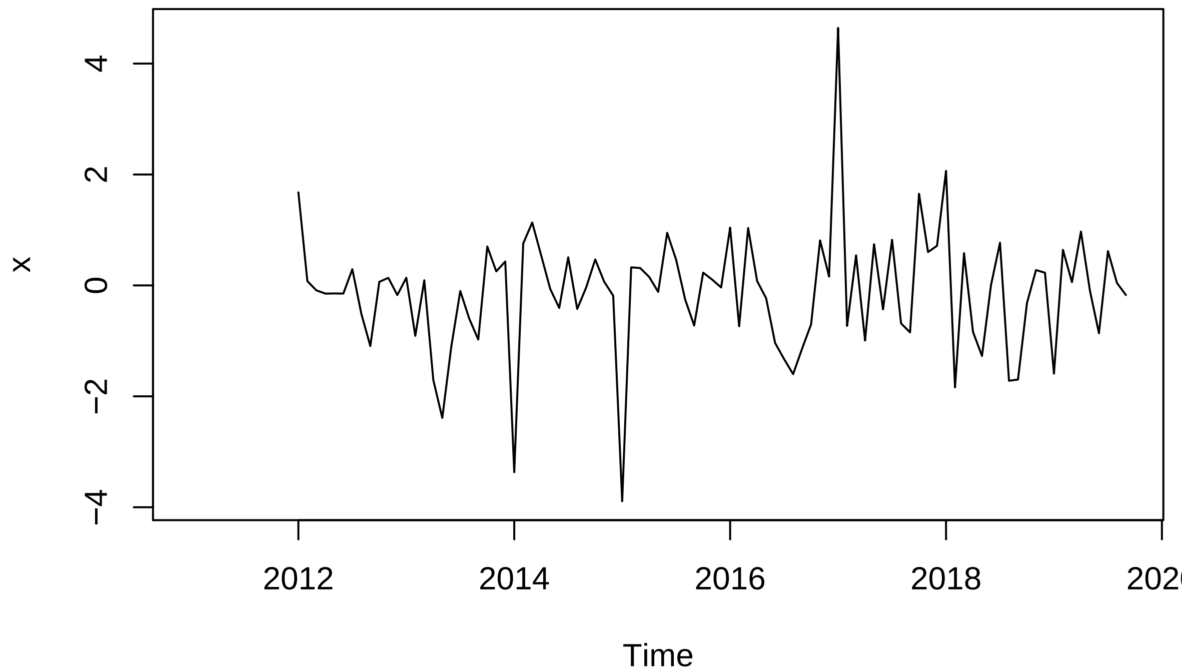

```
##
##  Box-Ljung test
##
## data:  ts_forcaste2$residuals
## X-squared = 19.024, df = 20, p-value = 0.5203
```

p-value is 0.5 instead of 0.9

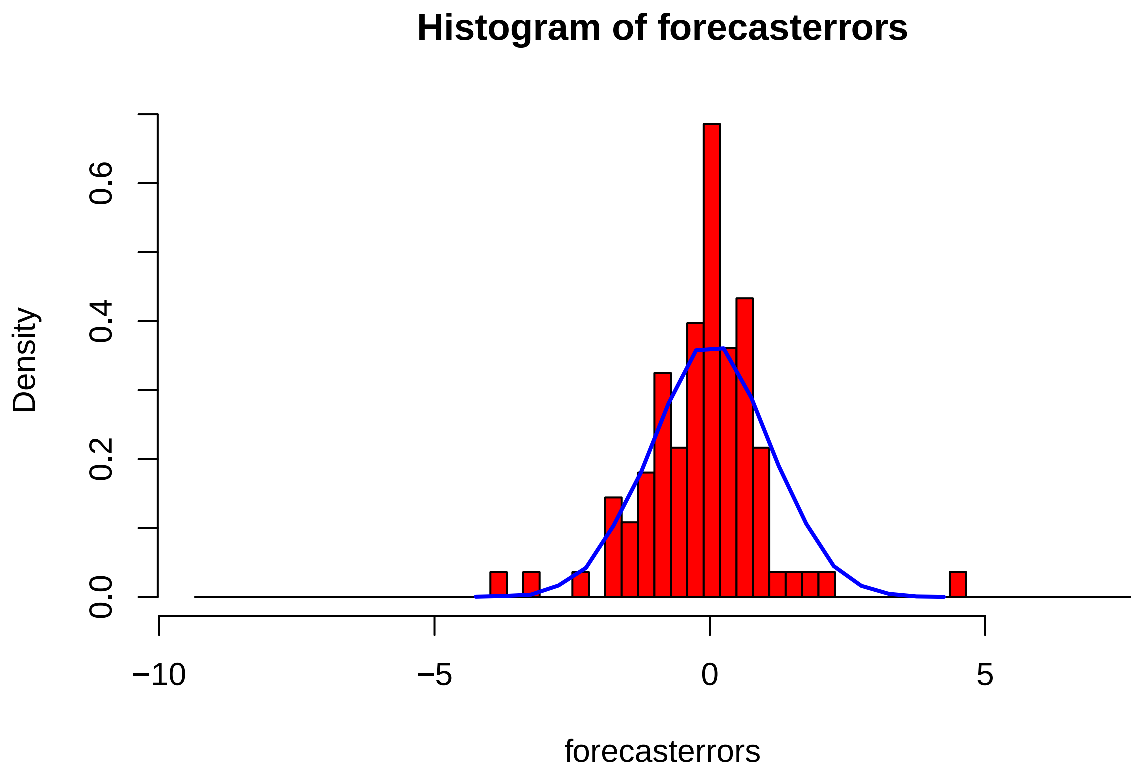
The correlogram shows that the autocorrelations for the in-sample forecast errors do not exceed the significance bounds for lags 1-20. Furthermore, the p-value for Ljung-Box test is 0.9, indicating that there is no evidence of non-zero autocorrelations at lags 1-20.

We can check whether the forecast errors have constant variance over time, and are normally distributed with mean zero, by making a time plot of the forecast errors and a histogram (with overlaid normal curve):

```r
plot.ts(ts_forcaste2$residuals)
```



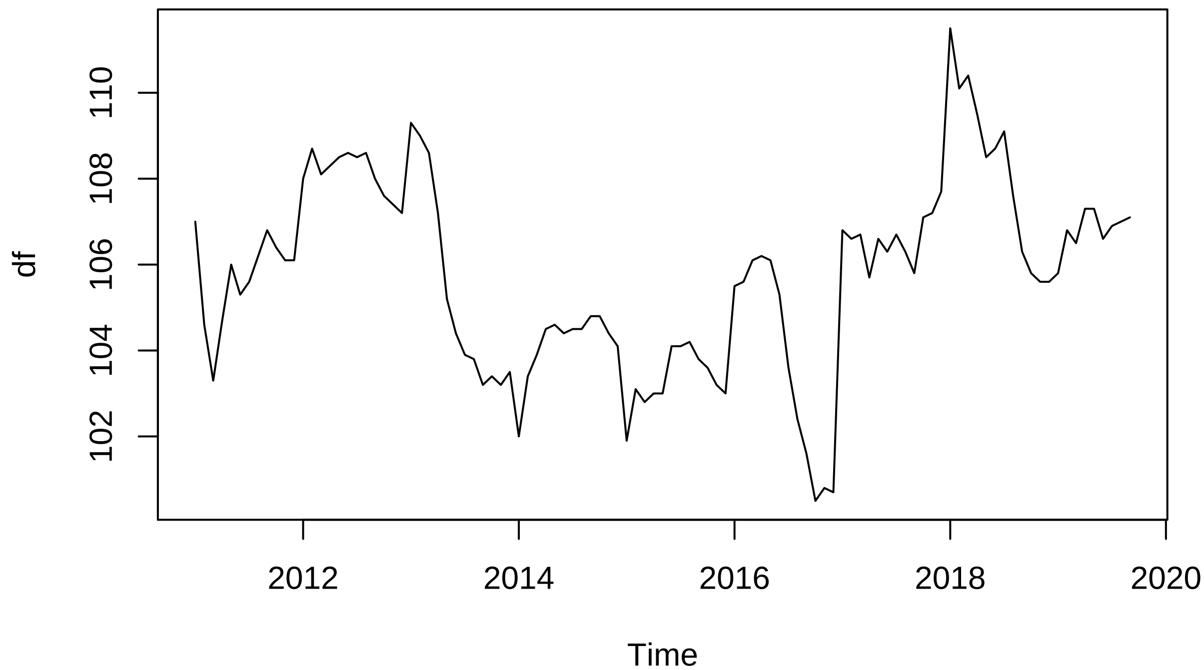

```r
plotForecastErrors(ts_forcaste2$residuals)
```



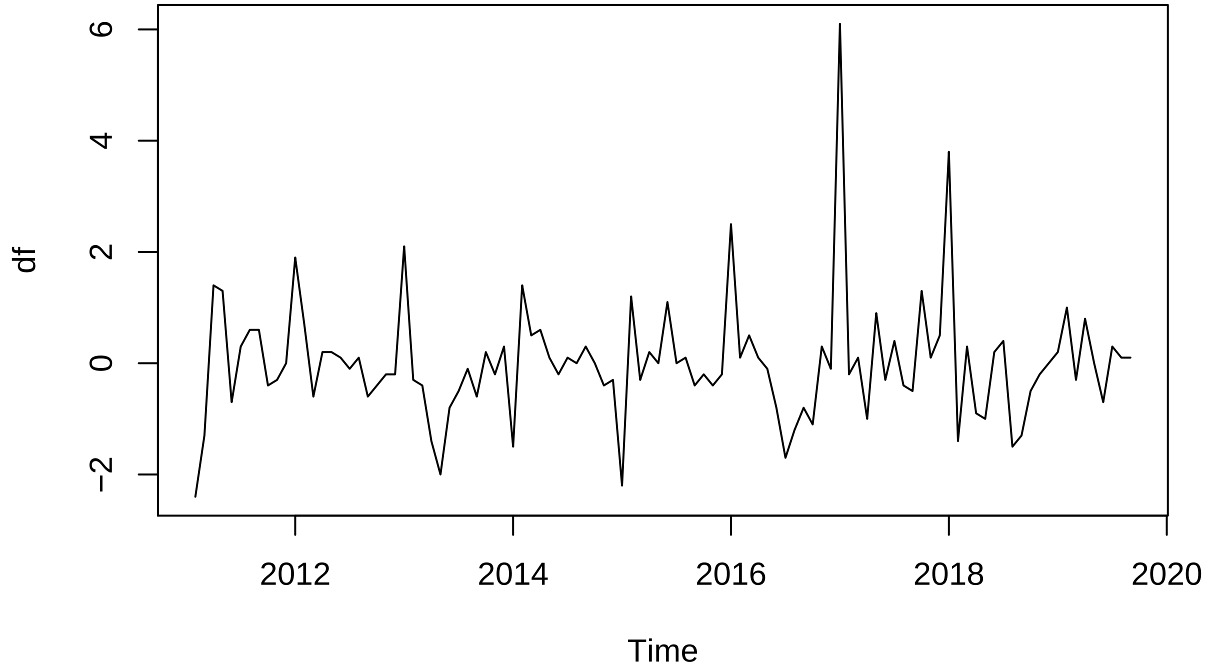
From the time plot, it appears plausible that the forecast errors have constant variance over time. From the histogram of forecast errors, it seems plausible that the forecast errors are normally distributed with mean zero.

Thus,there is little evidence of autocorrelation at lags 1-20 for the forecast errors, and the forecast errors appear to be normally distributed with mean zero and constant variance over time. This suggests that Holt-Winters exponential smoothing provides an adequate predictive model of the log of total productivity, which probably cannot be improved upon. Furthermore, the assumptions upon which the prediction intervals were based are probably valid.

```
plot.ts(ts)
```

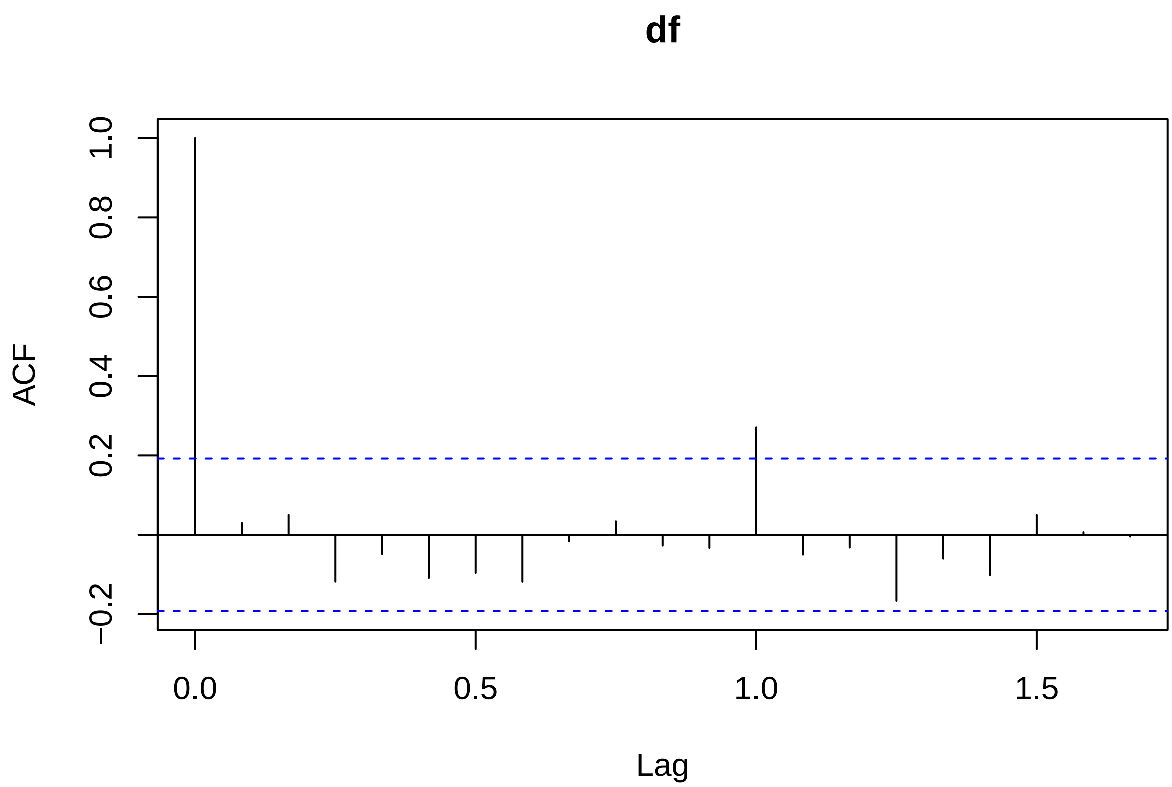



```
ts_diff1 <-  diff(ts, differences = 1)

plot.ts(ts_diff1)
```



The time series of differences (above) does appear to be stationary in mean and variance, as the level of the series stays roughly constant over time, and the variance of the series appears roughly constant over time

```
acf(ts_diff1, lag.max=20)              # plot a correlogram
```

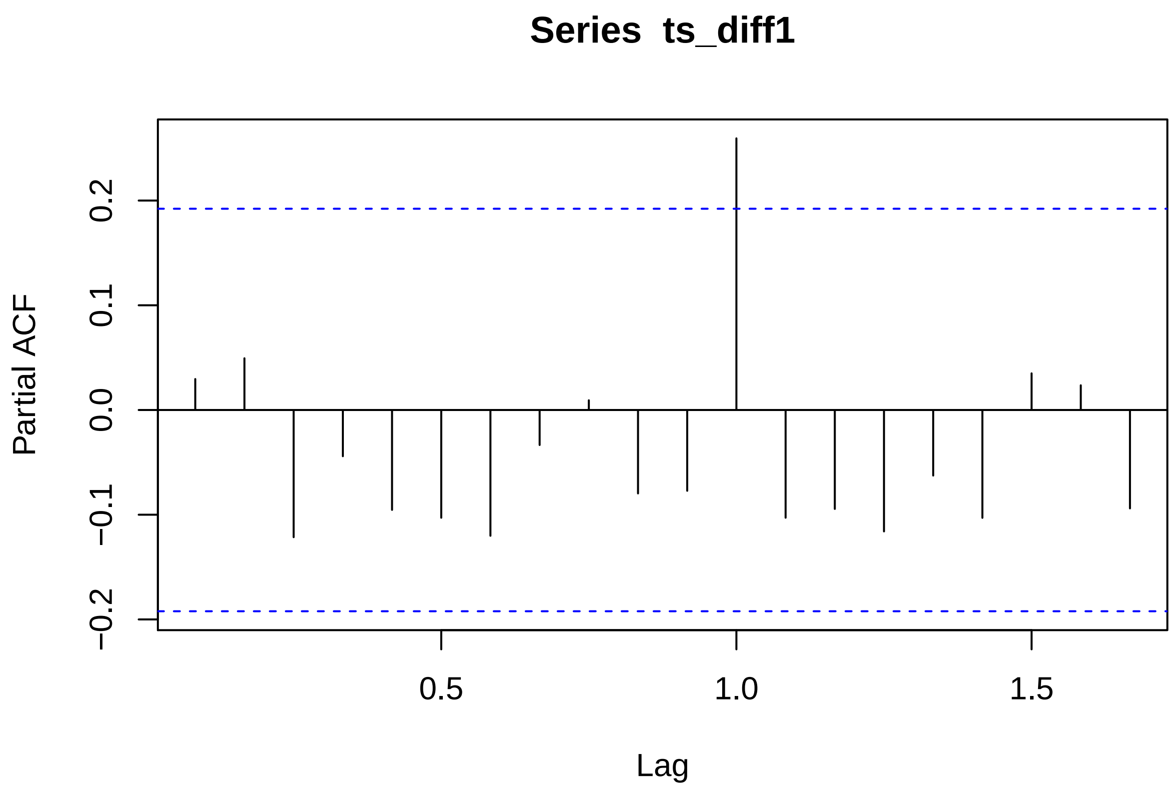


We see from the correlogram that the autocorrelation exceeds the significance bound 3 times but all the others do not exceed

```
acf(ts_diff1, lag.max=20, plot=FALSE) # get the autocorrelation values
```

```
##
## Autocorrelations of series 'ts_diff1', by lag
##
## 0.0000 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500
##  1.000  0.030  0.050 -0.118 -0.048 -0.109 -0.096 -0.118 -0.016  0.034
## 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833
## -0.027 -0.033  0.271 -0.050 -0.032 -0.166 -0.060 -0.101  0.050  0.006
## 1.6667
## -0.004
```

```
pacf(ts_diff1, lag.max=20)             # plot a partial correlogram
```




```r
pacf(ts_diff1, lag.max=20, plot=FALSE) # get the partial autocorrelation values
```

```
##
## Partial autocorrelations of series 'ts_diff1', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333
##  0.030  0.049 -0.121 -0.044 -0.095 -0.103 -0.120 -0.033  0.009 -0.080
## 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667
## -0.077  0.259 -0.103 -0.094 -0.116 -0.063 -0.103  0.035  0.024 -0.094
```

Arima, 0,1,0

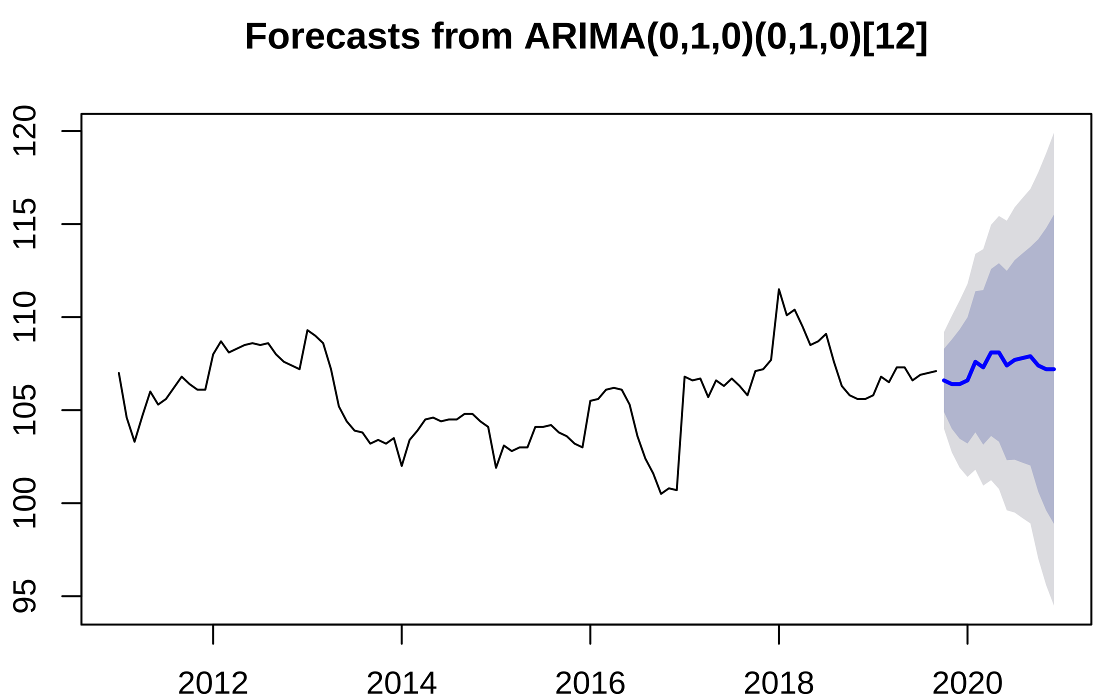
```r
ts_arima = Arima(ts, order=c(0,1,0),seasonal = list(order = c(0,1,0)))
ts_arima
```

```
## Series: ts
## ARIMA(0,1,0)(0,1,0)[12]
##
## sigma^2 estimated as 1.752:  log likelihood=-156.28
## AIC=314.57   AICc=314.61   BIC=317.09
```

```r
ts_arima_forecast = forecast(ts_arima,h = 15)
ts_arima_forecast
```

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Oct 2019          106.6 104.90392 108.2961 104.00607 109.1939
## Nov 2019          106.4 104.00138 108.7986 102.73162 110.0684
## Dec 2019          106.4 103.46230 109.3377 101.90717 110.8928
## Jan 2020          106.6 103.20783 109.9922 101.41213 111.7879
## Feb 2020          107.6 103.80744 111.3926 101.79979 113.4002
## Mar 2020          107.3 103.14546 111.4545 100.94618 113.6538
## Apr 2020          108.1 103.61259 112.5874 101.23709 114.9629
## May 2020          108.1 103.30275 112.8972 100.76325 115.4368
## Jun 2020          107.4 102.31175 112.4882  99.61820 115.1818
## Jul 2020          107.7 102.33652 113.0635  99.49726 115.9027
## Aug 2020          107.8 102.17473 113.4253  99.19689 116.4031
## Sep 2020          107.9 102.02460 113.7754  98.91435 116.8857
## Oct 2020          107.4 100.61567 114.1843  97.02426 117.7757
## Nov 2020          107.2  99.61489 114.7851  95.59957 118.8004
## Dec 2020          107.2  98.89093 115.5091  94.49237 119.9076
```

```r
forecast:::plot.forecast(ts_arima_forecast)
```



Growth

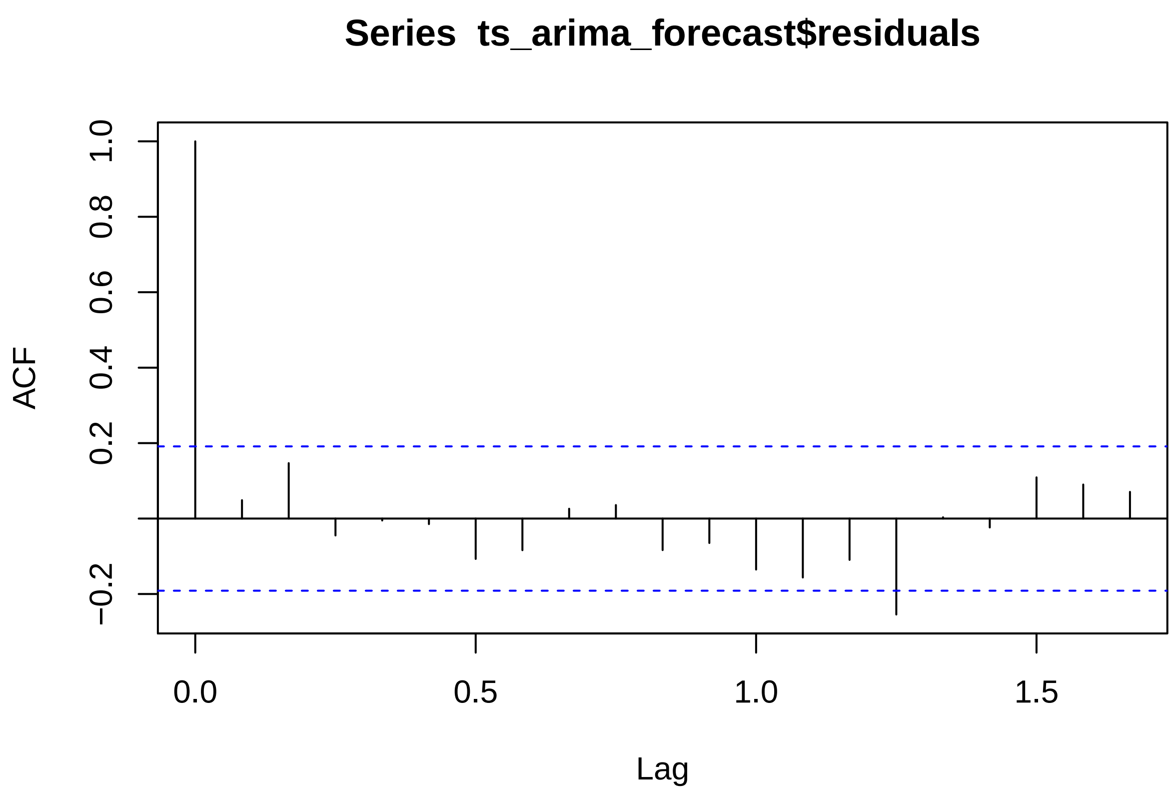
```r
this_year_predict_ARIMA <- (as.data.frame(ts_arima_forecast))[1]


year_2019_predict_ARIMA <- (as.data.frame(ts_arima_forecast))[1][c(1:3),]
sum_year_2019 = sum(c(year_2019,year_2019_predict_ARIMA))
year_2020 = (as.data.frame(ts_arima_forecast))[1][c(4:15),]
growth_ARIMA <- growth(sum_year_2019, sum(year_2020))
-growth_ARIMA
```

```
## [1] 0.00744013
```

As in the case of exponential smoothing models, it is a good idea to investigate whether the forecast errors of an ARIMA model are normally distributed with mean zero and constant variance, and whether the are correlations between successive forecast errors.

For example, we can make a correlogram of the forecast errors for our ARIMA(0,1,1) model, and perform the Ljung-Box test for lags 1-20, by typing:

```r
acf(ts_arima_forecast$residuals, lag.max=20)
```

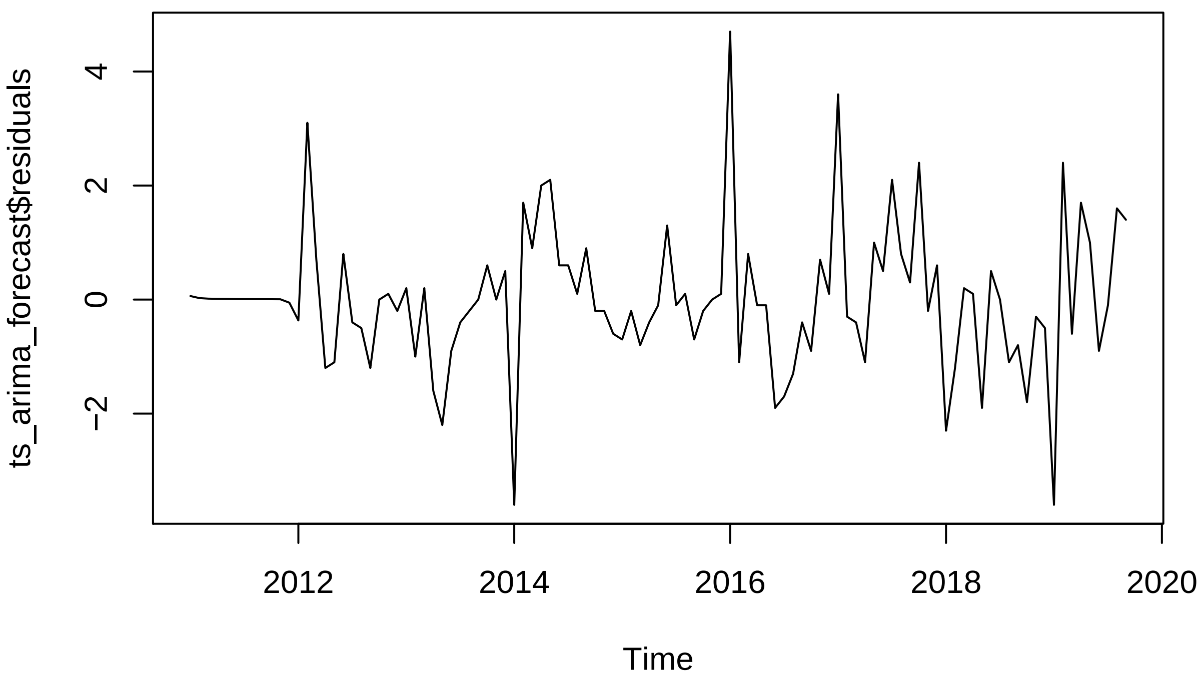


```r
Box.test(ts_arima_forecast$residuals, lag=20, type="Ljung-Box")
```

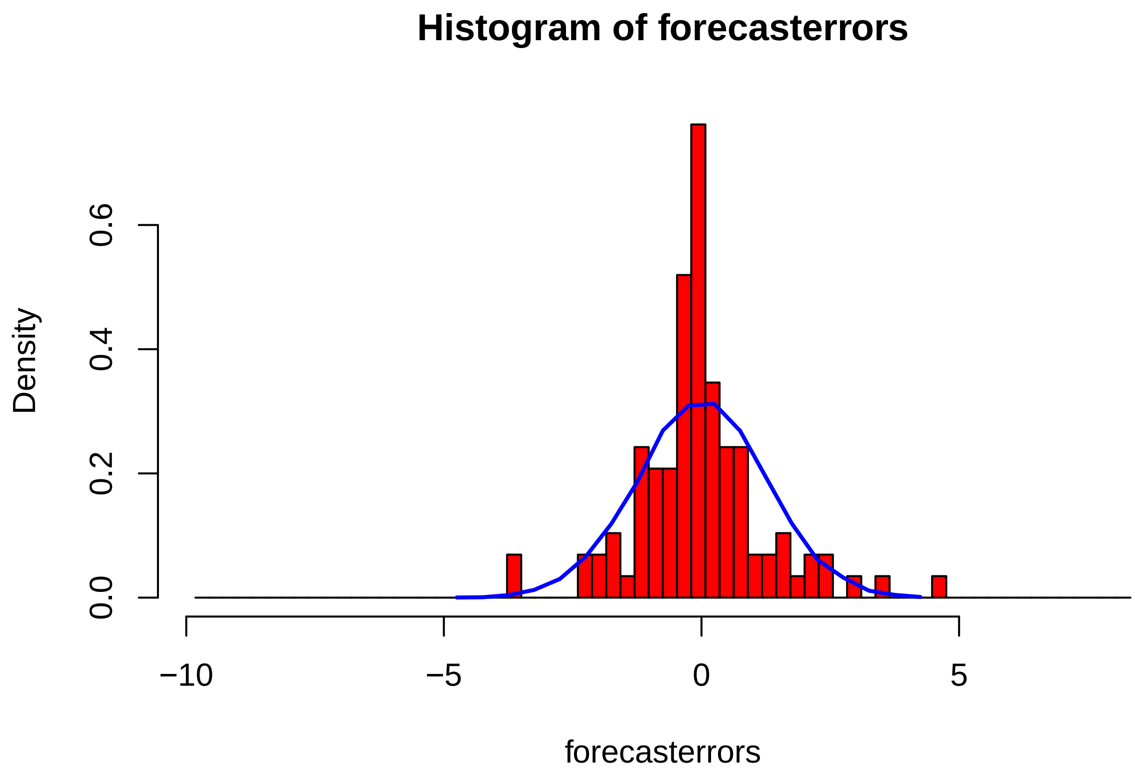
```
##
##  Box-Ljung test
##
## data:  ts_arima_forecast$residuals
## X-squared = 24.615, df = 20, p-value = 0.2166
```

p value too high to reject

```r
plot.ts(ts_arima_forecast$residuals)            # make time plot of forecast errors
```




```r
plotForecastErrors(ts_arima_forecast$residuals)
```

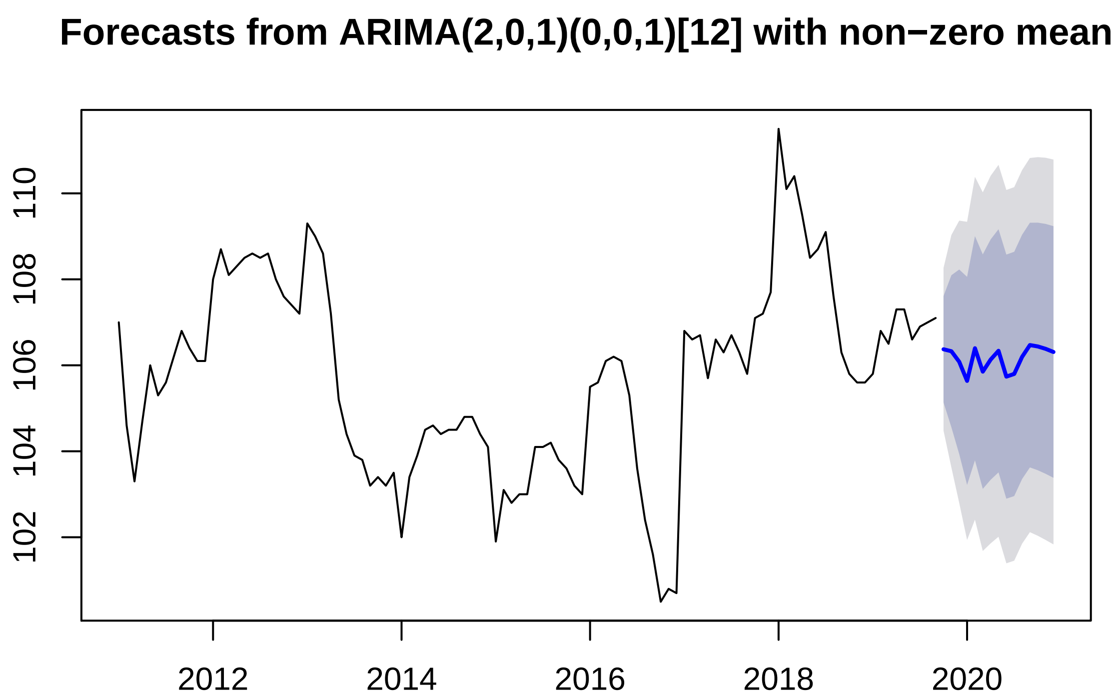


A model chosen automatically

```
fit3 <- auto.arima(ts, seasonal = T)
fit3
```

```
## Series: ts
## ARIMA(2,0,1)(0,0,1)[12] with non-zero mean
##
## Coefficients:
##          ar1      ar2      ma1    sma1      mean
##       1.7778  -0.8309  -0.7565  0.5550  105.8879
## s.e.  0.1010   0.0908   0.1269  0.1205    0.6375
##
## sigma^2 estimated as 0.9316:  log likelihood=-145.54
## AIC=303.08   AICc=303.94   BIC=319.01
```

```
fit_forecast = forecast(fit3,h=15)
plot(fit_forecast)
```




```
# str(fit)
```

Growth

```
year_2019_predict_auto.arima <- (as.data.frame(fit_forecast))[1][c(1:3),]
year_2019_predict_auto.arima_95_low <- (as.data.frame(fit_forecast))[4][c(1:3),]
year_2019_predict_auto.arima_95_high <- (as.data.frame(fit_forecast))[5][c(1:3),]

sum_year_2019 = sum(c(year_2019,year_2019_predict_auto.arima))
sum_year_2019_low = sum(c(year_2019,year_2019_predict_auto.arima_95_low))
sum_year_2019_high = sum(c(year_2019,year_2019_predict_auto.arima_95_high))


year_2020_predict_auto.arima <- (as.data.frame(fit_forecast))[1][c(4:15),]
year_2020_predict_auto.arima_95_low <- (as.data.frame(fit_forecast))[4][c(4:15),]
year_2020_predict_auto.arima_95_high <- (as.data.frame(fit_forecast))[5][c(4:15),]

growth_auto.arima <- growth(sum(year_2020_predict_auto.arima),sum_year_2019)
growth_auto.arima_95_low <- growth(sum(year_2020_predict_auto.arima_95_low),sum_year_2019_low)
growth_auto.arima_95_high <- growth(sum(year_2020_predict_auto.arima_95_high),sum_year_2019_high)

growth_auto.arima
```

```
## [1] -0.004994184
```

```r
growth_auto.arima_95_low
```

```
## [1] -0.03904977
```

```r
growth_auto.arima_95_high
```

```
## [1] 0.02864472
```

```r
-growth_ARIMA
```

```
## [1] 0.00744013
```

```r
growth_auto.arima
```

```
## [1] -0.004994184
```

```r
growth_HW
```

```
## [1] 0.01054675
```