Written Report – 6.419x Module 4

Name: (Felipe Mehsen Tufaile)

▪ The final model

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

Solution:

In order to calculate P_i, the following procedure was implemented:

- Time variable (t_i) was calculated following its definition in the problem statement: $t_i = \frac{i+0.5}{12}$;
- Missing CO2 concentration value were removed in order to create an interpolation function;
- An interpolation function was created using the available data (x=t_i 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 $F(t_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$;
- The deterministic trend $F(t_i)$ was removed from the series and the residual was averaged for each month. This average residual for each month is then the periodic signal P_i;
- 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.

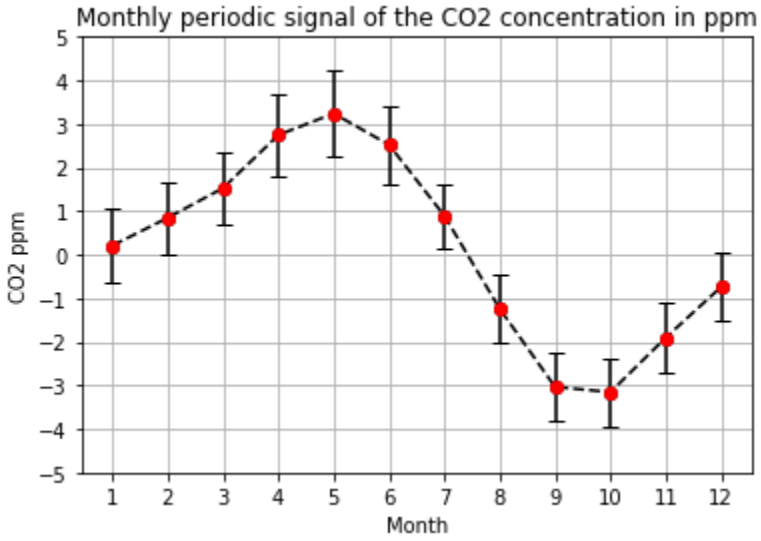

Figure 1 – Monthly periodic signal of the CO2 concentration (values in ppm).

2. *(2 points) Plot the final fit $F_n(t_i) + P_i$. 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).

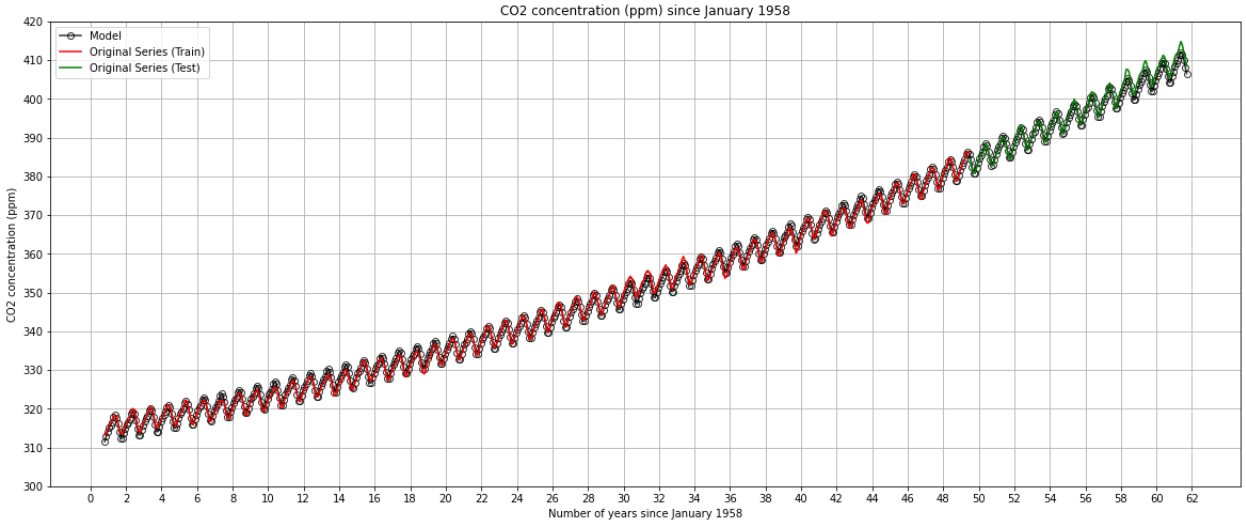

Figure 2 – CO2 concentration (ppm) since January 1958. First value shown correspond to the CO2 concentration of October 1958 (~0.79 years after Jan-1958) while the last value corresponds to September 2019 (~61.7 years after Jan-1958) which encompass 61-cycles of the CO2 concentration measurement.

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 $F_n(t_i)$ 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 P_i and the ratio of the amplitude of P_i to the range of the residual R_i (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 P_i (F/P_i) in the left chart and the ratio of the amplitude of P_i to the range of the residual R_i (Pi/R_i) in the right chart. If we calculated the median value of the two distributions, we would find median F/P_i ~ 32.56 and median Pi/R_i ~ 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.

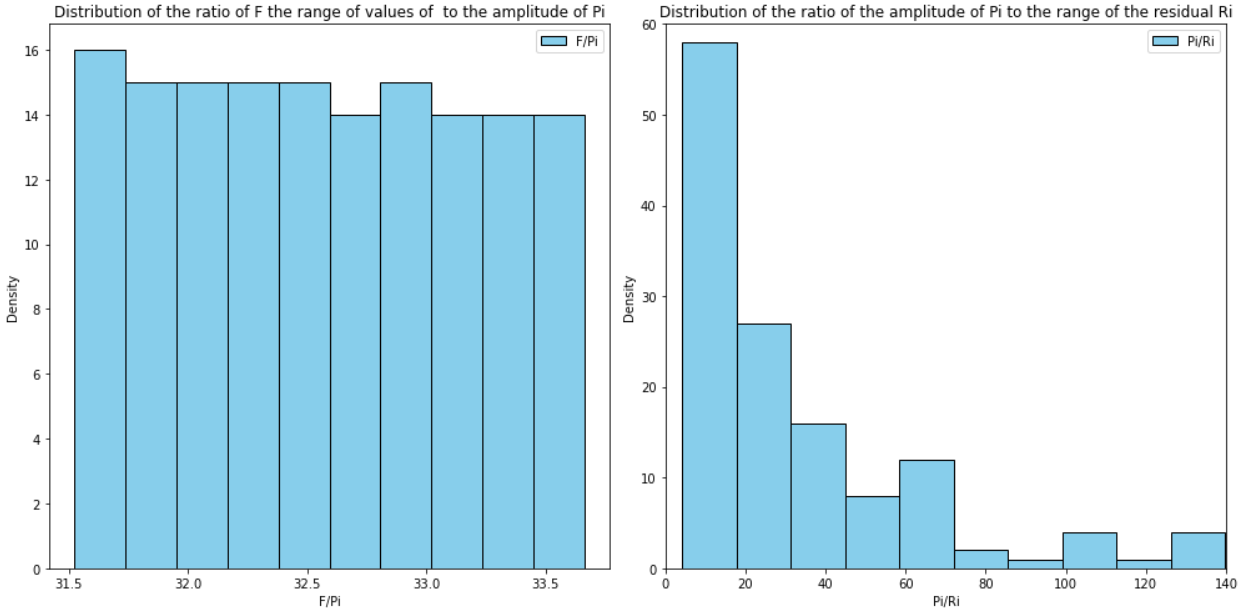

Figure 3 – Distribution of the ratio of the range of values of F to the amplitude of P_i (F/P_i) in the left chart and the ratio of the amplitude of P_i to the range of the residual R_i (Pi/R_i) in the right.

- **Autocovariance Function**

1. *(4 points) Consider the MA(1) model,* $X_t = W_t + \theta W_{t-1}$, *where* $\{W_t\} \sim W \sim N(0,\ \sigma^2)$. *Find the covariance function of* $\{X_t\}$. *Include all important steps of your computations in your report.*

Solution:

The autocovariance function of $\{X_t\}$ can be calculated implementing the covariance function as follows:

$$Cov(X_{t+h},\ X_t) = Cov(W_{t+h} + \theta W_{t+h-1},\ W_t + \theta W_{t-1})$$

Where h is the lag order.

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

$$Cov(X_t,\ X_t) = Cov(W_t + \theta W_{t-1},\ W_t + \theta W_{t-1})$$

$$Cov(X_t,\ X_t) = Cov(W_t, W_t) + Cov(W_t, \theta W_{t-1}) + Cov(\theta W_{t-1}, W_t) + Cov(\theta W_{t-1}, \theta W_{t-1})$$

$$Cov(X_t,\ X_t) = \sigma^2 + 0 + 0 + \theta^2\sigma^2 = \sigma^2(1 + \theta^2)$$

For h = -1:

$$Cov(X_{t-1},\ X_t) = Cov(W_{t-1} + \theta W_{t-2},\ W_t + \theta W_{t-1})$$

$$Cov(X_{t-1},\ X_t) = Cov(W_{t-1},\ W_t) + Cov(W_{t-1}, \theta W_{t-1}) + Cov(\theta W_{t-2},\ W_t) + Cov(\theta W_{t-2}, \theta W_{t-1})$$

$$Cov(X_{t-1},\ X_t) = 0 + \theta\sigma^2 + 0 + 0 = \theta\sigma^2$$

Similarly, for h=1:

$$Cov(X_{t+1},\ X_t) = \theta\sigma^2$$

For h = -2:

$$Cov(X_{t-2},\ X_t) = Cov(W_{t-2} + \theta W_{t-3},\ W_t + \theta W_{t-1})$$

$$Cov(X_{t-2},\ X_t) = Cov(W_{t-2},\ W_t) + Cov(W_{t-2}, \theta W_{t-1}) + Cov(\theta W_{t-3},\ W_t) + Cov(\theta W_{t-3}, \theta W_{t-1})$$

$$Cov(X_{t-1},\ X_t) = 0 + 0 + 0 + 0 = 0$$

Similarly, for h=2:

$$Cov(X_{t+2},\ X_t) = 0$$

Finally, for $|h| > 2$, $Cov(X_{t+h},\ X_t) = 0$

Therefore, the autocovariance function of $\{X_t\}$ is:

$$\gamma_{X_t}(t+h,t) = \begin{cases} \sigma^2(1+\theta^2), & for\ h = 0 \\ \theta\sigma^2, & for\ |h| = 1 \\ 0, & for\ |h| \geq 2 \end{cases}$$

2. *(4 points) Consider the AR(1) model,* $X_t = \emptyset X_{t-1} + W_t$*, where* $\{W_t\} \sim W \sim N(0,\ \sigma^2)$*. Suppose* $|\emptyset| < 1$*. Find the covariance function of* $\{X_t\}$*. (You may use, without proving, the fact that* $\{W_t\}$ *is stationary if* $|\emptyset| < 1$*). Include all important steps of your computations in your report.*

Solution:

The autocovariance function of $\{X_t\}$ can be calculated implementing the covariance function as follows:

$$Cov(X_{t+h},\ X_t) = Cov(\emptyset X_{t-1+h} + W_{t+h},\ \emptyset X_{t-1} + W_t)$$

Where h is the lag order.

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

$$Cov(X_t,\ X_t) = Cov(\emptyset X_{t-1} + W_t,\ \emptyset X_{t-1} + W_t)$$

$$Cov(X_t,\ X_t) = Cov(\emptyset X_{t-1}, \emptyset X_{t-1}) + Cov(\emptyset X_{t-1}, W_t) + Cov(W_t, \emptyset X_{t-1}) + Cov(W_t, W_t)$$

$$Cov(X_t,\ X_t) = \emptyset^2 Cov(X_{t-1},\ X_{t-1}) + 0 + 0 + \sigma^2$$

Since, $Cov(X_t,\ X_t) = Cov(X_{t-1},\ X_{t-1}) = \gamma(0)$, we have:

$$\gamma(0) = \emptyset^2\gamma(0) + \sigma^2$$

Rearranging:

$$\gamma(0) = \frac{\sigma^2}{1-\emptyset^2}$$

Now, in order to $\gamma(0)$ be defined, $1 - \emptyset^2 > 0$. Therefore, $\emptyset$ must meet the condition $|\emptyset| < 1$, which is assumed in the problem statement.

If we now calculate the autocovariance for h = -1, we will find:

$$Cov(X_{t-1},\, X_t) = Cov(\emptyset X_{t-2} + W_{t-1},\, \emptyset X_{t-1} + W_t) = Cov(\emptyset X_{t-2} + W_{t-1}, \emptyset(\emptyset X_{t-2} + W_{t-1}) + W_t)$$

$$Cov(X_{t-1},\, X_t) = Cov(\emptyset X_{t-2} + W_{t-1}, \emptyset^2 X_{t-2} + \emptyset W_{t-1} + W_t)$$

$$Cov(X_{t-1},\, X_t) = \emptyset^3 Cov(X_{t-2}, X_{t-2}) + \emptyset^2 Cov(X_{t-2}, W_{t-1}) + \emptyset Cov(X_{t-2}, W_t) + \emptyset^2 Cov(W_{t-1}, X_{t-2}) + \emptyset Cov(W_{t-1}, W_{t-1}) + Cov(W_{t-1}, W_t)$$

$$Cov(X_{t-1},\, X_t) = \emptyset^3 Cov(X_{t-2}, X_{t-2}) + 0 + 0 + 0 + \emptyset\sigma^2 + 0$$

Since, $Cov(X_t,\, X_t) = Cov(X_{t-2},\, X_{t-2}) = \gamma(0)$, we have:

$$Cov(X_{t-1},\, X_t) = \emptyset^3\gamma(0) + \emptyset\sigma^2 = \emptyset(\emptyset^2\gamma(0) + \sigma^2)$$

As seen before, $\emptyset^2\gamma(0) + \sigma^2 = \gamma(0)$, therefore:

$$Cov(X_{t-1},\, X_t) = \emptyset\gamma(0) = \emptyset\frac{\sigma^2}{1-\emptyset^2}$$

Similarly, for any given "h" value, we have:

$$Cov(X_{t-h},\, X_t) = \emptyset^h\frac{\sigma^2}{1-\emptyset^2}$$

- **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 $IR = log(CPI_t) - log(CPI_{t-1})$. The resulting time series of inflation rates spans from Sep-2008 to Oct-2019, as shown in the left visualization of **Figure 4** below.

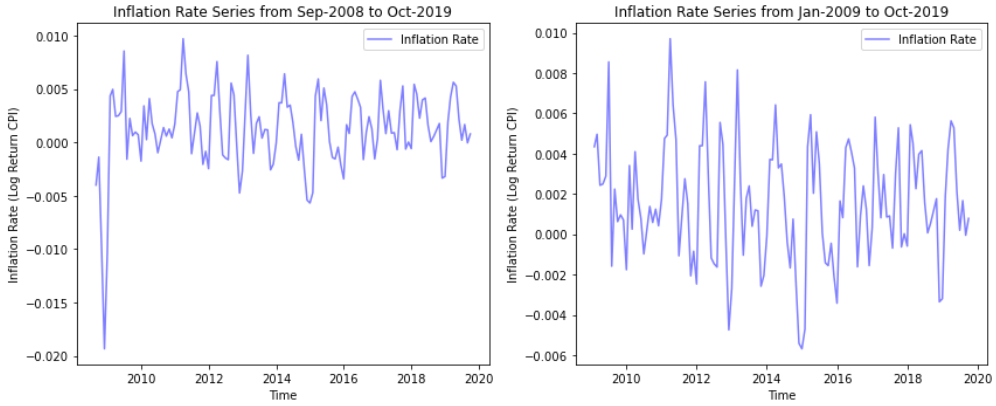

Figure 4 – Inflation rates calculated by implementing the log return of the CPI. Left visualization: data from Sep-2008 to Oct-2019; Right visualization: data from Jan-2009 to Oct-2019.

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**. To verify which polynomial trend would best fit the inflation rate series, the RMSE value of each detrended series was calculated and compared.

- Linear trend RMSE: 0.00312;
- Quadratic trend RMSE: 0.00292;
- Cubic trend RMSE: 0.0488;

Looking at the RMSE values calculated we verify that the quadratic trend is the one that best fit the inflation rate time series (lowest RMSE). **Therefore, the model will be detrended using a quadratic trend.**

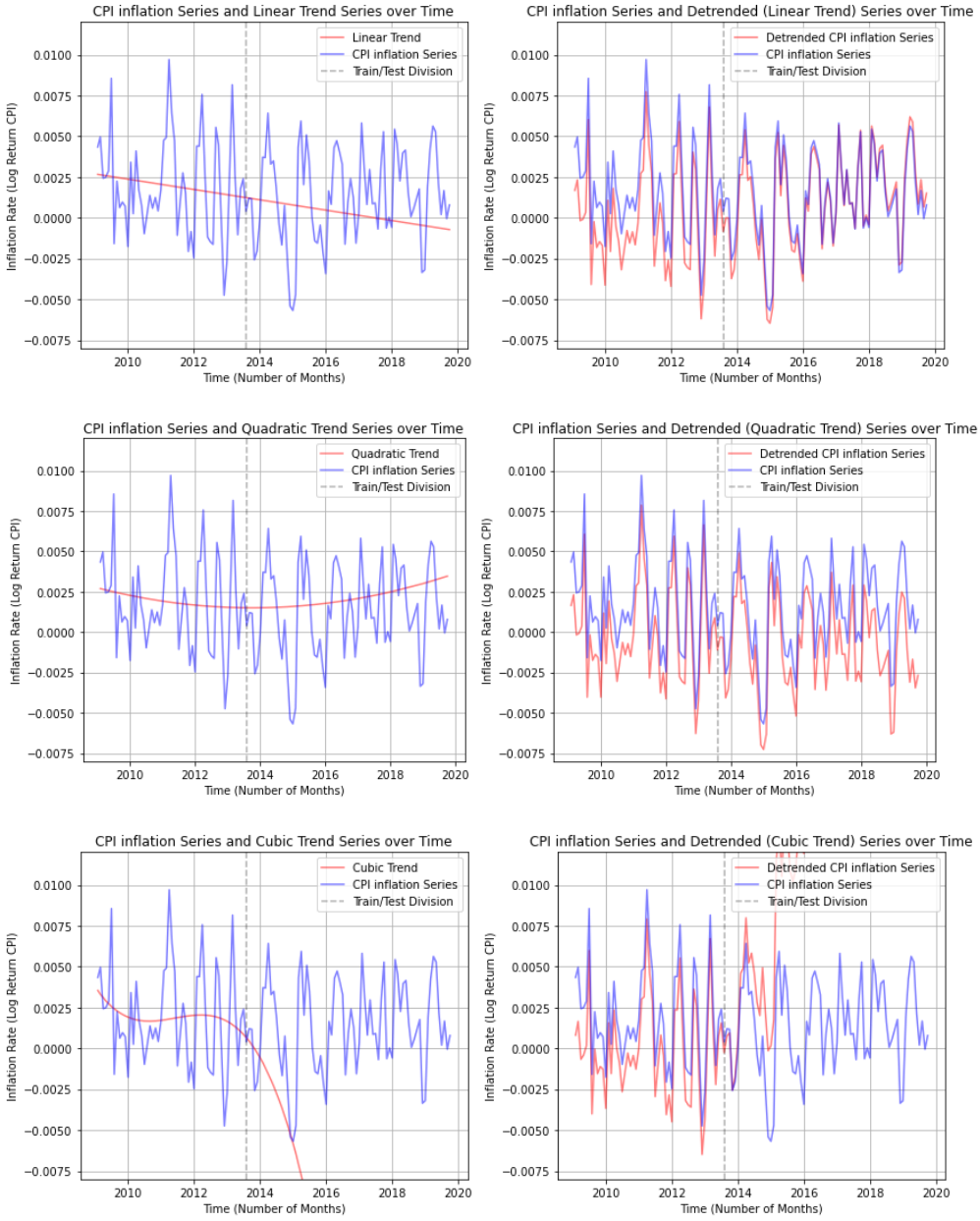

Figure 5 – Detrend inflation rate timeseries considering three different trend polynomials: first order polynomial (linear trend), second order polynomial (quadratic trend) and third order polynomial (cubic trend).

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 component that should be removed before checking for auto regressive components.

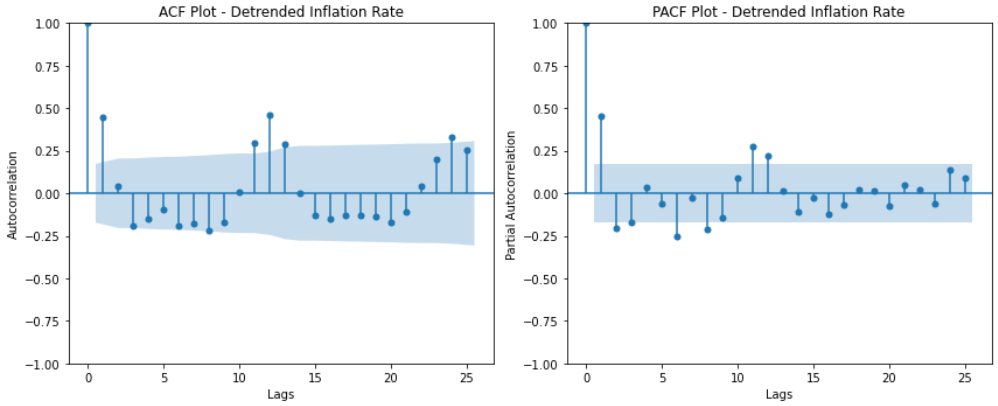

Figure 6 – ACF and PACF plot of the detrend inflation rate.

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

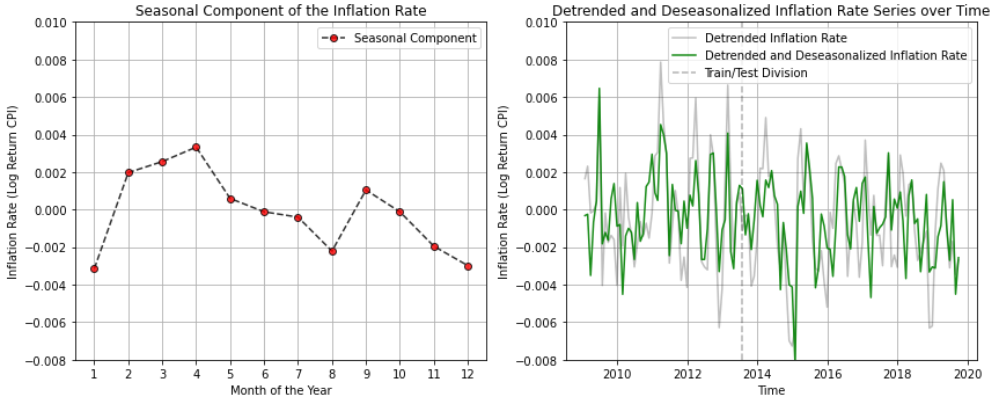

Figure 7 – Seasonal component of the inflation rate series and the detrended and deseasonalized inflation rate series.

Figure 8 now shows the ACF and PACF plots of the detrended and deseasonalized inflation rate series. The ACF plot suggests that the periodic pattern of the detrended inflation rate series is mostly gone. Furthermore, the ACF does not show an exponential decaying pattern as it would be expected if the model had auto regressive components. Instead, it only shows significant correlation for the lag 1 term. This indicates that the model that best fit the residual series is a MA(1) model. Therefore, since the problem statement specifically asks for a AR model, the best order of the AR model is p = 0, that is AR(0).

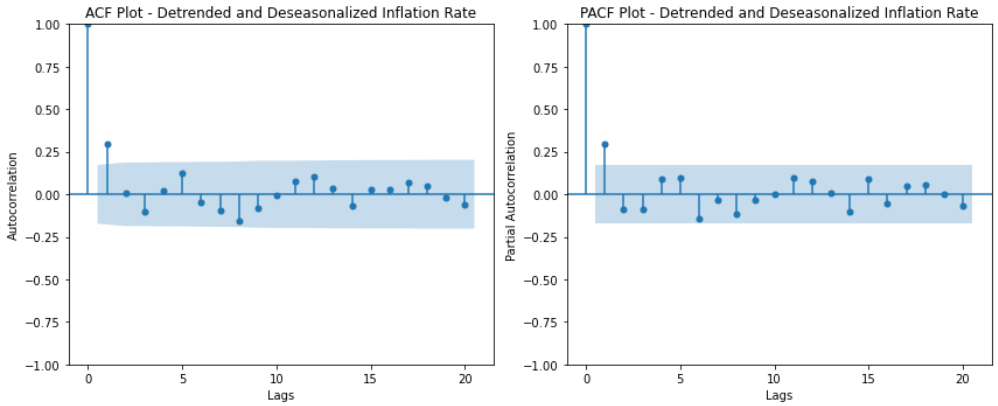

Figure 8 – ACF and PACF plot of the detrended and deseasonalized inflation rate series.

D. Since last question showed that the model that best fit the detrended and deseasonalized inflation rate series is a MA(1) model and that the question is not considering MA components at this moment, the final model is a quadratic trend model with seasonal component:

$$IR_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + P_i$$

Where:

- IR_t: is the inflation rate at time "t"
- $\alpha_0 + \alpha_1 t + \alpha_2 t^2$: is the quadratic trend;
- P_i: is the seasonal component of the inflation rate at month "i";

Figure 9 shows the actual inflation rate (black) against the 1 month-ahead forecast (red) using the final model. Forecasting starts in Sep-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 0.00244.

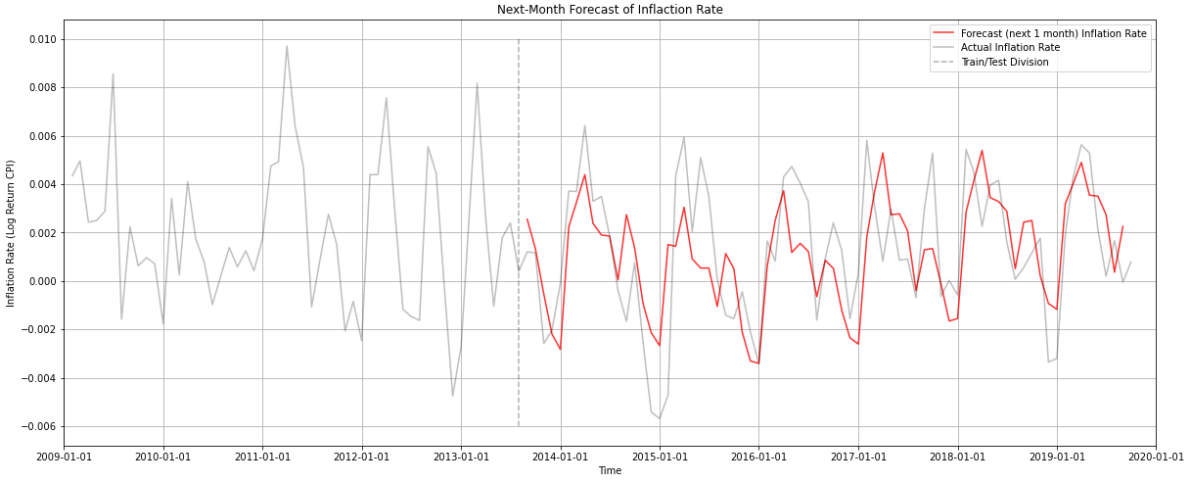

Figure 9 – Actual inflation rate (black) against the 1 month-ahead forecast (red).

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 various AR(p) models with p ranging from 0 to 3. If we calculate the RMSE score for the validation set for all AR(p) models, we will note that the AR(0) model is the one that generates the lower RMSE and, therefore, gives the best predictions, which confirms the assumption that the best auto-regressive term "p" is zero.

- AR(0) model - RMSE: 0.00244;
- AR(1) model - RMSE: 0.00257;
- AR(2) model - RMSE: 0.00263;
- AR(3) model - RMSE: 0.00265;

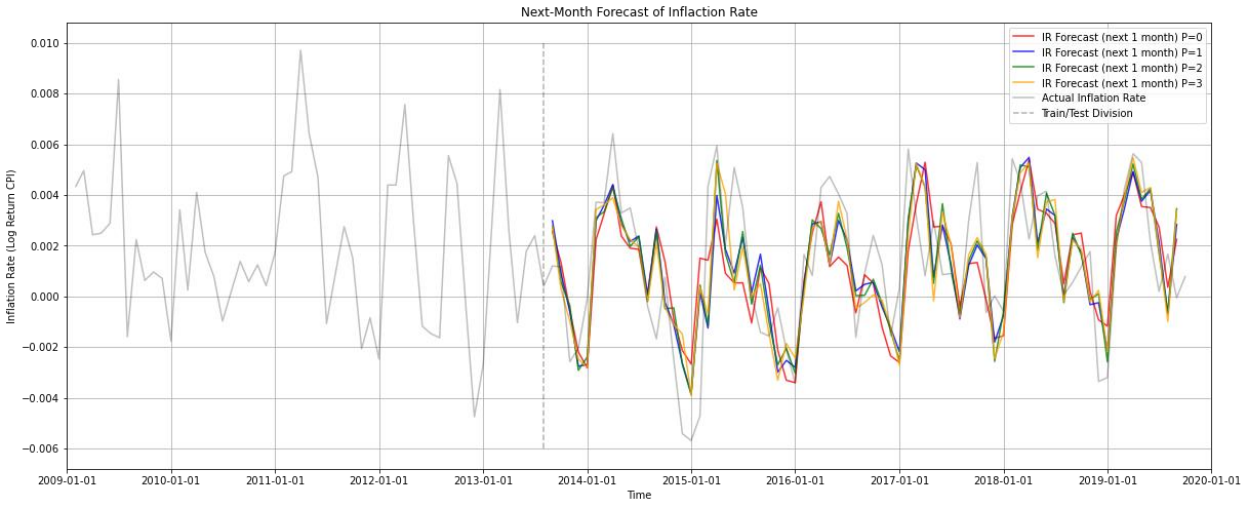

Figure 10 – Actual inflation rate (black) against the 1-month-ahead forecast from multiple AR(p) models.

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).

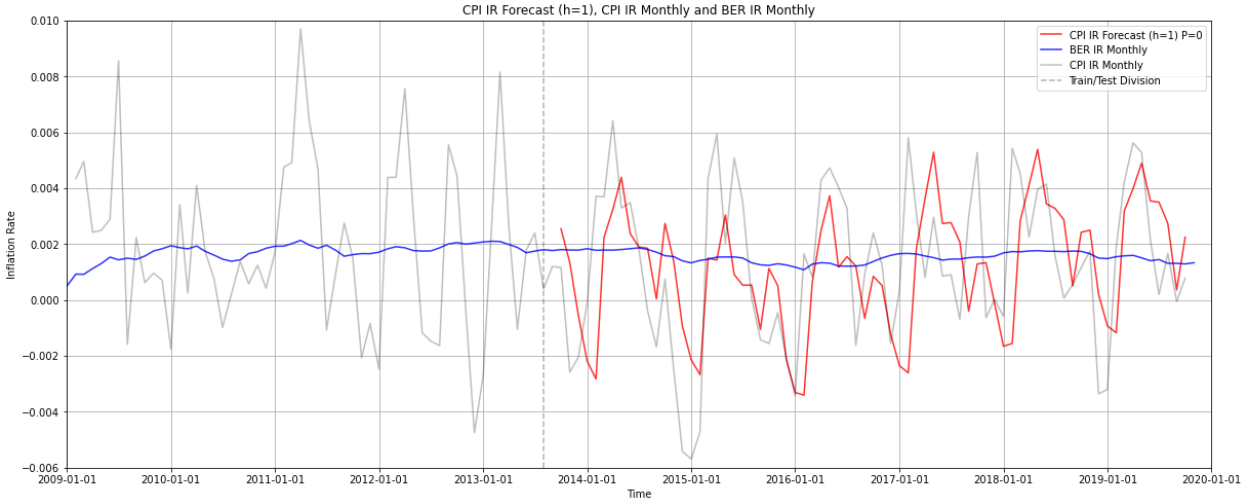

Figure 11 – CPI IR Forecast (h=1), CPI IR Monthly and BER IR Monthly. CPI and BER inflation rate series shown in the same y-axis.

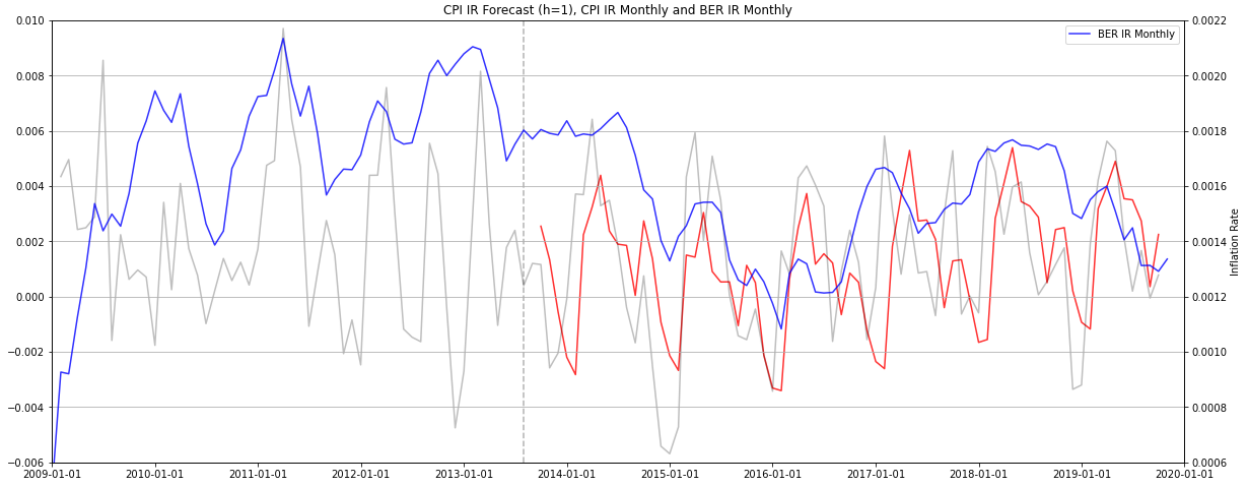

Figure 12 – CPI IR Forecast (h=1), CPI IR Monthly and BER IR Monthly. CPI and BER inflation rate series shown in different y-axis.

▪ External Regressors and Model Improvements

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 X_t by*

$$X_t = \sum_{i=1}^{p} \emptyset_i X_{t-i} + \varphi Y_{t-r} + W_t$$

where Y_t is the BER inflation rate at time t, $r \geq 0$ is the lag of BER rate w.r.t. CPI rate, and W_t 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, $CPI\ IR_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + P_i$, 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: 6.23e-11;
- p-value ADF test BER IR residuals: 3.38e-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 for lags 1 and 2, that is $r_1 = 1$ and $r_2 = 2$.

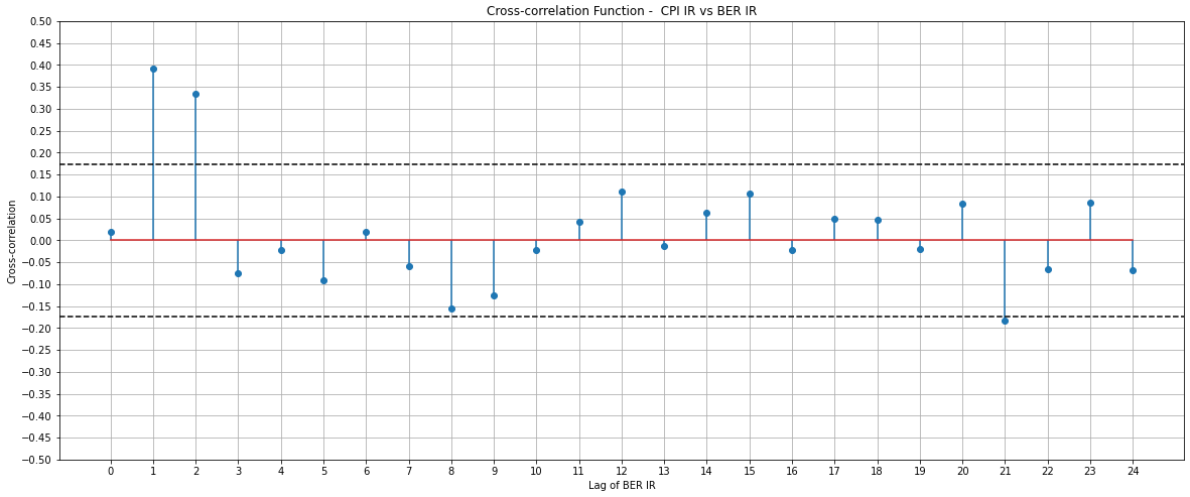

Figure 13 – Cross correlation function between CPI IR (residuals) and BER IR (residuals).

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 BER IR is a model with seasonal component and the lag 1 and lag 2 series of BER IR as described in the equation below. It is interesting to mention that the model gave a good result even though no polynomial trend was used to detrend the series. The reason is probably because the exogenous BER IR variable is accounting for the trend in the model which eliminates the use of polynomial trends.

$$CPI_t = 12.8836Y_{t-1} - 11.9676Y_{t-2} + P_i$$

Where,

- $CPI_{t,T}$: CPI inflation rate at time "t";
- Y_{t-1}: BER monthly inflation rate at time "t-1" (lag 1);
- Y_{t-2}: BER monthly inflation rate at time "t-2" (lag 2);

The plot of the CPI IR forecast using the model mentioned above (red curve) on top of the actual CPI IR (black curve) can be seen in **Figure 14** below. The image also shows the forecast of the previous quadratic trend model (green curve). Looking at the forecast, it is noticeable that both forecasts approximate the original curve reasonably well.

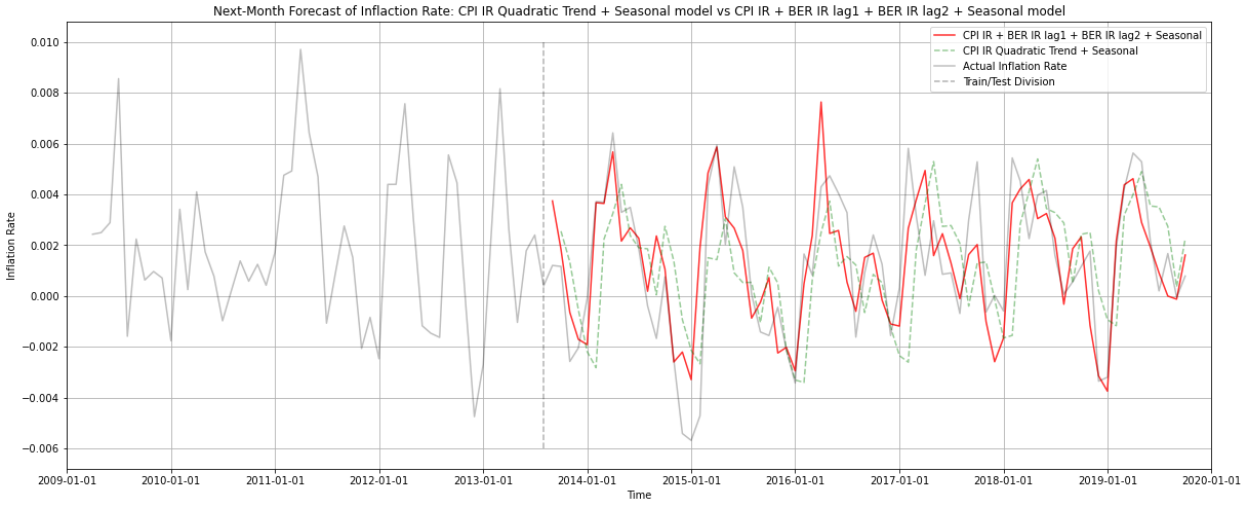

Figure 14 – 1-month-ahead forecasts of the CPI inflation rate series using a quadratic trend model with exogenous variable.

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

Solution:

The RMSE values obtained for the forecast is shown below. It can be seen that in terms of RMSE, the model with exogenous variables have a better performance.

- Quadratic Trend + Seasonal Component model - RMSE: 0.00244;
- BER IR lag1 + BER IR lag2 + Seasonal Component model - RMSE: 0.00179;

▪ Improving your model

(5 points) What other steps can you take to improve your model from part III? What is the smallest prediction error you can obtain? Describe the model that performs best. You might consider including MA terms, adding a seasonal AR term, or adding multiple daily values (or values from different months) of BER data as external regressors.

Solution:

During the process to improve the model, several approaches were taken considering:

- Including MA terms to the original ARIMA model;
- Including AR and MA terms using both ARIMA and SARIMA models;
- Adding other lagged terms of the BER IR feature;

In the end, the process that provided the best result was model with BER IR term from lag 1 to 3 and additional seasonal component. The final RMSE of this model was 0.00168 whereas the second best RMSE was 0.00179 (BER IR term from lag 1 to 2 and additional seasonal component). **Figure 15** shows the 1-month-ahead forecast of the main models tested. The dark blue curve indicates the performance of the best model.

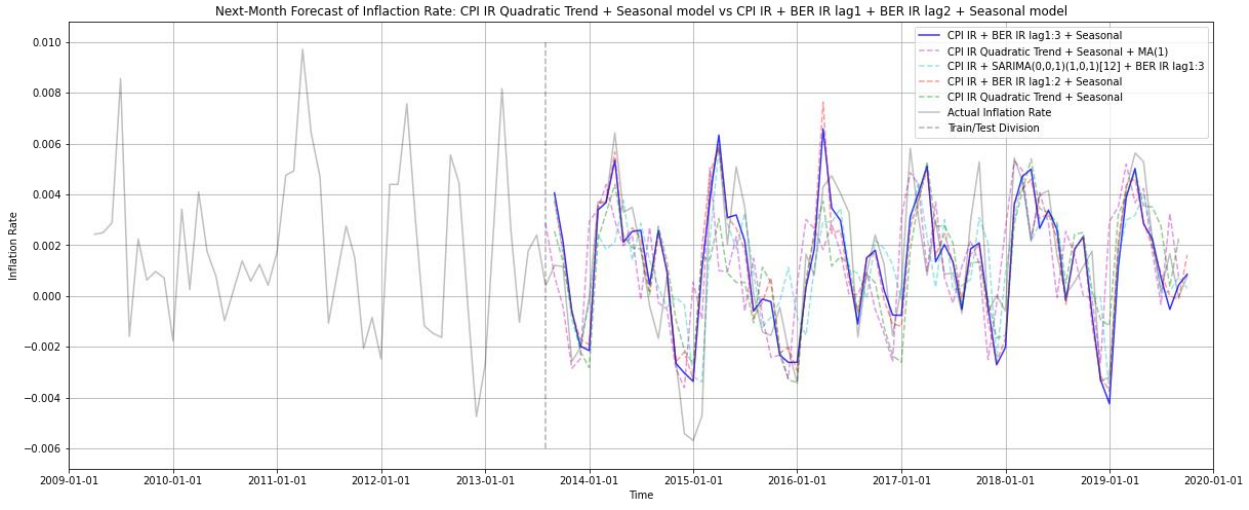

Figure 15 – 1-month-ahead forecasts of the CPI inflation rate series for various models.

Reference

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