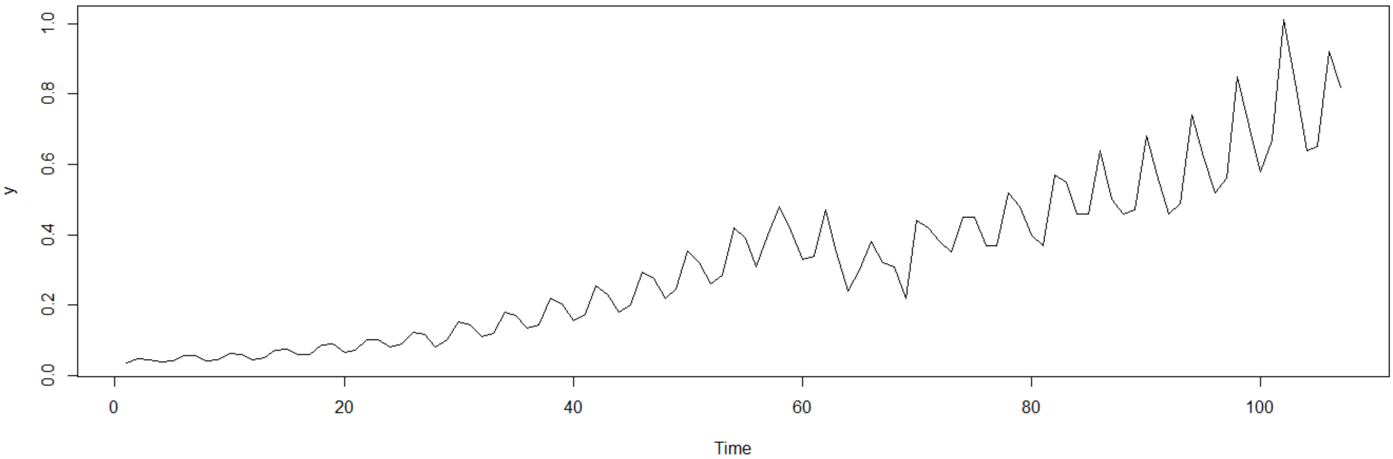
# Forecasting Time Series

**Professor: Lorenzo Pascual Caneiro**

# Homework 2

**Coca-Cola quarterly earnings** (1983-2009)



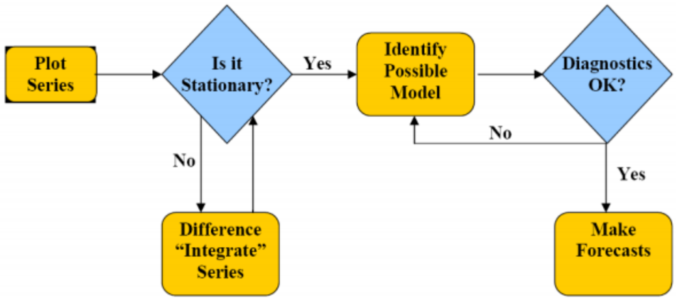

**Group B**

Stefano Magnan
Nicolas Boer
Eleonora Jimenez
Rebecca Rosser
Alberto de Roni
Laura Frazer

# Task 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)*

To complete the exercise, we will follow the Box-Jenkins methodology:

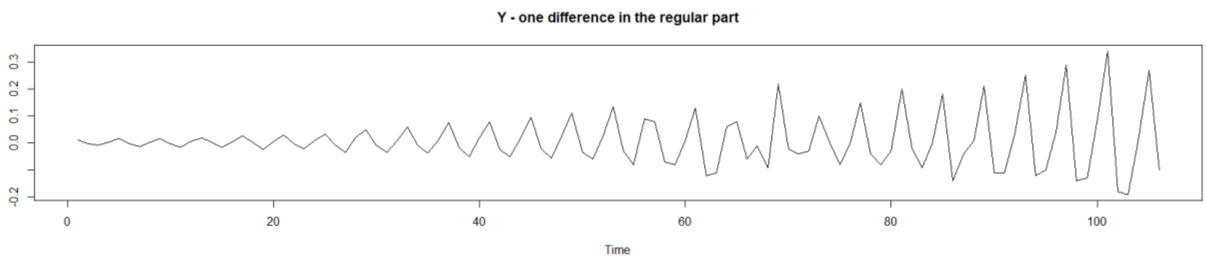



Before plotting the data, we define s=4 as the data is 'quarterly' data.

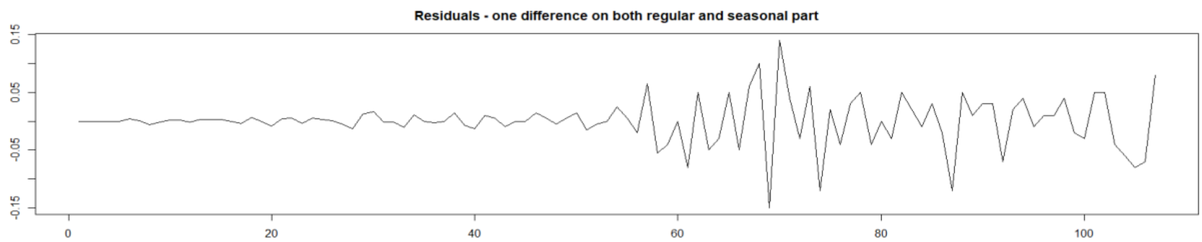
After plotting the Coca-Cola quarterly earnings (as seen in the cover page) we can say without doing any test that the **data** is **not stationary** in the mean as it has an upward trend. We will check the variance after taking the differences.

We calculated the number of differences on the regular and seasonal data using the Augmented Dickey–Fuller test ("adf") and the Osborn, Chui, Smith, and Birchenhall ("ocsb") test. We identified that **we need to take 1 difference on both the regular and the seasonal data**.

After taking the difference in the regular data, we see that the **data is now stationary in the mean,** but the "ocsb" test tells us that we still need to take one difference in the seasonal part to achieve fully stationary data**.** We also notice that data is **characterized** by **an increasing variance**. After removing seasonality, we can assess if we need to perform a log transformation at the beginning.

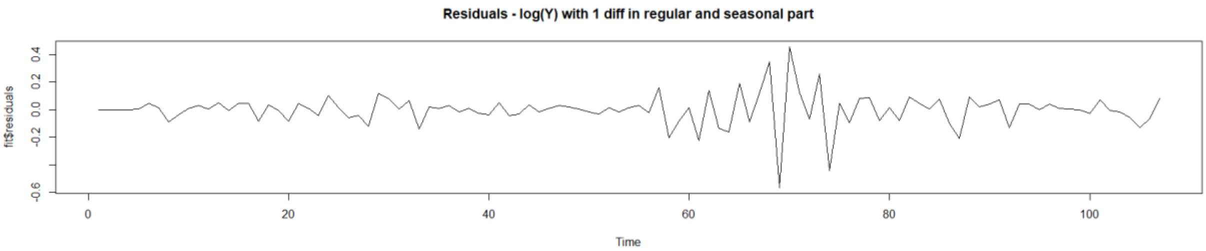



We identified the following model: **ARIMA(0,1,0)x(0,1,0)s=4**, to take one difference in both the regular and seasonal part. In the following plot, we can see the residuals:

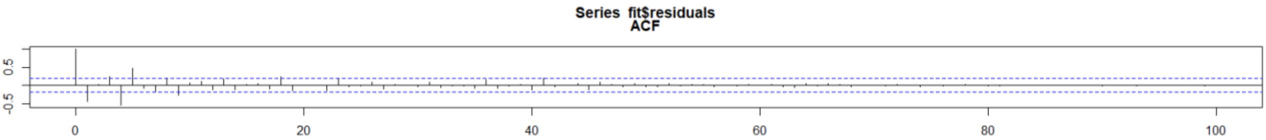


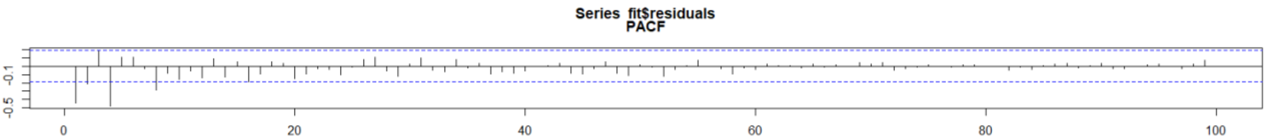
**The data is still not stationary in the variance**, so we now **need to transform the original data using the 'log operator',** to make it stationary.

**After taking the log operator** and making **1 transformation** in the **regular part** and **1 transformation** in the **seasonal part**, we have the following model **ARIMA(0,1,0)x(0,1,0)s=4** but **with the logs** of the original data. We can now see that **we have stationary data in both the mean and the variance:**




Now we proceed to plot the ACF and PACF of the residuals:




Looking at the ACF, we can see that for the seasonal part, we have only lag 4 out of bounds which identifies a SMA(1), because s = 4. Additionally, with lag 1 out of bounds, that can identify some relationships in the regular part.

Looking at the PACF, we can see that for the seasonal part, we have lags 4 and 8 out of bound, which identifies a SAR(2).

In total, we identified 12 models:

| N° | MODEL | COEFF. | ACF/PACF | WN (BOX-TEST RESIDUALS) |
|---|---|---|---|---|
| **1** | ARIMA(0,1,0)(2,1,0) s = 4 | Significant | 1st and 5th lag out of bound | No linear dependence (p-value>5%) |
| **2** | ARIMA(1,1,0)(2,1,0) s = 4 | Significant | 1st lag is no longer out of bound, but still the 5th is | No linear dependence (p-value>5%) |
| **3** | ARIMA(0,1,1)(2,1,0) s = 4 | Significant | 1st lag is no longer out of bound, but still the 5th is | No linear dependence (p-value>5%) |
| **4** | ~~ARIMA(5,1,0)(2,1,0) s = 4~~ | AR3, AR5, S1 are not significant | No lags out of boundary | No linear dependence (p-value>5%) |
| **~~5~~** | ~~ARIMA(0,1,5)(2,1,0) s = 4~~ | MA2, MA3, S1, S2 are not significant | 5th lag out of bound in both ACF and PACF | No linear dependence (p-value>5%) |
| **~~6~~** | ~~ARIMA(1,1,1)(2,1,0) s = 4~~ | AR1, MA1 are not significant | 5th lag out of bound in both ACF and PACF | No linear dependence (p-value>5%) |
| **7** | ARIMA(0,1,0)(0,1,1) s = 4 | Significant | 1st lag out of bound in both ACF and PACF | No linear dependence (p-value>5%) |
| **8** | ARIMA(1,1,0)(0,1,1) s = 4 | Significant | 5th lag out of bound in both ACF and PACF | No linear dependence (p-value>5%) |
| **9** | ARIMA(0,1,1)(0,1,1) s = 4 | Significant | No lags out of boundary | No linear dependence |

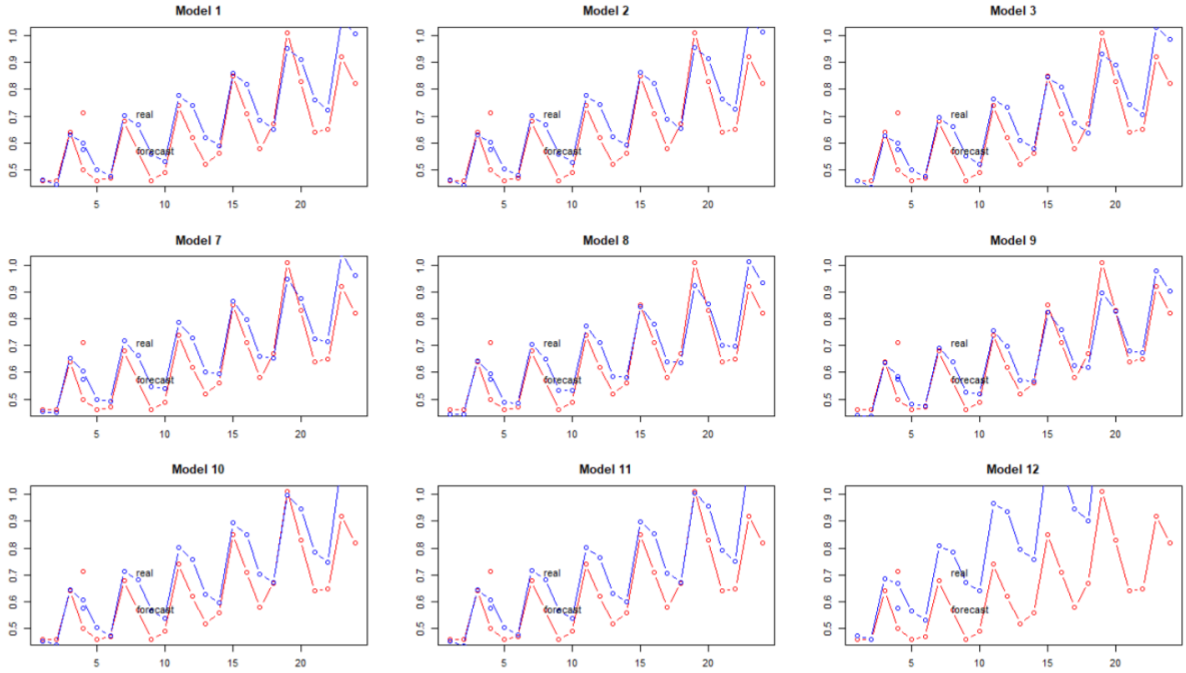
| | | | | |
|---|---|---|---|---|
| | | | | (p-value>5%) |
| **10** | ARIMA(0,1,0)(1,1,0) s = 4 | Significant | 1st lag is no longer out of bound, but still the 5th is | No linear dependence (p-value>5%) |
| **11** | ARIMA(1,1,0)(1,1,0) s = 4 | Significant | No lags out of boundary | No linear dependence (p-value>5%) |
| **12** | ARIMA(0,1,5)(1,1,0) s = 4 | Significant | 5th lag out of bound in both ACF and PACF | No linear dependence (p-value>5%) |

We discarded the models 4, 5, and 6 because the coefficients that refer to the parameters chosen in the model are not significant (e.g. Model 5 coefficient S2 is not significant and refers to the parameter P=2 of the model).

## Task 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?*

Forecasting exercise:




After visualizing the comparison between point predictions and real values, we see that **models 8 and 9 seem to be the best performant ones**. To have a proper understanding of the models, we calculated the MAPE and MSFE of each model forecasting 8 periods ahead (2 years) with both recursive and rolling scheme.

We choose to use the **MAPE to evaluate our models** because as it is a percentage it allows us to compare the models in a better way and is easier to interpret than MSFE.

**Forecasting results:**

**MAPE Recursive**

| | Model 1 | Model 2 | Model 3 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Average() | Min() | Max() |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period 1 | 5.2379 | 5.4571 | 5.0603 | 5.0817 | 5.6679 | 5.2246 | 5.2809 | 5.4564 | 5.3757 | 5.3158 | 5.0603 | 5.6679 |
| Period 2 | 7.2895 | 7.0440 | 6.7770 | 7.2683 | 7.0793 | 6.7831 | 7.3801 | 7.5226 | 6.7882 | 7.1036 | 6.7770 | 7.5226 |
| Period 3 | 8.3046 | 8.0558 | 7.6820 | 8.1911 | 7.5104 | 7.2021 | 9.0626 | 9.1458 | 7.4602 | 8.0683 | 7.2021 | 9.1458 |
| Period 4 | 8.5381 | 7.9585 | 7.6162 | 7.7376 | 7.1381 | 6.8361 | 9.6256 | 9.2758 | 8.2079 | 8.1038 | 6.8361 | 9.6256 |
| Period 5 | 8.8501 | 8.9309 | 8.5617 | 7.5211 | 7.0733 | 6.9636 | 9.9827 | 9.4826 | 9.3042 | 8.5189 | 6.9636 | 9.9827 |
| Period 6 | 10.0899 | 9.7609 | 9.0037 | 8.1609 | 7.7407 | 7.1262 | 10.3522 | 10.1401 | 10.9118 | 9.2540 | 7.1262 | 10.9118 |
| Period 7 | 10.1744 | 9.4513 | 8.8058 | 8.5413 | 7.9176 | 7.4713 | 11.6979 | 11.3483 | 11.3509 | 9.6399 | 7.4713 | 11.6979 |
| Period 8 | 10.4479 | 9.1194 | 8.5329 | 8.2849 | 7.4648 | 7.2212 | 12.1719 | 11.3407 | 12.9859 | 9.7300 | 7.2212 | 12.9859 |

**MAPE Rolling**

| | Model 1 | Model 2 | Model 3 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Average() | Min() | Max() |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period 1 | 5.2422 | 5.4757 | 5.0683 | 5.0727 | 5.6913 | 5.2269 | 5.2805 | 5.4676 | 6.1269 | 5.4058 | 5.0683 | 6.1269 |
| Period 2 | 7.2882 | 7.0472 | 6.7710 | 7.2791 | 7.1160 | 6.7915 | 7.3816 | 7.5264 | 7.9971 | 7.2443 | 6.7710 | 7.9971 |
| Period 3 | 8.3119 | 8.0503 | 7.6753 | 8.1695 | 7.5414 | 7.1990 | 9.0640 | 9.1456 | 9.1810 | 8.2598 | 7.1990 | 9.1810 |
| Period 4 | 8.5489 | 7.9471 | 7.6154 | 7.7081 | 7.1233 | 6.8271 | 9.6240 | 9.2673 | 10.3274 | 8.3321 | 6.8271 | 10.3274 |
| Period 5 | 8.8525 | 8.9439 | 8.5695 | 7.4562 | 7.1205 | 6.9936 | 9.9773 | 9.4915 | 11.2706 | 8.7417 | 6.9936 | 11.2706 |
| Period 6 | 10.0846 | 9.7380 | 8.9943 | 8.1471 | 7.7619 | 7.1433 | 10.3461 | 10.1318 | 12.3518 | 9.4110 | 7.1433 | 12.3518 |
| Period 7 | 10.1649 | 9.4178 | 8.7801 | 8.4957 | 7.9673 | 7.4657 | 11.6914 | 11.3457 | 11.8615 | 9.6878 | 7.4657 | 11.8615 |
| Period 8 | 10.4322 | 9.0943 | 8.5217 | 8.2437 | 7.5142 | 7.2349 | 12.1696 | 11.3523 | 13.3851 | 9.7720 | 7.2349 | 13.3851 |

To conclude, we **selected model 9 (ARIMA(0,1,1)(0,1,1) s = 4)** because it is the **best model** for predicting the Coca-Cola quarterly revenues as it is the one which **has the lowest MAPE on 6 of 8 periods and is lower than the average for the other 2.**