**Written Report – 6.419x Module 4**

**Name:** (Felipe Mehsen Tufaile)

* **The final model**

**1.** *(3 points) Plot the periodic signal . (Your plot should have 1 data point for each month, so 12 in total.) Clearly state the definition the , and make sure your plot is clearly labeled*.

**Solution:**

In order to calculate , the following procedure was implemented:

* Time variable () was calculated following its definition in the problem statement: ;
* Missing CO2 concentration value were removed in order to create an interpolation function;
* An interpolation function was created using the available data (x= and y=CO2 concentration) and previous missing values were substituted by the interpolated values;
* The remaining data was filtered so values would range from Oct 1958 to Sep 2019. This was done to ensure the dataset would contain 61 records of each month, that is, 61 complete cycles;
* The trend was modeled fitting the 61-cycles data in a quadratic model of the form ;
* The deterministic trend was removed from the series and the residual was averaged for each month. This average residual for each month is then the periodic signal ;
* Note: since interpolation was done initially, there was no need to interpolate the values again to calculate the periodic signal.

The visualization of the period signal calculated can be seen in **Figure 1**. Looking at the signal, it is possible to note a sinusoidal pattern with the highest concentration of CO2 occurring around May and the lowest concentration of CO2 occurring around October.

A graph with lines and dots

Description automatically generated

**Figure 1** – Monthly periodic signal of the CO2 concentration (values in ppm).

**2.** *(2 points) Plot the final fit . Your plot should clearly show the final model on top of the entire time series, while indicating the split between the training and testing data.*

**Solution:**

**Figure 2** shows the training data (red), the test data (green) and the predictions made by the final model (black markers). In general, it is noticeable that the predicted data fits quite well the original series (train and test sets taken together), which validates both trend and periodic components calculated. However, the predictions seem to deviate from the original series after Jan-2016 (58 years after Jan-1958).

A graph showing a curve

Description automatically generated

**Figure 2** – CO2 concentration (ppm) since January 1958. First value shown correspond to the CO2 concentration of October 1958 (~0.79 years after Jan-1958) while the last value corresponds to September 2019 (~61.7 years after Jan-1958) which encompass 61-cycles of the CO2 concentration measurement.

**3.** *(4 points) Report the root mean squared prediction error RMSE and the mean absolute percentage error MAPE with respect to the test set for this final model. Is this an improvement over the previous model   without the periodic signal? (Maximum 200 words.)*

**Solution:**

The root mean squared error **RMSE** obtained with the final model is **1.120**, whereas the mean absolute percentage error **MAPE** is **0.202%**. If we compare to the quadratic model without the periodic term (RMSE = 2.501 and MAPE = 0.532%) we see that **the final model with the periodic term outperforms the model without the periodic term**. The reason is because the model without seasonal adjustment only accounts for the average increase in CO2 for each point in time, disregarding any periodic fluctuation of the CO2 concentration. When we add the periodic term to the model, we give it the ability to differentiate the average increase in CO2 depending on the month of the year.

**4.** *(3 points) What is the ratio of the range of values of F to the amplitude of  and the ratio of the amplitude of   to the range of the residual  (from removing both the trend and the periodic signal)? Is this decomposition of the variation of the CO2 concentration meaningful? (Maximum 200 words.)*

**Solution:**

**Figure 3** shows the distribution of the ratio of the range of values of F to the amplitude of  (F/) in the left chart and the ratio of the amplitude of   to the range of the residual  (Pi/) in the right chart. If we calculated the median value of the two distributions, we would find median F/ ~ 32.56 and median Pi/ ~ 25.73. Since these ratios are significantly greater than 1, it is possible to say that the trend, seasonal and residuals components have different orders of magnitude, which justifies the decomposition of the CO2 concentration series into the mentioned components. If some of the ratios were close to 1, the decomposition would be less meaningful since the overall pattern or direction of the data would not be clearly distinguishable.

A screenshot of a graph

Description automatically generated

**Figure 3** – Distribution of the ratio of the range of values of F to the amplitude of  (F/) in the left chart and the ratio of the amplitude of   to the range of the residual  (Pi/) in the right.

* **Autocovariance Function**

**1.** *(4 points) Consider the MA(1) model, , where . Find the covariance function of . Include all important steps of your computations in your report.*

**Solution:**

The autocovariance function of can be calculated implementing the covariance function as follows:

Where h is the lag order.

Calculating the covariance function indicated above for h = 0, gives:

For h = :

Similarly, for h=1:

For h = :

Similarly, for h=2:

Finally, for ,

Therefore, the autocovariance function of is:

**2.** *(4 points) Consider the AR(1) model, , where . Suppose . Find the covariance function of . (You may use, without proving, the fact that is stationary if ). Include all important steps of your computations in your report.*

**Solution:**

The autocovariance function of can be calculated implementing the covariance function as follows:

Where h is the lag order.

Calculating the covariance function indicated above for h = 0, gives:

Since, , we have:

Rearranging:

Now, in order to be defined, . Therefore, must meet the condition , which is assumed in the problem statement.

If we now calculate the autocovariance for h = , we will find:

Since, , we have:

As seen before, , therefore:

Similarly, for any given “h” value, we have:

* **Converting to Inflation Rates**

**1.** *(9 points) Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI.*

*Your response should include:*

* *A. (1 point) Description of how you compute the monthly inflation rate from CPI and a plot of the monthly inflation rate. (You may choose to work with log of the CPI.)*
* *B. (2 points) Description of how the data has been detrended and a plot of the detrended data.*
* *C. (3 points) Statement of and justification for the chosen AR(p) model. Include plots and reasoning.*
* *D. (3 points) Description of the final model; computation and plots of the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.*

**Solution:**

**A.** To calculate the inflation rate (IR), it was used the CPI data from the first day of each month. Any row with missing CPI data was excluded. The inflation rate was then calculated by implementing the log return of the CPI, represented as . The resulting time series of inflation rates spans from Sep-2008 to Oct-2019, as shown in the left visualization of **Figure 4** below.

A screenshot of a graph

Description automatically generated

**Figure 4** – Inflation rates calculated by implementing the log return of the CPI. Left visualization: data from Sep-2008 to Oct-2019; Right visualization: data from Jan-2009 to Oct-2019.

Looking at the left visualization of **Figure 4,** we note that there is a significantly drop in the inflation rate around Dec-2008 (probably because of the financial crisis that happened hit in Sep-2008). In order to ensure that this event does not affect the time series analysis, all datapoints prior to Jan-2009 will be excluded from the time series. The resulting time series is shown in the right visualization of **Figure 4** above.

**B.** The selection of the detrending method for the time series involved comparing the fit of three different trend polynomials. These polynomials were: first order polynomial (linear trend), second order polynomial (quadratic trend) and third order polynomial (cubic trend). The model was trained in the first 43% of the series, that is, all values prior to Sep-2013.

The result of the detrend process considering the three approaches mentioned can be seen in **Figure 5**. Looking at Figure 5, the quadratic trend appears to provide the best fit for the time series. To validate this observation, the mean value of the each detrended series was calculated and compared. Ideally, the mean value should be close to zero, as detrending the series using a polynomial with both time-dependent terms and an independent term aims to eliminate any trend and constant value.

* Mean value of the trended series: 0.00190;
* Mean value of the detrended series using a linear trend: 0.000456;
* Mean value of the detrended series using a quadratic trend: -0.000355;
* Mean value of the detrended series using a cubic trend: 0.0188;

Looking at the mean values calculated we can confirm that the quadratic trend best fit the inflation rate time series (lower mean value). Therefore, the final detrended time series will be the residual after removing the **quadratic trend** of the trended time series.

A graph of a graph

Description automatically generated with medium confidence A close-up of a graph

Description automatically generated

A close-up of a graph

Description automatically generated

**Figure 5** – Detrend inflation rate timeseries considering three different trend polynomials: first order polynomial (linear trend), second order polynomial (quadratic trend) and third order polynomial (cubic trend).

**C.** Looking at the ACF plot of the detrended inflation rate in **Figure 6** we note that there are statistically significant correlations for some of the lags (e.g. lag 11 and lag 12 ), which also seems to occur in a pattern that repeats every 12 months. This suggests that the detrended inflation rate under analysis has a periodic term that should be removed before applying an auto regressive model.

A diagram of a diagram

Description automatically generated with medium confidence

**Figure 6** – ACF and PACF plot of the detrend inflation rate.

The seasonal component was calculated by averaging the inflation rate values for each month using all values prior to Sep-2013, the training set, which resulted in the pattern shown in the left visualization of **Figure 7.** The seasonal component was removed from the detrended series which resulted in the detrended and deseasonalized series shown by the green curve in the right visualization of **Figure 7**.

A graph of a graph and a graph of a graph

Description automatically generated

**Figure 7** – Seasonal component of the inflation rate series and the detrended and deseasonalized inflation rate series.

Looking now at the ACF plot of the detrended and deseasonalized inflation rate series in **Figure 8** we notice that the periodic pattern observed in the ACF plot of the detrended inflation rate series is mostly gone. The ACF and PACF plot now indicates that the model is likely an ARMA(0, 1), or MA(1), model since we do not observe an exponential decaying pattern in the ACF plot and there is only a statistically significant correlation with lag 1 in the ACF plot, which indicates the “q” term of the MA model. Therefore, the “p” term of the AR(p) model is 0.

A diagram of a diagram of a line

Description automatically generated with medium confidence

**Figure 8** – ACF and PACF plot of the detrended and deseasonalized inflation rate series.

**D.** Given that the problem statement specifically asks for an AR model, although the best model may be a MA(1) model, the final model with a quadratic trend and a seasonal component of the form:

Where:

* : is the inflation rate at time “t”
* : is the quadratic trend;
* : is the seasonal component of the inflation rate at month “i”;
* : is MA(1) component;

**Figure 9** shows the actual inflation rate (black) against the 1 month-ahead forecast (red). Forecasting starts in Oct-2013 and goes on until Oct-2019 (validation set). The forecast process follows an “evaluation on a rolling forecasting origin” approach where the training set increases one data point for each new iteration and forecast is performed for the next data point outside of the training set (see chapter 5.10, time series cross-validation, of reference [1] for more details). The RMSE of the validation ser is 0.00186.

A graph showing a graph

Description automatically generated with medium confidence

**Figure 9** – Actual inflation rate (black) against the 1 month-ahead forecast (red).

**2.** *(3 points) Which* ***AR****(p) model gives the best predictions? Include a plot of the RMSE against different lags p for the model.*

**Solution:**

**Figure 10** shows the 1-month-ahead inflation ration forecast for AR(p) model with p ranging from 0 to 3. Looking at the figure, it is hard to tell the difference in performance since the forecasts looks very similar for all AR models tested. However, if we calculate the RMSE score for the validation set for all AR(p) models, we will not that the AR(0) model is the one that generates the lower RMSE and, therefore, gives the best predictions:

* AR(0) model - RMSE: 0.00206;
* AR(1) model - RMSE: 0.00207;
* AR(2) model - RMSE: 0.00218;
* AR(3) model - RMSE: 0.00210;

A graph showing a graph

Description automatically generated with medium confidence

**Figure 10** – Actual inflation rate (black) against the 1-month-ahead forecast from multiple AR(p) models.

Additionally, if we would calculate the BIC value for each model in order to account for model complexity as well we would confirm that the AR(0) model is the best since it has the lowest value. This result is expected since the AR(0) model not only has the lowest RMSE score, but also has the lowest number of terms (less complex).

* AR(0) model - BIC: -1216.1;
* AR(1) model - BIC: -1211.5;
* AR(2) model - BIC: -1211.1;
* AR(3) model - BIC: -1203.9;

**3.** *(3 points) Overlay your estimates of monthly inflation rates and plot them on the same graph to compare. (There should be 3 lines, one for each dataset, plus the prediction, over time from September 2013 onward).*

**Solution:**

**Figure 11** shows the CPI IR series, BER IR series (monthly) and the CPI IR forecast using the AR(0) model from the previous question (red curve). The forecast shows the 1-month-ahead forecast (h=1) as calculated in the previous question. All inflation rate series are plotted using a common y-axis. It is noticeable that the amplitude of oscillation of the CPI series is significantly greater than the amplitude of oscillation of the BER series.

**Figure 12**, on the other hand, shows the same CPI and BER series, but plotted in different y-axis. It is visible that the overall oscillation pattern is somewhat similar for all curves: in general, when there is a local maximum in the CPI series, there is usually a local maximum in the BER series (the same is valid for local minimums).

**A graph with red and blue lines

Description automatically generated**

**Figure 11** – CPI IR Forecast (h=1), CPI IR Monthly and BER IR Monthly. CPI and BER inflation rate series shown in the same y-axis.

A graph showing a graph

Description automatically generated with medium confidence

**Figure 12** – CPI IR Forecast (h=1), CPI IR Monthly and BER IR Monthly. CPI and BER inflation rate series shown in different y-axis.

* **External Regressors and Model Improvements**

**1.** *(9 points) Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI.*

***External Regressors***

*Next, we will include monthly* ***BER*** *data as an external regressor to try to improve the predictions of inflation rate. Here we only consider to add one* ***BER*** *term in the AR(p) model of CPI inflation rate. In specific, we model the CPI inflation rate by*

*where is the BER inflation rate at time , is the lag of BER rate w.r.t. CPI rate, and is white noise.*

*1. (4 points) Plot the cross correlation function between the CPI and BER inflation rate, by which find r, i.e., the lag between two inflation rates. (As only one external regressor term is involved in the model, we only consider the peak in the CCF plot.)*

**Solution:**

Before creating the cross-correlation function between the CPI IR and BER IR series, the two series were converted to a stationary. This process is done to ensure that these two series contain no trend, seasonality or autocorrelation structure that could lead to spurious correlation.

To make the CPI IR series stationary, the CPI IR series was fit to the model described in previous questions, , and the residual was calculated. On the other hand, to make the BER IR series stationary the series was fit to a AR(2) model with seasonal component and the residual was calculated.

In order to ensure that the residuals of each model were, in fact, stationary, the residuals were submitted to the Augmented Dickey-Fuller (ADF) test, which, in summary, check the stationarity of a time series by evaluating the presence of a unit root (null hypothesis states that there is unit root). Since the p-value obtained for both series of residuals were less than 0.05 we reject the hypothesis that the series have unit root and, therefore, we may consider the series stationary.

* p-value ADF test CPI IR residuals: 3.52e-20;
* p-value ADF test BER IR residuals: 5.27e-25;

Finally, after ensuring that the two series are stationary, the cross-correlation function between the series was calculated and the result is shown in **Figure 14**. It is visible that the two series have significant correlation when BER IR is lagged once, that is .

A graph with blue and white lines

Description automatically generated

**Figure 13** – Cross correlation function between CPI IR (residuals) and BER IR (residuals).

*2. (3 points) Fit a new AR model to the CPI inflation rate with these external regressors and the most appropriate lag. Report the coefficients, and plot the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.*

**Solution:**

The best model found using the external regressor is a SARIMAX(1,0,1)(1,0,1)[12] model, as it can be seen the table of **Figure 14** below.

A screenshot of a calendar

Description automatically generated

A table of numbers and letters

Description automatically generated

**Figure 14** – Description of the best SARIMAX model created to predict future CPI IR values using the BER IR lag 1 time series (x1).

Given the information mentioned in the table, the SARIMAX model can be represented as follows:

Where,

* : CPI inflation rate at time “t” and seasonal period T;
* : CPI inflation rate at time “t-1”;
* : Error term at time “t-1”;
* : CPI inflation rate for the corresponding month at cycle “T-1”;
* : Error term for the corresponding month at cycle “T-1”;
* : BER monthly inflation rate at time “t-1” (lag 1);

The plot of the CPI IR forecast using the SARIMAX model mentioned above (red curve) on top of the actual CPI IR (black curve) can be seen in **Figure 15** below. The image also shows the forecast of the MA(1) model created in previous question (green curve). Looking at the forecast, it is noticeable that the general pattern of both forecasts are very similar and approximates the original curve reasonably well. However, the SARIMAX model seems to underestimate the local maximum values more often than the MA(1) model.

**A graph showing a graph

Description automatically generated with medium confidence**

**Figure 15** – 1-month-ahead forecasts of the CPI inflation rate series using a SAMIRAX model with exogenous variable.

*3. (3 points) Report the mean squared prediction error for 1 month ahead forecasts.*

**Solution:**

The RMSE values obtained for the forecast of the validation set can be seen listed below. It can be seen that in terms of the RMSE, both MA(1) and SARIMAX(1,0,1)(1,0,1)[12] + BER IR lag1 have very similar performance.

* MA(1) model - RMSE: 0.00206;
* SARIMAX(1,0,1)(1,0,1)[12] + BER IR lag1 model - RMSE: 0.00205;

**Reference**

[1] Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on Nov 2023.