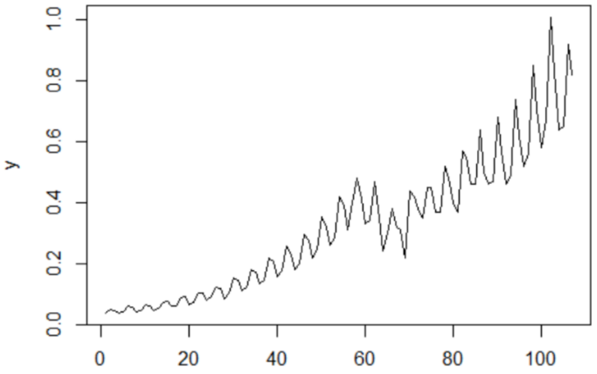
Forecasting Time Series

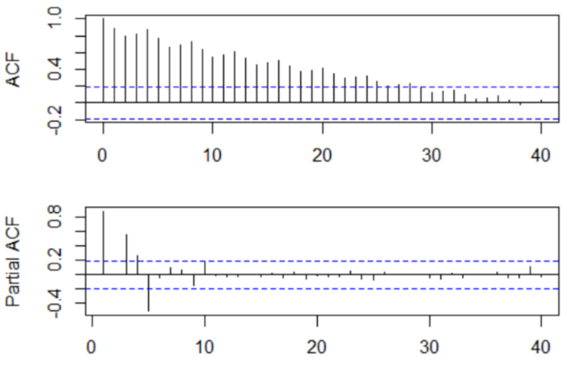
Homework 2

Group D

1. Find at least two linear time series models, using the Box-Jenkins methodology, for the quarterly earnings per share of Coca-Cola Company from the first quarter of 1983 to the third quarter of 2009. Identify your models using the entire available sample (coca_cola_earnings.csv)

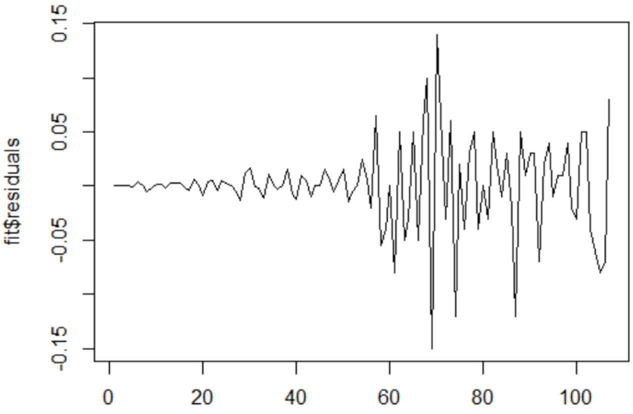
First, we plot the data and its corresponding ACF and Partial ACF to check for stationary.





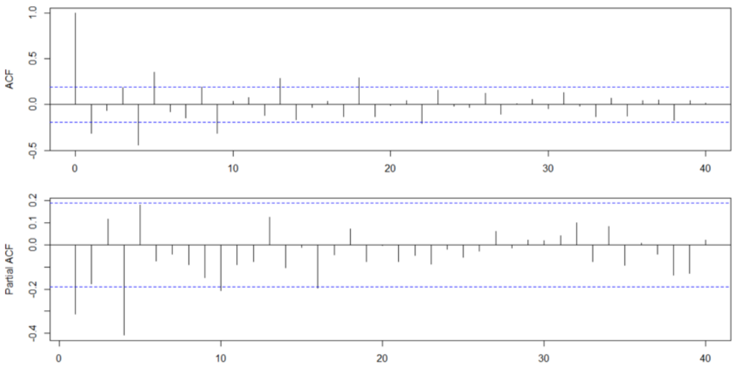
We can infer the data is not stationary, so we proceed to check how many differences we must do in order to have stationary data.

When running the dickey fuller test, the results indicate that we must take 1 regular difference and 1 seasonal difference.



After taken both differences we can see from the above graph that we have stationary data.

Now we have a look to the ACF and Partial ACF to decide which possible models could fit our data.



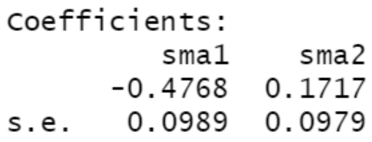
Considering the plots, we can suggest 4 possible models for our data:

- (16,1,0)x(0,1,0) S = 4
- (0,1,18)x(0,1,0) S = 4
- (0,1,0)x(0,1,2) S = 4
- (0,1,0)x(1,1,0) S = 4

However, because of the complexity of the first two models we decided to evaluate the last two models.

(0,1,0)x(0,1,2) S = 4

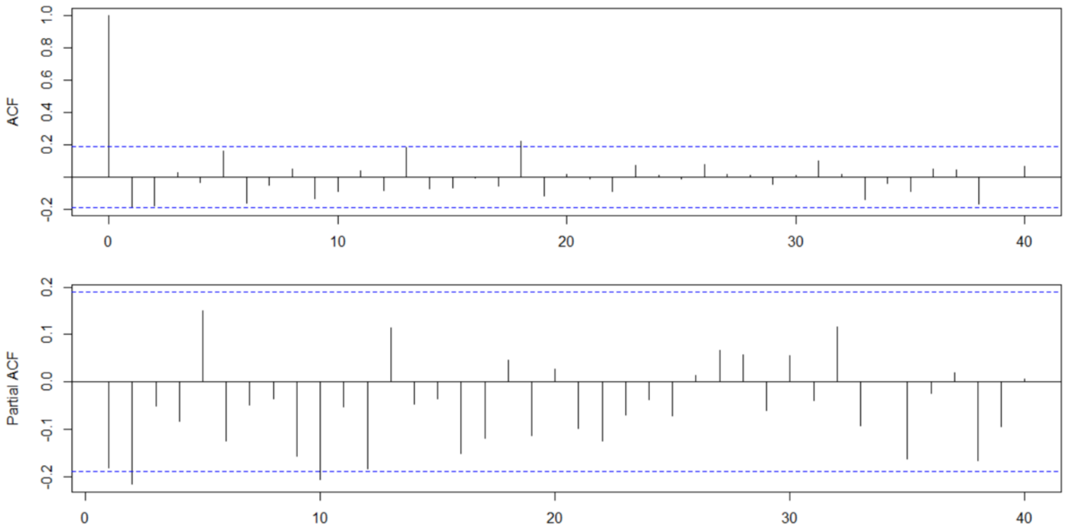
As a result, we get the following coefficients:



```
Coefficients:
         sma1    sma2
      -0.4768  0.1717
s.e.   0.0989  0.0979
```

For the first one we can easily conclude that is significant, for the second coefficient we can observe that is in the borderline, so we continue with the model

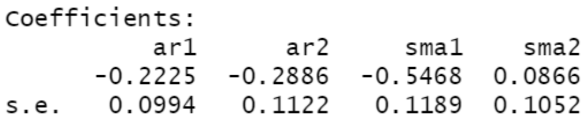
We again check for the ACF and Partial ACF for the residuals.



Because we still have lags out of limits, we decide to prove a new model. Looking to the Partial ACF, the second lag suggest that the following model might be better model:

(2,1,0)x(0,1,2) S = 4

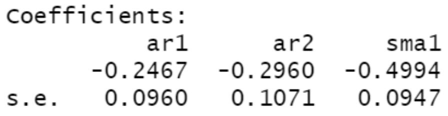
Because of the multiplication of the polynomial we are not considering lag 10.



```
Coefficients:
          ar1      ar2     sma1    sma2
      -0.2225  -0.2886  -0.5468  0.0866
s.e.   0.0994   0.1122   0.1189  0.1052
```

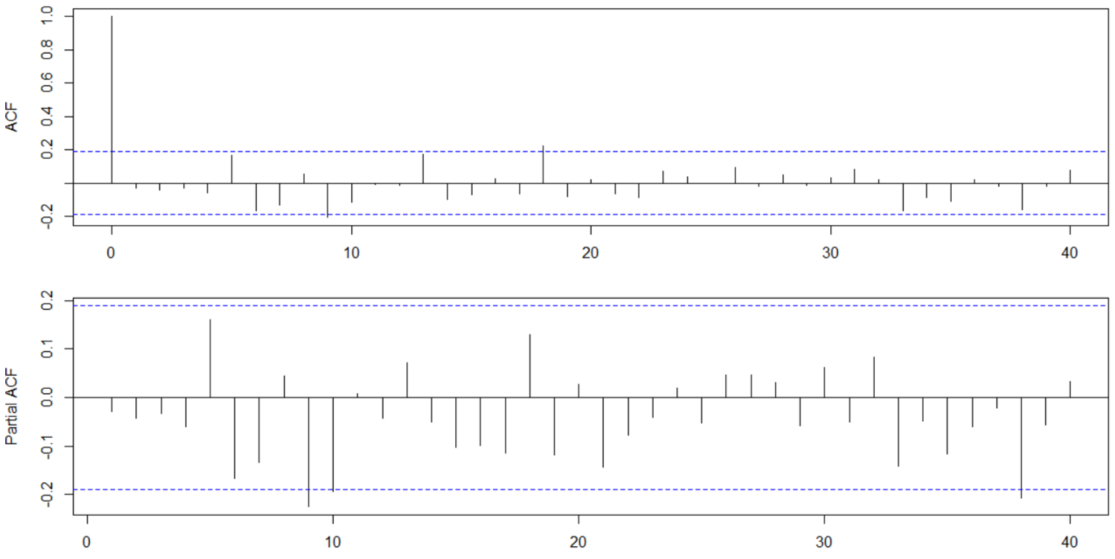
When checking for the coefficients we clearly see the last coefficient is not significant, so the model is discarded, and we proceed to try the same model but with a SMA(1).

(2,1,0)x(0,1,1) S = 4

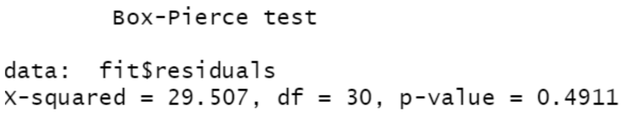


```
Coefficients:
          ar1      ar2     sma1
      -0.2467  -0.2960  -0.4994
s.e.   0.0960   0.1071   0.0947
```

When checking the coefficients, we see that all of them are significant, so we continue to evaluate the residuals to check for white noise



The ACF and PACF still show some lags out of limits, however those lacks implies to evaluate more complex AR or MA models, so we decide to continue with this model and run the Box test to check for white noise

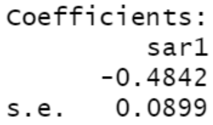


```
        Box-Pierce test

data:  fit$residuals
X-squared = 29.507, df = 30, p-value = 0.4911
```

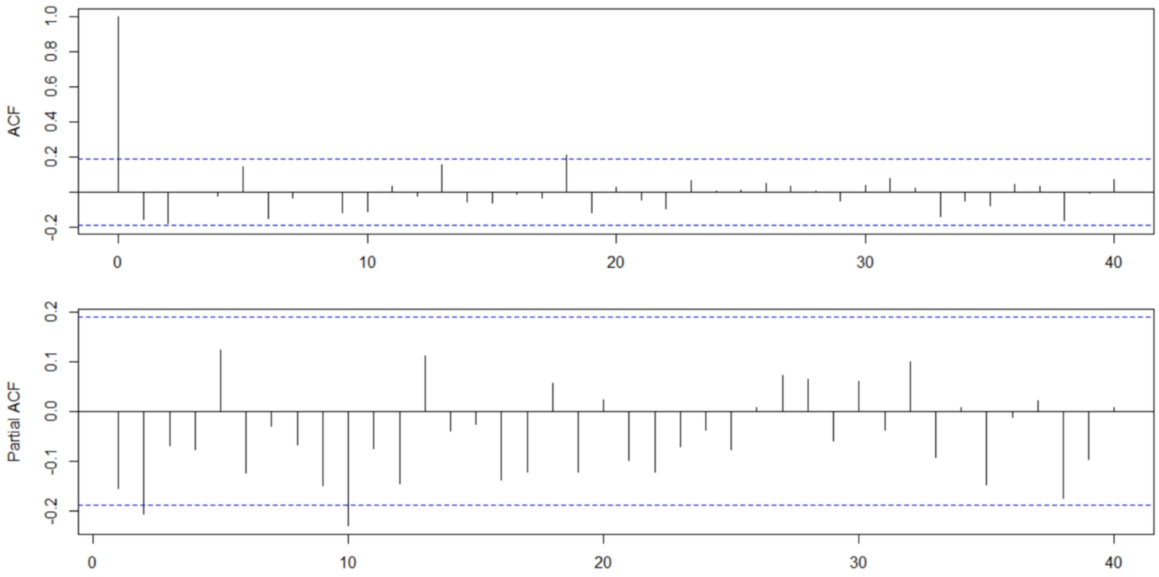
The test indicate that we do not reject the null hypothesis that means that we have white noise in the residuals, and this is a potential model to predict.

(0,1,0)x(1,1,0) S = 4



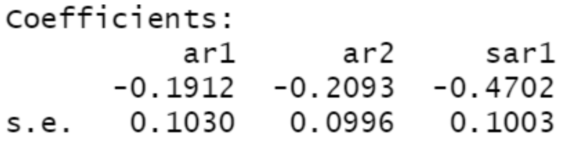
```
Coefficients:
         sar1
      -0.4842
s.e.   0.0899
```

When we look the coefficient of the model, we can say it is significant so we can proceed with the model.



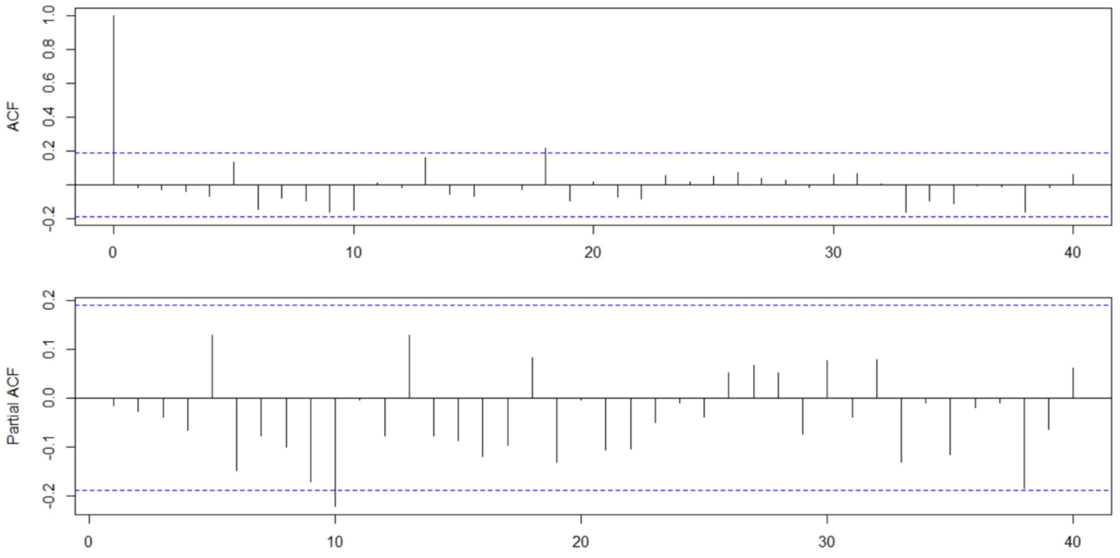
The residuals still do not show white noise, so we decide to test a new model. Considering the lags out of limit in the Partial ACF we decide to modify the model with a AR(2). Again, because of the multiplication of the polynomial we do not need to consider lag 10.

(2,1,0)x(1,1,0) S = 4

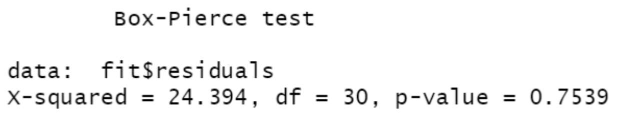


```
Coefficients:
          ar1      ar2     sar1
      -0.1912  -0.2093  -0.4702
s.e.   0.1030   0.0996   0.1003
```

When checking the coefficients, all of them are significant therefore we decide to continue with this model.



Looking at the ACF and Partial ACF we still have some lags out of limits; however, those lags will imply to run more complex models that probably are not as intuitive as the current one. For that reason, we finally run the formal test for white noise.



```
        Box-Pierce test

data:  fit$residuals
X-squared = 24.394, df = 30, p-value = 0.7539
```

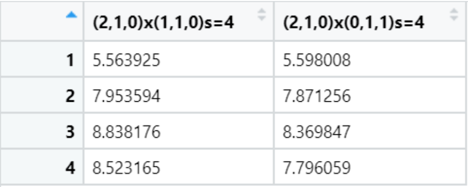
The test tells us that we have white noise for the residuals, so we conclude this is another potential model to predict.

2. For the models identified in the previous step, leave for example the last 24 real values to compare all the models in terms of forecasting (out of sample forecasting exercise). What is the best model and why is this your choice?

Recursive Method:

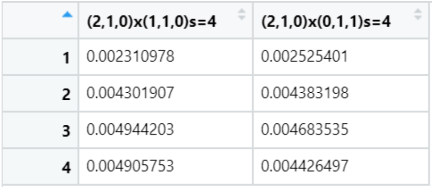
Using the Recursive Method, we calculate the Mean Absolute Percentage Error (MAPE) and the Mean Squared Predicted Error (MSPE)

MAPE



	(2,1,0)x(1,1,0)s=4	(2,1,0)x(0,1,1)s=4
1	5.563925	5.598008
2	7.953594	7.871256
3	8.838176	8.369847
4	8.523165	7.796059

MSPE

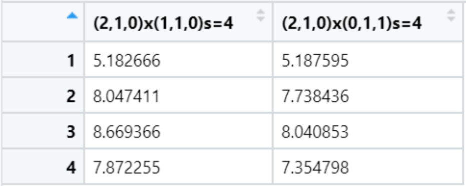


	(2,1,0)x(1,1,0)s=4	(2,1,0)x(0,1,1)s=4
1	0.002310978	0.002525401
2	0.004301907	0.004383198
3	0.004944203	0.004683535
4	0.004905753	0.004426497

Rolling Method:

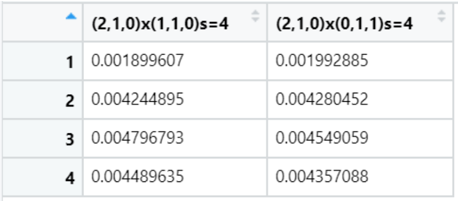
Using the Rolling Method, we again calculate the Mean Absolute Percentage Error (MAPE) and the Mean Squared Predicted Error (MSPE). In this occasion we tested every possible box size (from 20 to 78) to optimize the errors. We found that for the first model the optimal number of lags is 22 and for the second is 25. Each model was executed considering its corresponding number of lags.

MAPE



	(2,1,0)x(1,1,0)s=4	(2,1,0)x(0,1,1)s=4
1	5.182666	5.187595
2	8.047411	7.738436
3	8.669366	8.040853
4	7.872255	7.354798

MSPE



	(2,1,0)x(1,1,0)s=4	(2,1,0)x(0,1,1)s=4
1	0.001899607	0.001992885
2	0.004244895	0.004280452
3	0.004796793	0.004549059
4	0.004489635	0.004357088

According to the results we decided to go with the second model (**(2,1,0)x(1,1,1)s=4**) that except for the first prediction is the one that has less absolute error in its predictions, independently of the predicting method. Also, the square error is, in general, lower in this model which means that our model is more precise in the predictions.