**BIG DATA & ARTIFICIAL INTELLIGENCE**

**IN OPERATIONS MANAGEMENT**

**Homework 2**

**Group 4**

**October 2022**

# Part 1: Identifying time series models

## Dataset evaluation

Given the Coca Cola dataset, we first plot the quarterly earnings values to get a look at stationarity:

Chart

Description automatically generated

At first glance we can observe that there is no stationarity present, as the pattern trends away from the mean. Variance is also rising, but it is not immediately clear if this pattern is significant. Upon running an ADF test on the dataset we get the following:

Text

Description automatically generated

With a p-value > 0.05 by a large amount, we can confidently say that there is in fact no stationarity. We can also identify that the dataset likely requires both one regular difference as well as a seasonal one. In order to understand the type of SARIMA model that would best fit this series, we plot the ACF and PACF to look for their behavior:

Timeline

Description automatically generated

## Identifying orders

Looking at the ACF and PACF we can identify the following clues:

* The ACF trends exponentially towards 0.
* The ACF has ‘peaks’ every four lags, indicating seasonality.

Based on these, we run an ARIMA(0,1,0)x(0,1,0) on the data to transform it into stationary (Box test confirmed it as such with p-value < 0.05) and evaluate the ACF/PACF of the residuals:

Chart

Description automatically generatedChart

Description automatically generated

We get some more clues:

* The PACF at lag 1 is identical to the ACF at lag 1.
* Both the ACF ≠ 0 and PACF ≠ 0 at lags 1 and 4, indicating that adding a seasonal component to the ARIMA would make sense.
* The ACF ≠ 0 at lag 5 as well, so the multiplicator factor of having an additional MA on top of the seasonal would allow us to account for this variation.
* There are late lags (>30) in the PACF that deviate from 0, however, them being so far makes them irrelevant in a way as we don’t want too much weight from time points from too far back.

Taking all clues into account and the behavior of the ACF/PACF, we tested out a few models and have identified two promising ones:

* ARIMA(1,1,0)x(1,1,0)s
* ARIMA(0,1,1)x(0,1,1)s
* Where s = 4

## Model 1: ARIMA(1,1,0)x(1,1,0)s

When fitting the dataset with this model, we obtain the following results:

Graphical user interface

Description automatically generated with low confidence

With all p-values < 0.05, we can say that they are all significant. This means that, if there is no correlation within the residuals, this model could be accurate in predicting the time series. Looking at the stationarity and the ACF/PACF after the ARIMA:

Chart, line chart

Description automatically generated Chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

In this situation, affirming that the data is stationary and that there is no white noise based only on first glance is impossible. This is because there is a growing variation after half the dataset, although it is not clear if it is growing or just a pattern that changed over time. Because of this and the fact that certain PACF points are outside of bounds, we perform a box test to ensure stationarity, which resulted in a p-value of 0.63. This means that the data is uncorrelated, so we are now free to plot the result and see what comes out:

Chart, line chart, scatter chart

Description automatically generated

## Model 2: ARIMA(0,1,1)x(0,1,1)s

The second model proposed returned very similar results:

Text

Description automatically generated

P-values all under 0.05 indicate significance. Upon checking the residuals, we also have a similar situation as with model 1 with a box-test p-value of 0.22 (uncorrelated and WN):

Chart, line chart

Description automatically generatedChart

Description automatically generated Chart

Description automatically generated

Lastly, the predictive pattern also makes sense:

Chart, line chart

Description automatically generated

# Part 2: Comparing results of models

## SARIMA(1,1,0)x(1,1,0)s:

When applying the model found using AR fully, we identify the following pattern when backtracking 24 points and testing against those:

Chart, line chart

Description automatically generated

As can be seen, the trend of the prediction follows the same general direction as the actual results that occurred with those 24 points. It is observable that the prediction has a slightly different variation from the actual points, but this can be explained by a growing variation in the actual results which is not accounted for in the model.

## SARIMA(0,1,1)x(0,1,1)s

A very similar pattern to the other model can also be observed in this one:

Chart, line chart, scatter chart

Description automatically generated

## Comparing Results

With both models behaving in a very similar fashion, we need to investigate the errors to see which one is ‘better’, although it could be argued that with every additional point in the future either could beat the other because of future randomness. Upon using several metrics, we get the following results:

Text

Description automatically generatedText

Description automatically generated with low confidence

With these metrics in hand, we can say that Model 2 is only marginally better than its counterpart after checking for differences in the errors. However, both are almost equally fit to be used for this time series. If we had to pick, we would go for Model 2 for the time being, but with further observations in the future this could change.