Written Report – 6.419x Module 4

Name: ajland

Problem 2

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: The periodic signal, P_i, is defined as the interpolation of the average monthly residuals. It can be computed by taking $C_i - F_n(t_i)$ for each month, i, then aggregating and averaging by month. The result is plotted in Figure 1.

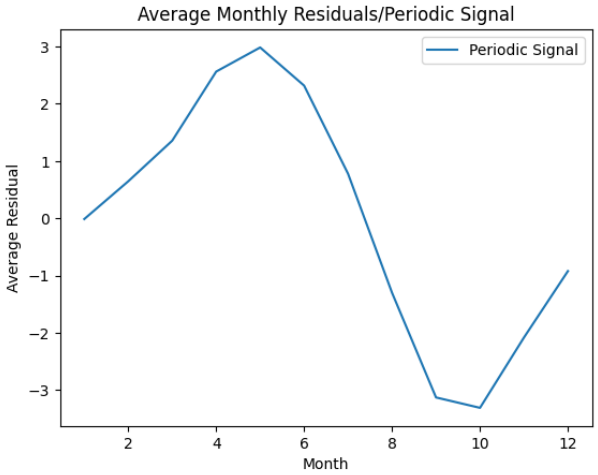

Figure 1: Average monthly residuals with linear interpolation between adjacent points

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: Taking the best quadratic fit and adding the periodic component for each month observed in Figure 1 results in the plot shown in Figure 2.

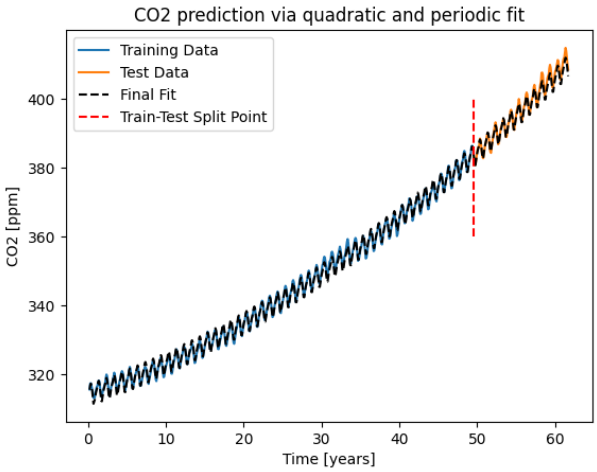

Figure 2: Quadratic (order 2) trend plus periodic component overlaid on training and test data

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 RMSE and MAPE for the final fit was 1.15 and 0.21, respectively. Comparatively, RMSE and MAPE for the quadratic trend was 2.50 and 0.53, respectively. In this way, both measures decreased by more than half their original value, indicating that the final fit is indeed an improvement.

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 CO_2 concentration meaningful? (Maximum 200 words.)

Solution: The ratio of the range of F to the amplitude of P_i is 32. The ratio of the amplitude of P_i to the range of R_i is 0.632. The first tells us that the trend spans values significantly larger than the periodic variation accounts for, meaning that the CO_2 concentration depends on more than seasonal variation can account for. The second ratio is harder to interpret, but if the amplitude of P_i and range of R_i represent kinds of variance in the data, then the ratio of the two indicates that the variation due to noise is greater than the seasonal, or periodic, variation by about 37%. However, it may be more appropriate to take the ratio of the two ranges for

this latter scenario, which yields 1.33, and would indicate that periodic variation is greater than random variation.

3. Autocovariance Functions

1. (4 points) Consider the MA(1) model,

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

where $\{W_t\} \sim W \sim \mathcal{N}(0, \sigma^2)$. *Find the autocovariance function of* $\{X_t\}$. *Include all important steps of your computations in your report.*

Solution:

$$\begin{aligned} \gamma_x(t, t-1) &= \mathsf{Cov}(X_t, X_{t-1}) \\ &= \mathsf{Cov}(W_t + \theta W_{t-1}, W_{t-1} + \theta W_{t-2}) \\ &= \mathsf{Cov}(W_t, W_{t-1}) + \mathsf{Cov}(W_t, \theta W_{t-2}) + ... \\ &\quad \mathsf{Cov}(\theta W_{t-1}, W_{t-1}) + \mathsf{Cov}(\theta W_{t-1}, \theta W_{t-2}) \\ &= \mathsf{Cov}(\theta W_{t-1}, W_{t-1}) \\ &= \theta \sigma^2 \end{aligned} \tag{1}$$

The ACF is zero at all gaps greater than 1 and σ^2 for a gap of 0.

2. (4 points) Consider the AR(1) model,

$$X_t = \phi X_{t-1} + W_t,$$

where $\{W_t\} \sim W \sim \mathcal{N}(0, \sigma^2)$. *Suppose* $|\phi| < 1$. *Find the autocovariance function of* $\{X_t\}$. *(You may use, without proving, the fact that* $\{X_t\}$ *is stationary if* $|\phi| < 1$.*) Include all important steps of your computations in your report.*

Solution: Since an **AR**(1) is stationary, the ACF is a function of the gap,

$$
\begin{aligned}
\gamma_x(t-s) &= \mathsf{Cov}(X_t, X_{t-s}) \\
&= \mathsf{Cov}(\phi X_{t-1} + W_t, X_{t-s}) \\
&= \mathsf{Cov}(\phi(\phi X_{t-2} + W_{t-1}) + W_t), X_{t-s}) \\
&= \mathsf{Cov}(\phi(\phi(\phi(\ldots\phi(\phi X_{t-s} + W_{t-s-1})))) + W_t), X_{t-s}) \\
&= \mathsf{Cov}(\phi^s X_{t-s} + \sum_{h=0}^{s-1} \phi^h W_{t-h}, X_{t-s}) \\
&= \mathsf{Cov}(\phi^s X_{t-s}, X_{t-s}) \\
&= \phi^s \mathsf{Var}(X_{t-s}) \\
&= \phi^s \mathsf{Var}(X_t)
\end{aligned}
\tag{2}
$$

where the correlation between all cross-term noises and X_{t-s} are 0 since the noise terms come after X_{t-s} and are thus independent of it.

5. Converting to Inflation Rates

1. Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI. Your response should include:

- *(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.)*
- *(2 points) Description of how the data has been detrended and a plot of the detrended data.*
- *(3 points) Statement of and justification for the chosen $AR(p)$ model. Include plots and reasoning.*
- *(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: The monthly inflation rate can be computed using the following forumula,

$$
IR_t = \frac{\text{CPI}_t - \text{CPI}_{t-1}}{\text{CPI}_{t-1}} \tag{3}
$$

where the first timestamp will have no rate value. Computing the inflation rate and plotting yields Figure 3. The line $y = 0$ is shown to illustrate that the data is not zero-centered. Several transformations then linear regressions were performed on the data with results reported in Table 1.

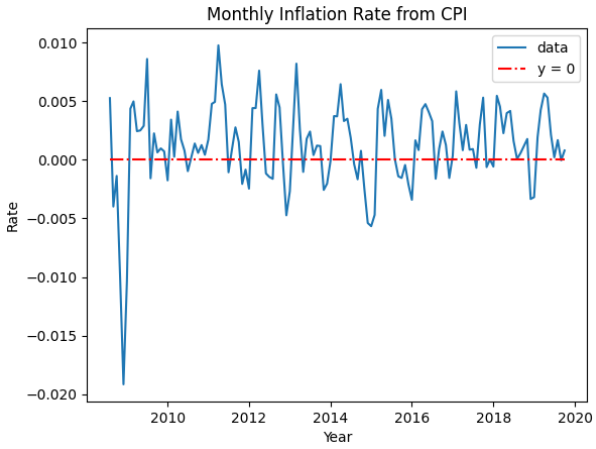

Figure 3: Raw inflation rate data over several years

Table 1: Error reporting for various fits

Transformation	Linear	Quadratic	Sqrt	Log
RMSE	0.003039	0.002869	0.002916	0.002861
MAPE	148.3	431.7	187.5	235.3

A log transformation yields the best RMSE and appears to be the most zero-centered after removing the trend. The result is shown in Figure 4. Note that regression was performed only on data after the large drop below zero, which corresponds to February 2009 and later.

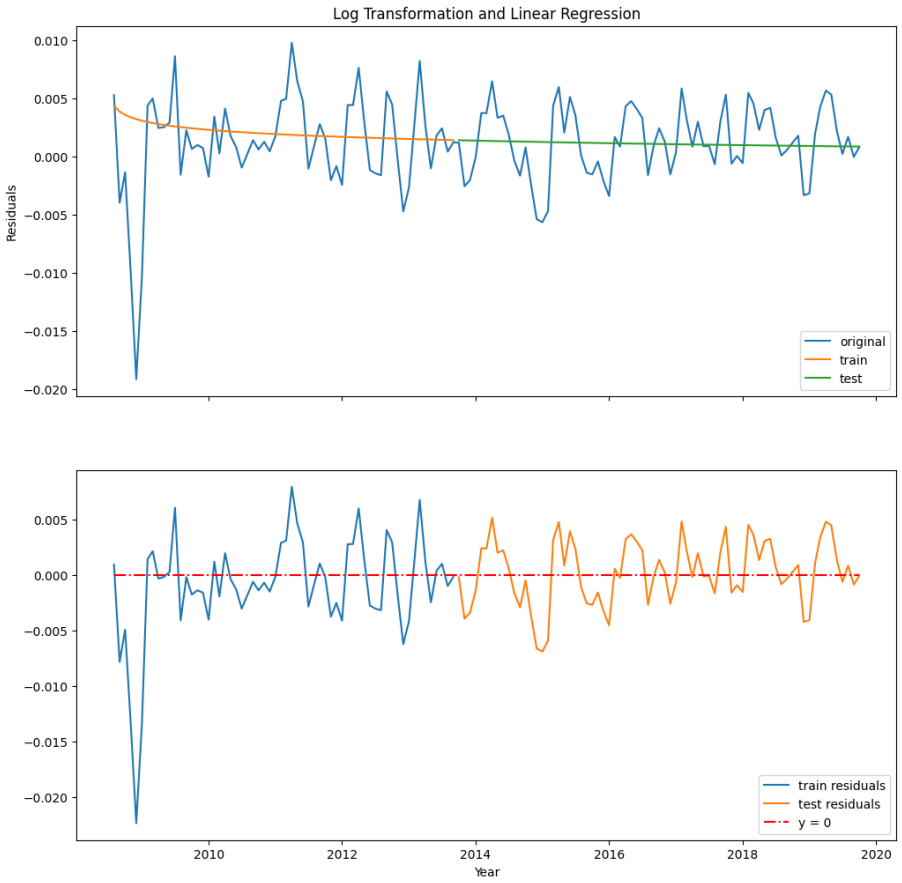

Figure 4: (Top) original data with log transformed regression line overlaid; (Bottom) Residuals with y=0 plotted for reference

For this analysis, I will choose an AR(1) model since the ACF plot shows a decay pattern and the PACF plot inicates non-significant correlation after one lag (see Figure 5).

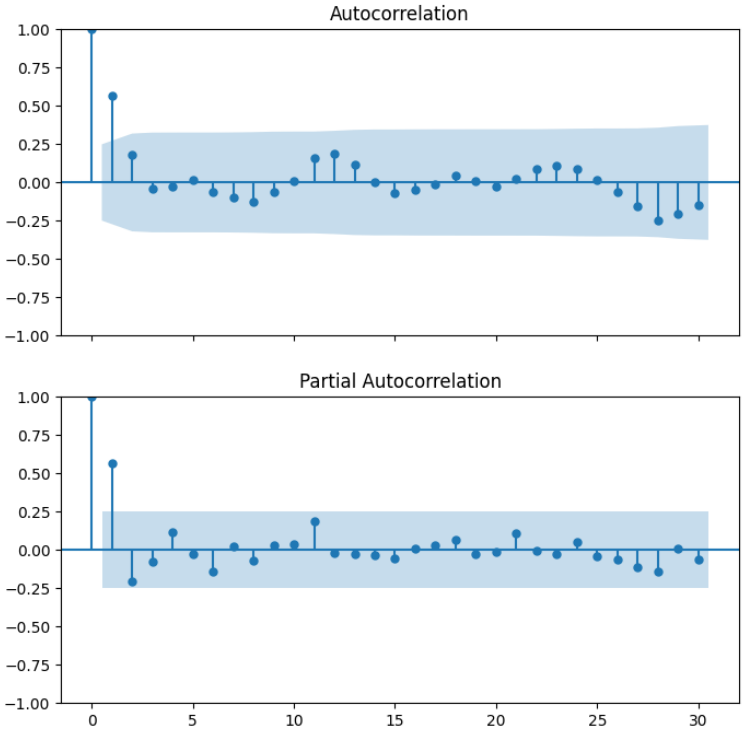

Figure 5: ACF and PACF plot with up to 30 lags shown. ACF plot indicates an auto-regressive model and the PACF plot indicates order $p = 2$.

After fitting the trend and AR(1) components, the model can be summarized by the following equation,

$$IR_i = F(t_i) + S_i + R_i$$

where IR_i is the inflation rate, $F(t_i)$ is the trend, S_i is the auto-regressive/seasonal component and R_i is the random noise, all at time-step i. The plot of IR is overlaid on the sum of fitted components, $F(t_i) + S_i$, in Figure 6. The RMSE is reduced from 0.002861 to 0.002405.

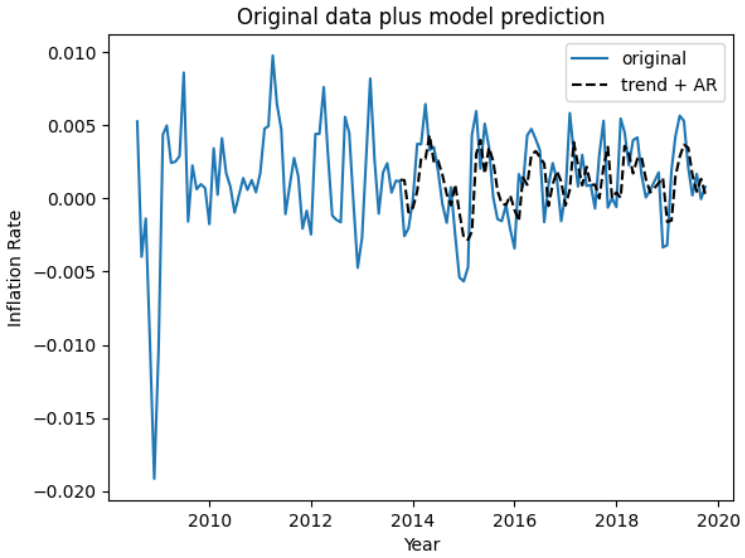

Figure 6: Plot showing original inflation rate data with 1 month-ahead forecast for validation data overlaid.

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: A lag of $p = 2$ in the AR(p) model yields the best RMSE of 0.002342. Referring back to the PACF plot in Figure 5, this may not be surprising as the lag 2 correlation was nearly considered significant. In fact, using a lag of 2, 3 or 4 yields a better RMSE than lag 1, as shown in Figure 7. This indicates that the PACF is merely one tool for determining the optimal lag value.

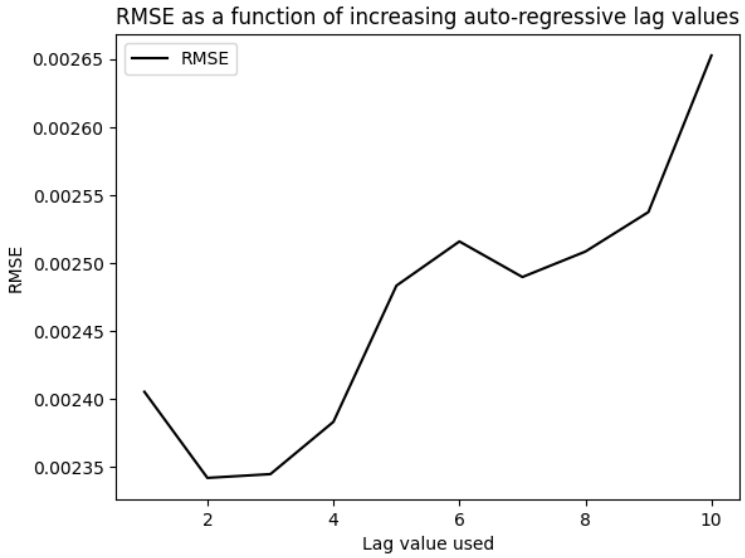

Figure 7: RMSE is minimized at lag $p = 2$, indicating an AR(2) model yields the best fit.

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 datasets, plus the prediction, over time from September 2013 onward.)

Solution: Figure 8 contains a plot of the inflation rates derived from botht the CPI and BER dataset, along with the monthly inflation rate estimates from before. Notably, the BER derived inflation rates are less noisy than the CPI derived rates. This may be due to taking the average of all monthly BER values as the representative as opposed to the first monthly value like we did for the CPI. Additionally, the formula for converting to inflation rates for both methods is different. The IR_BER line appears to lie near the center of the other two plots.

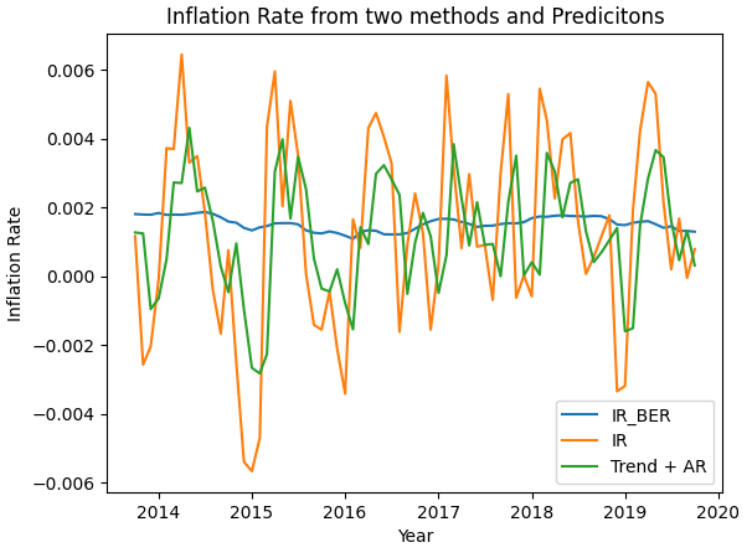

Figure 8: Inflation computed from CPI and BER datasets, plus the 1 month-ahead predictions from the CPI derived inflation rate.

Problem 6. External Regressors and Model Improvements (Written Report)

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

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \psi Y_{t-r} + W_t,$$

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. ***Note:*** *In general, multiple external terms $\sum_{i=1}^{m} \psi_i Y_{t-r_i}$ can be incorporated in the model if there are multiple peaks in CCF plots.)*

Solution: The cross-correlation function is computed using the BER based inflation rate as the sliding time-series and the CPI based IR as the fixed. The result is shown in Figure 9. The CCF is maximized at a lag of 1. Note that the BER_IR

window was reduced to match the time interval of the CPI_IR data; that is, both time series' were aligned and range from 2008-08 to 2019-10.

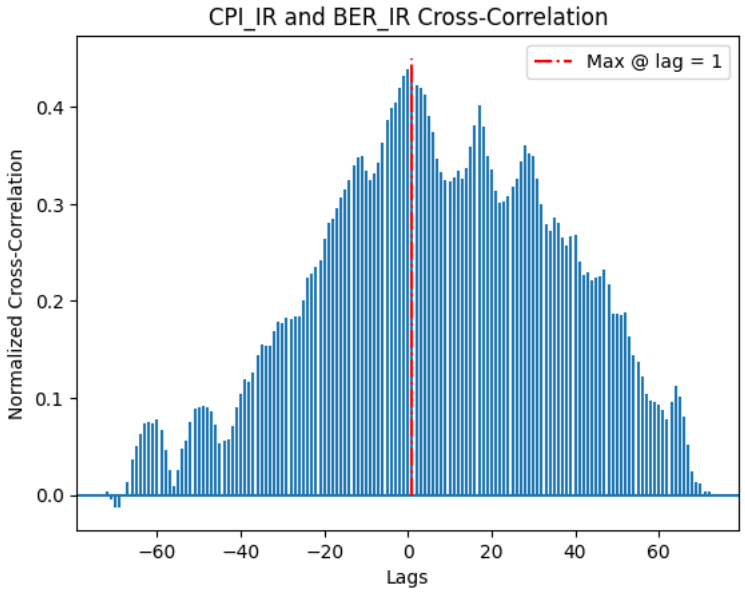

Figure 9: Cross-correlation with BER_IR set as the sliding time-series

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. Python Tip: You may use sm.tsa.statespace.SARIMAX.

Solution: Using the SARMIAX model, the CPI_IR is regressed onto the BER_IR with order (1,0,0). The 1-month ahead forecast along with original CPI_IR data are shown in Figure 10. The coefficients for the regression are given in Table 2.

Table 2: SARIMAX(1,0,0) coefficients

	coef	std err	z	P>|z|	[0.025	0.975]
IR_ber	0.9689	0.648	1.496	0.135	-0.301	2.239
ar.L1	0.4915	0.082	6.011	0.000	0.331	0.652
sigma2	1.408e-05	1.94e-06	7.271	0.000	1.03e-05	1.79e-05

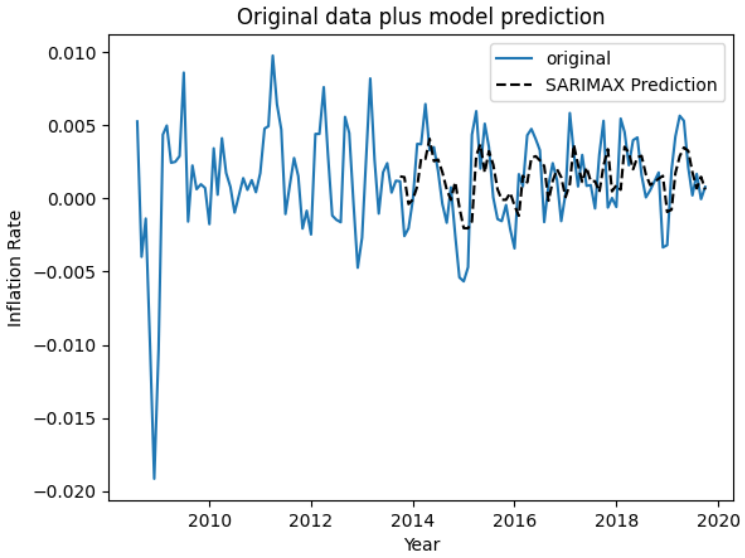

Figure 10: Caption

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

Solution: Using the new previously stated SARIMAX(1,0,0) model, the new 1 month-ahead RMSE is 0.002376, which is lower than the original AR(1) model (0.002405) but slightly higher than the AR(2) model (0.002342). As the AR(1) is more comparable to the SARMIMAX(1,0,0) model, it is safe to say that using the BER_IR as an external regressor improved the RMSE.

(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: There are several hyper-parameters in the SARIMAX model that are tunable and therefore can improve the prediction error. By performing a grid search over order (p, d, q) and seasonality parameters P and setting $s = 12$ (as the data is given monthly), I found an optima at (p, d, q) = (6, 0, 0), P = 3 and s = 12. These hyper-parameters yielded an RMSE of 0.002089. The final 1 month-ahead prediction values are overlaid on the original data in Figure 11 and the summary of the fit, along with model parameters, are shown in Figure 12.

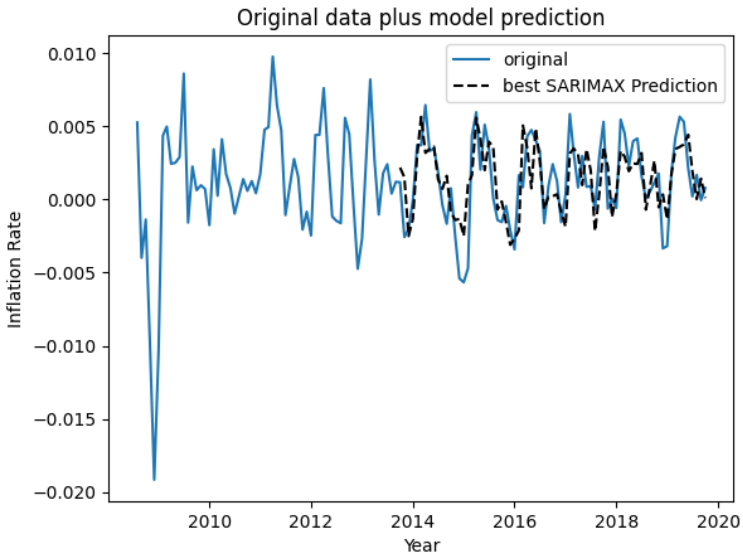

Figure 11: Best SARIMAX fit found via grid search over p, d, q and P hyperparameters with s=12. RMSE = 0.002089.

SARIMAX Results

Dep. Variable:	IR	No. Observations:	135
Model:	SARIMAX(6, 0, 0)x(3, 0, 0, 12)	Log Likelihood	597.558
Date:	Tue, 06 Aug 2024	AIC	-1173.115
Time:	19:52:45	BIC	-1141.157
Sample:	08-01-2008 - 10-01-2019	HQIC	-1160.128
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
IR_ber	1.0138	0.602	1.684	0.092	-0.166	2.194
ar.L1	0.4071	0.148	2.744	0.006	0.116	0.698
ar.L2	-0.0011	0.128	-0.008	0.993	-0.252	0.250
ar.L3	-0.1918	0.182	-1.052	0.293	-0.549	0.166
ar.L4	0.1192	0.197	0.606	0.545	-0.266	0.505
ar.L5	0.0197	0.194	0.102	0.919	-0.360	0.400
ar.L6	-0.3306	0.150	-2.205	0.027	-0.625	-0.037
ar.S.L12	0.4056	0.246	1.646	0.100	-0.077	0.889
ar.S.L24	0.0334	0.342	0.098	0.922	-0.637	0.704
ar.S.L36	0.1413	0.220	0.643	0.520	-0.289	0.572
sigma2	1.119e-05	2.58e-06	4.331	0.000	6.13e-06	1.63e-05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	33.14
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.30	Skew:	-0.65
Prob(H) (two-sided):	0.00	Kurtosis:	5.05

Figure 12: Best SARIMAX model fit summary

References