UNIVERSITÀ DEGLI STUDI DI MILANO

Data Science for Economics

Time Series and Forecasting

Final project

Guglielmo Berzano, id: 13532A

Nicolò Pignatelli, id: 13831A

E-mail: guglielmo.berzano@studenti.unimi.it

nicolo.pignatelli@studenti.unimi.it

December 2023

Abstract

This work aims to forecast in the most accurate way the *Real Gross Domestic Product* (GDP) of the United States in a time window going from 1985Q3 to 2018Q4 given data that go back to 1960Q2. Among the adopted models you will find Vector AutoRegressive (VAR), eXogenous AutoRegressive (AR-X) and simple AutoRegressive (AR) models. By looking at the root mean square error (RMSE) of the predictions, we found that it is not true that more complex models yield necessarily better results.

Contents

1 Data exploration

In this first section we will look at which series we are dealing with.

The first analysed series is the *Real Gross Domestic Product* taken in logs to compress the scale and make the trend, if any, more linear. The series is depicted in the following plot.

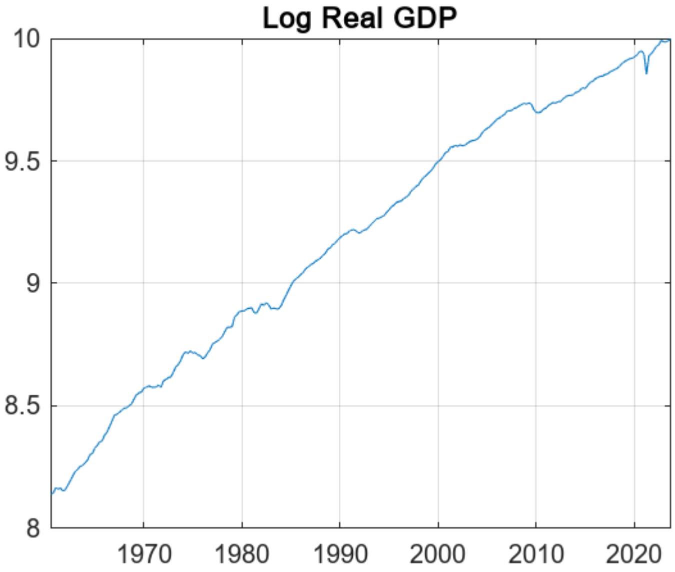

Figure 1: $ln(GDP)$ in real terms

Clearly the series presents a very persistent trend that will be analysed more in details later. We can also notice some structural breaks in conjunction with recessions like around 2010 and 2020.

In the following plot we look at the $\Delta ln(GDP)$, namely the first difference of the $ln(GDP)$. This is obtained simply by doing $\Delta y_t = ln(y_t) - ln(y_{t-1})$, where y_t represents the real GDP at time t. Results are depicted in Figure 2.

After having taken the first difference, the series looks almost stationary with mean slightly higher than 0. There are clearly two spikes due to the Covid-19 pandemic but these did not influence much the series since it came back to stationarity immediately afterwards.

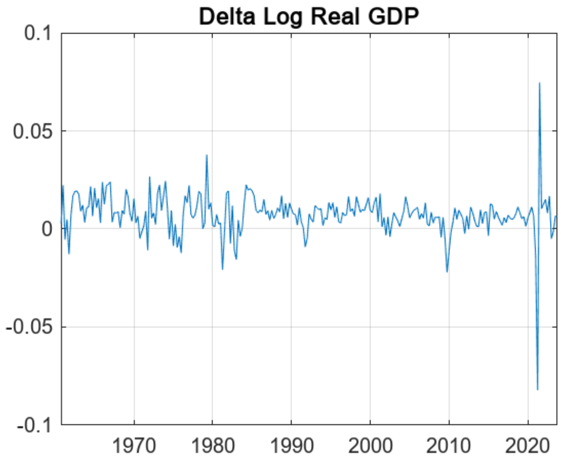

Figure 2: ΔLog GDP in real terms

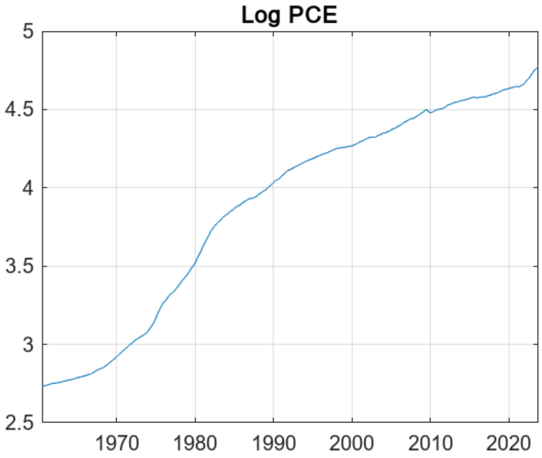

Figure 3: Log of Personal Consumption Expenditures

The next series is *Personal Consumption Expenditures*, shortened as *PCE* taken in logs. The plot is in Figure 3 and clearly there is no seasonality nor cycles but a non-linear trend.

The next series we will analyse is *inflation*, obtained by doing: $\pi_t = ln(PCE_t) - ln(PCE_{t-1})$, thus we compare the natural logarithm of the consumption at time t with time $t-1$. Results are summarised in Figure 4. It is not easy to clearly identify an overall pattern.

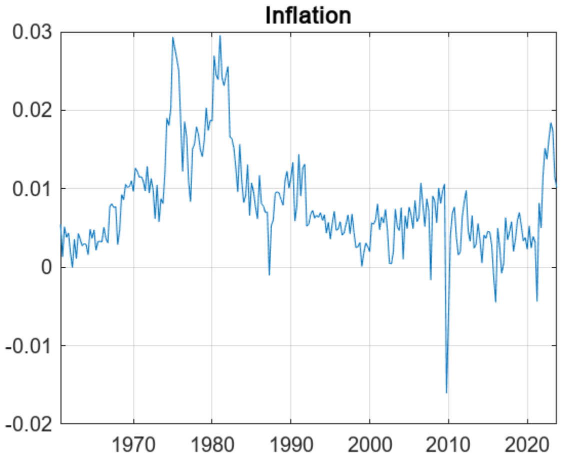

Figure 4: Inflation

The last series we analyse is the *Term Spread*, meaning the difference between the 10-years-rate of the treasury stocks and the 3-months treasury bill. This series is shown in Figure 5. Even though there are many fluctuations, these cannot be considered cycles since their persistence is almost none.

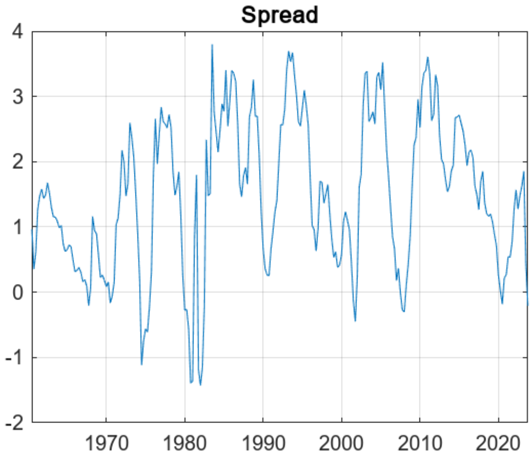

Figure 5: Term Spread

2 Vector AutoRegressive models

In this section we will implement first a VAR(1) model and then a VAR(p) model by choosing the parameter p of autoregressed lags based on the Akaike information criterion (AIC).

Variables entering the VAR models are the same for both models, the only change is the p parameter. These variables are:

- $\Delta ln(GDP)$
- Inflation: π_t
- Term spread: $Tspread_t$

The VAR(p) model will have the following formula:

$$y_t = \mu_t + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + U_t \tag{1}$$

Where:

$$y_t = \begin{bmatrix} \Delta ln(GDP) \\ \pi_t \\ Tspread_t \end{bmatrix}$$

is the vector of variables,

$$A_p = \begin{bmatrix} a_{p,11} & a_{p,12} & a_{p,13} \\ a_{p,21} & a_{p,22} & a_{p,23} \\ a_{p,31} & a_{p,32} & a_{p,33} \end{bmatrix}$$

is the matrix of coefficients and U_t is the vector of errors which should be distributed as a white noise with mean 0 and variance Σ_u.

To continue, we first look at their autocorrelation functions computed on a sample of 100 observations from 1960Q2 up to 1985Q2. Results are depicted in the following plots. The first shows that for $\Delta ln(GDP)$ the coefficients decay already after 2 periods, the second that *inflation* has a very persistent behavior and lastly the third that the behavior of *Tspread* is peculiar and so quite hard to interpret.

In the following analyses we forecast up until 2018Q4 for a total of 134 predictions.

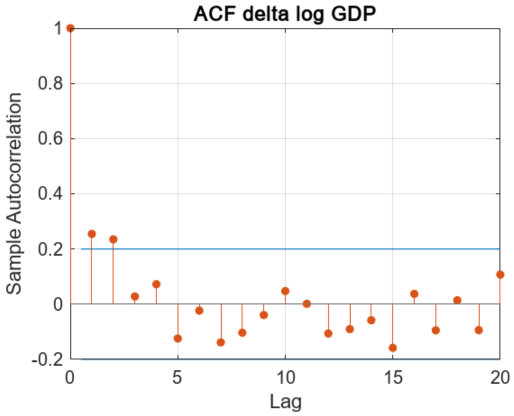

Figure 6: ACF $\Delta ln(GDP)$

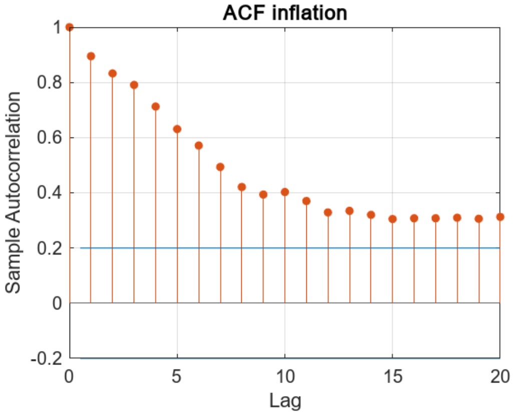

Figure 7: ACF Inflation π_t

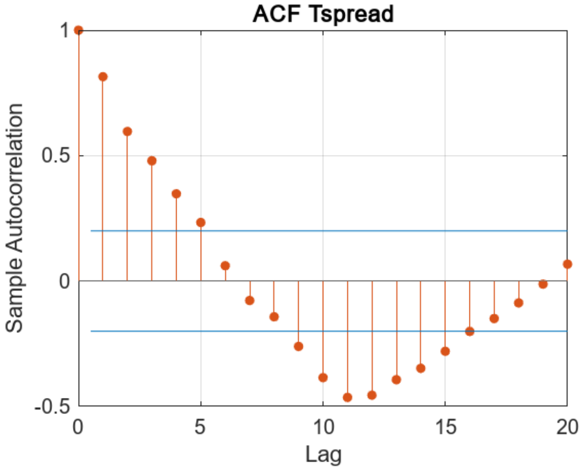

Figure 8: ACF $Tspread_t$

2.1 VAR(4)

We have first implemented a VAR(4) model obtaining a quite good result, with an $\boldsymbol{RMSE = 93.35}$. The plot is represented in Figure 9, residuals in the plots from 10 to 12 and ACFs and PACFs from plot 13 to 18. In general we can notice that the residuals seem to have mean equal to zero, which means that the model is not systematically over or under predicting, but there is some autocorrelation, so there is still some information that can be exploited.

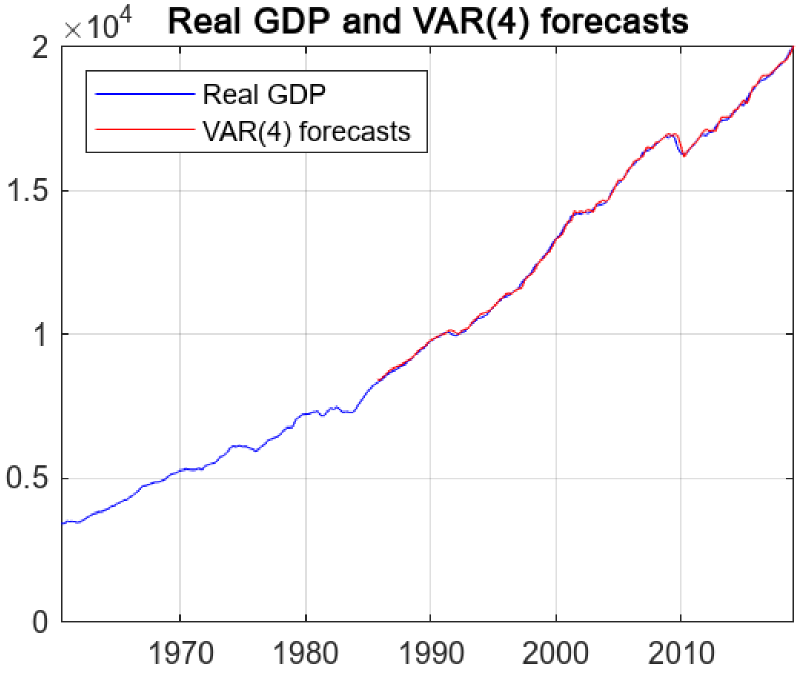

Figure 9: Vector AutoRegressive process of order 4

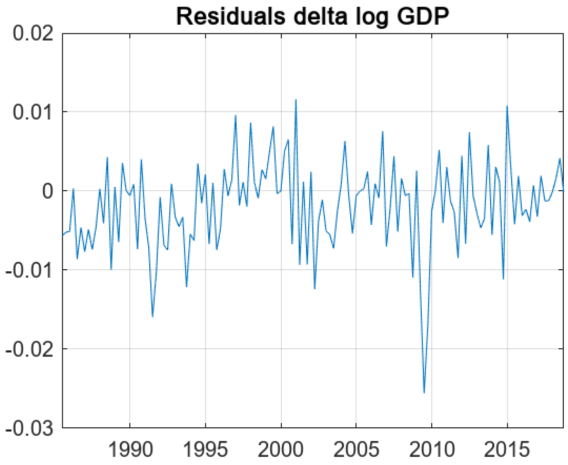

Figure 10: Residuals $\Delta ln(GDP)$

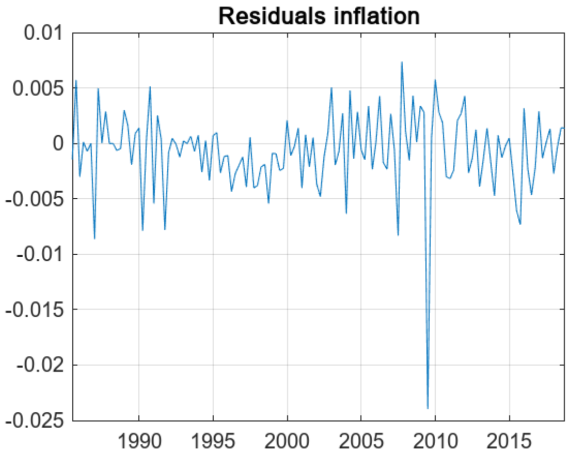

Figure 11: Residuals Inflation π_t

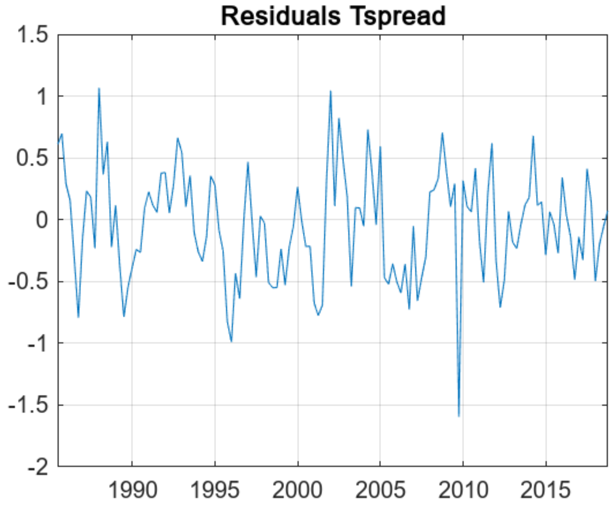

Figure 12: Residuals $Tspread_t$

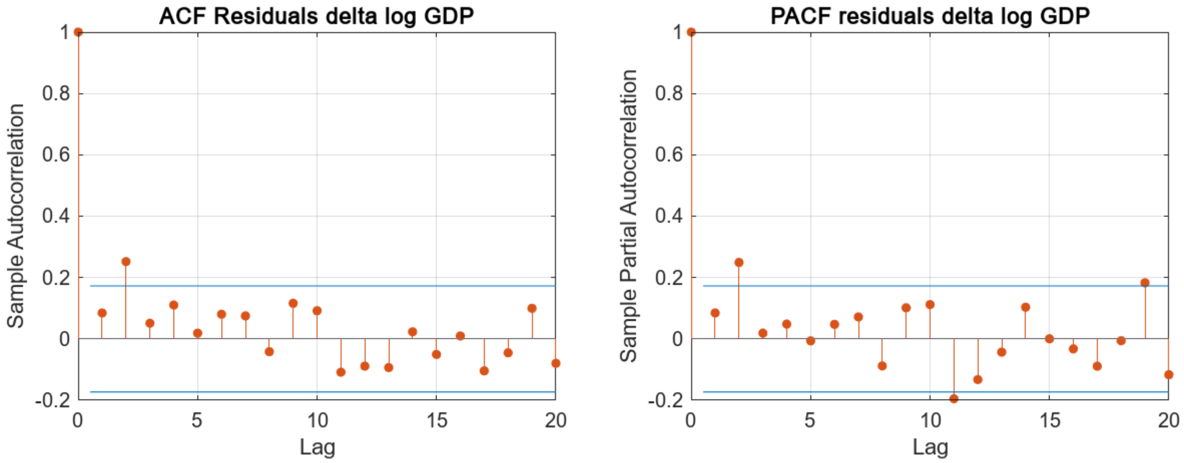

Figure 13: ACF residuals $\Delta ln(GDP)$

Figure 14: PACF residuals $\Delta ln(GDP)$

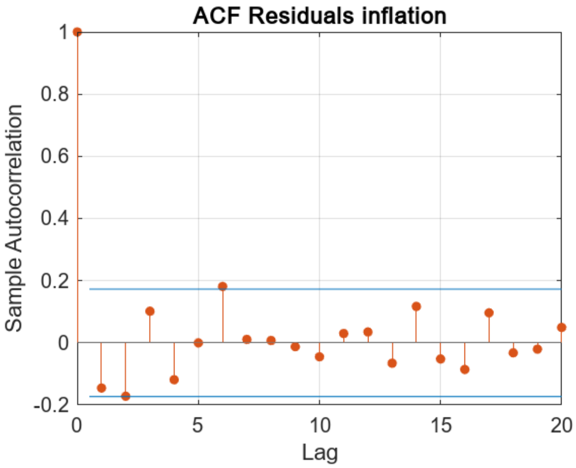

Figure 15: ACF residuals Inflation π_t

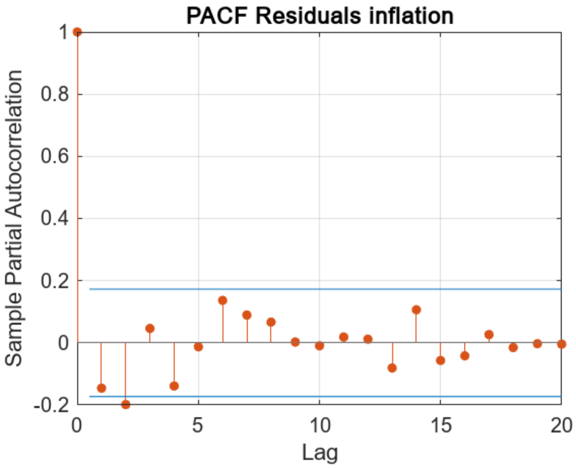

Figure 16: PACF residuals Inflation π_t

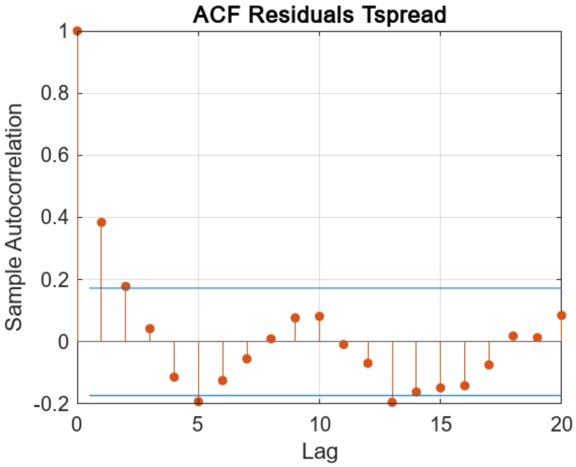

Figure 17: ACF residuals $Tspread_t$

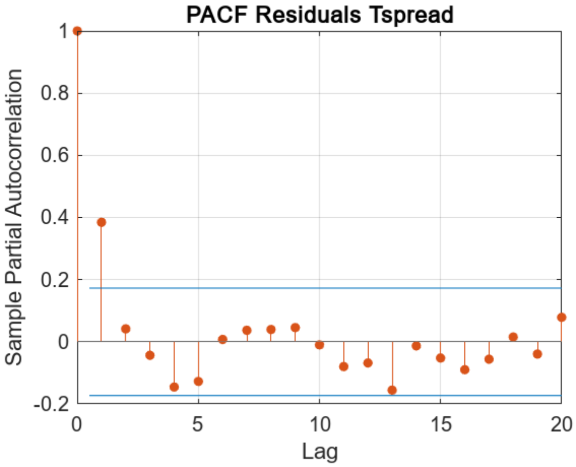

Figure 18: PACF residuals $Tspread_t$

2.2 VAR(p) with optimal p

In order to compute the optimal number of autoregressed lags p for this VAR model, we used the Akaike information criterion (AIC). We decided to do this in two ways: in the first case we decided to pick the best p according to the first 100 observsations, obteining a **VAR(1)** and then we decided to pick, at each iteration, a new p, obtaining a **VAR(p)** with varying p.

2.2.1 VAR(1)

According to the results of the AIC run on the first 100 observations, presented in Figure 19, the AIC selects $p = 1$ as the best lag order, with a value equal to -1287.90, so in the following analysis we will select a **VAR(1)**.

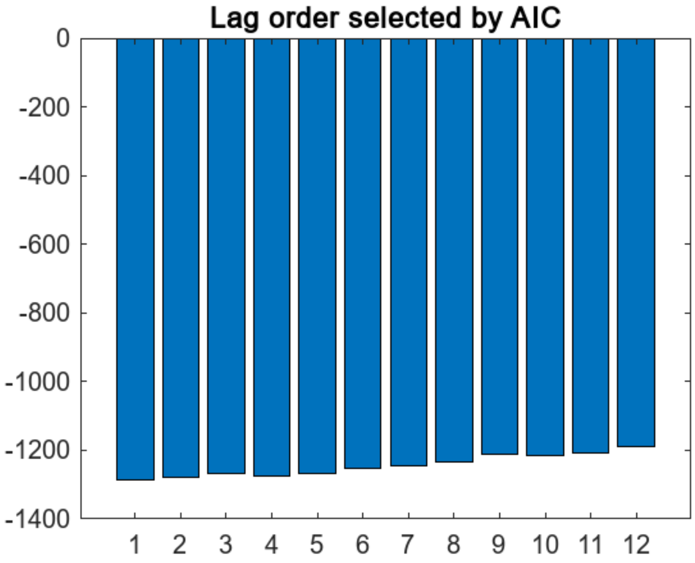

Figure 19: AIC value for each p for VAR models

By selecting $p = 1$ we get $\boldsymbol{RMSE = 84.99}$, thus thanks to AIC we have been able to slightly improve the model seen in subsection 2.1. The plot is represented in 20.

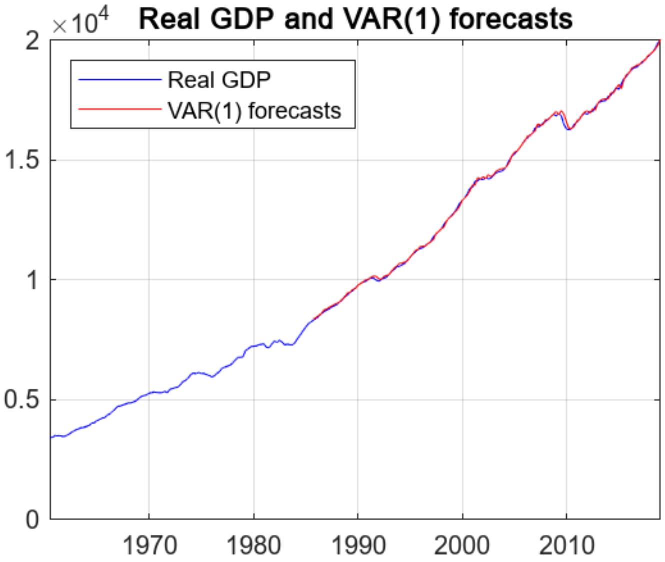

Figure 20: Var(1) forecasts

2.2.2 VAR(p) with varying p

In this case, instead of doing as we did before, we chose a different p for each iteration according to the best value computed by AIC. If we look at the values of p, we will find that 1 is the most chosen but it is not the only one. Strangely the $\boldsymbol{RMSE = 85.84}$, thus it is worse than its VAR(1) counterpart. The plot is the following:

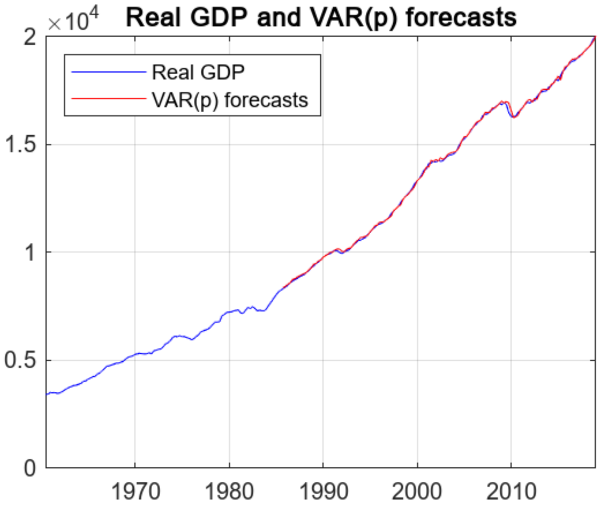

Figure 21: VAR(p) with varying p

3 Principal component analysis and AR(4)-X model

In order to improve the results of the previous models, we can analyse the relation between *GDP* and many economic factors, 245 to be precise. To do this, we used a data dimensionality reduction technique known as *principal component analysis*, shortened as *PCA*.

In particular we analysed the value of the principal components obtained by applying the PCA on just the first 100 observations. After having obtained the values, we computed the eigenvalues for each PC and only selected the ones whose eigenvalues were greater than 1, meaning that they are able to explain more than one variable. By doing so we have obtained a dataset with just 41 columns, showing a great improvement with respect to the initial 235-column dataset. In the following plot we presented how much variablity the first 10 principal components are able to explain of the original dataset in percentage. We notice that the first PC explains about 22% of the total variability and, in particular, the variables that contributed the most to the creation of this principal component can be checked in Figure 23.

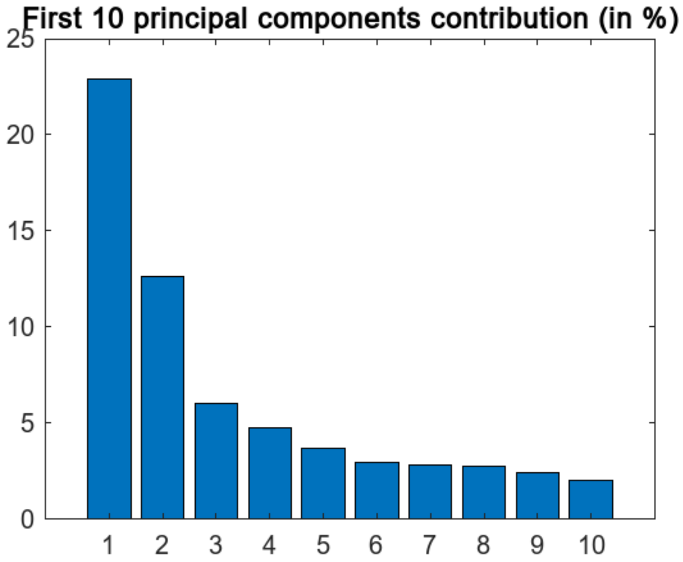

Figure 22: Contribution in percentage of the first ten principal components

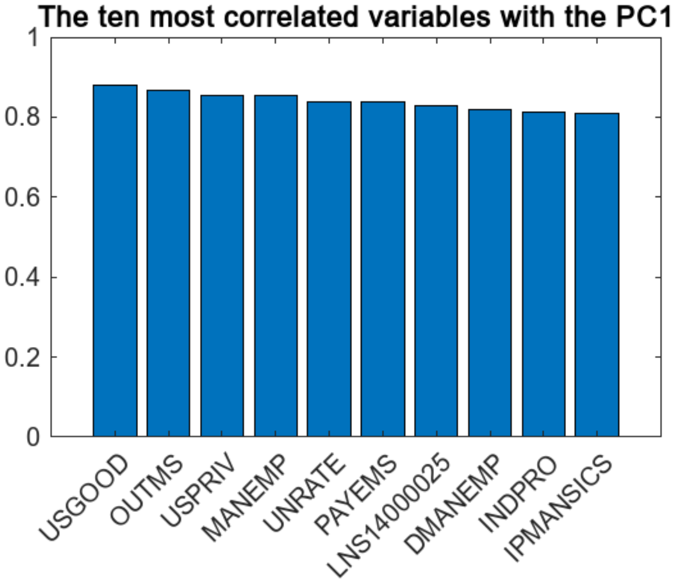

Figure 23: Most contributing variables to the first principal component

We will now proceed similarly to what we did in section 2 but in this case, instead of using variables like inflation or the term spread, we use the values computed by the first principal component. In particular, we implemented an autoregressive model with $p = 4$ with the addition of an exogenous variable, namely the first pc. From here, we call this model AR(4)-X and has the formula:

$$\Delta ln(GDP)_t = a + \sum_{j=1}^{4} b_j \Delta ln(GDP)_{t-j} + c\hat{F}_{1,t-1} \tag{2}$$

The plot is represented in Figure 24 and presents the best error so far with an ***RMSE = 84.60***.

In this case the prediction follows closely the real values but seems slow at catching up with the 2008 financial crash.

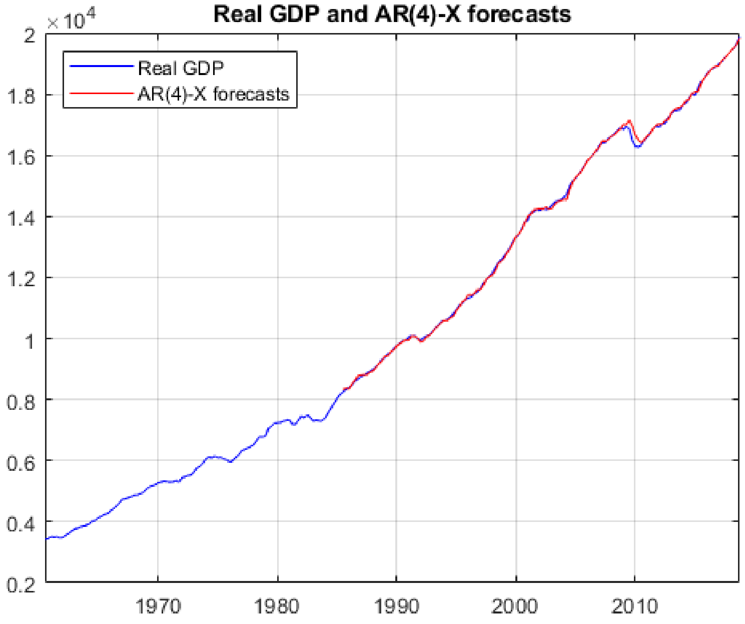

Figure 24: AR(4)-X model

4 Other models

In this section we present and briefly comment all the models we estimated.

4.1 Random Walk

The first model we estimated is the Random Walk. The formula is the following:

$$ln(GDP)_t = ln(GDP)_{t-1} + u_t \tag{3}$$

with $u_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

The result is poor with respect to the previous ones, indeed ***RMSE = 120.06***. Probably the model is too simple. The plot is the following:

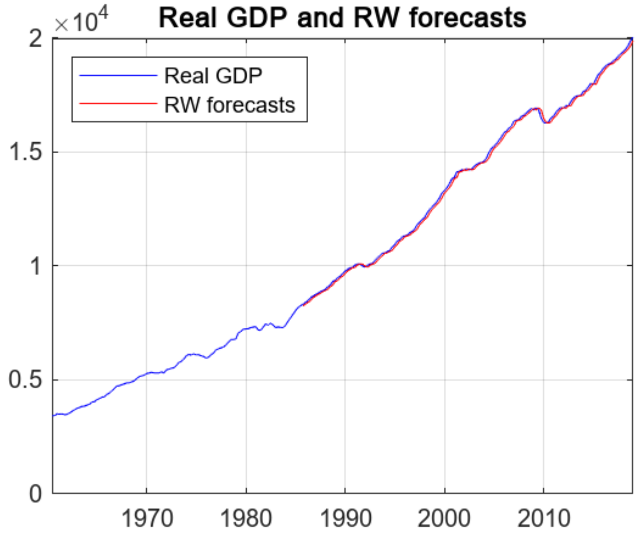

Figure 25: Random Walk forecasts

4.2 AR(4)

The second model we estimated is an AutoRegressive process of order 4, in short AR(4). The formula is the following:

$$\Delta ln(GDP)_t = a + \sum_{j=1}^{4} b_j \Delta ln(GDP)_{t-j} + u_t \tag{4}$$

with $u_t \sim WN(0, \sigma^2)$.

This time the result is the best one, indeed $\boldsymbol{RMSE = 79.34}$ and the plot is the following:

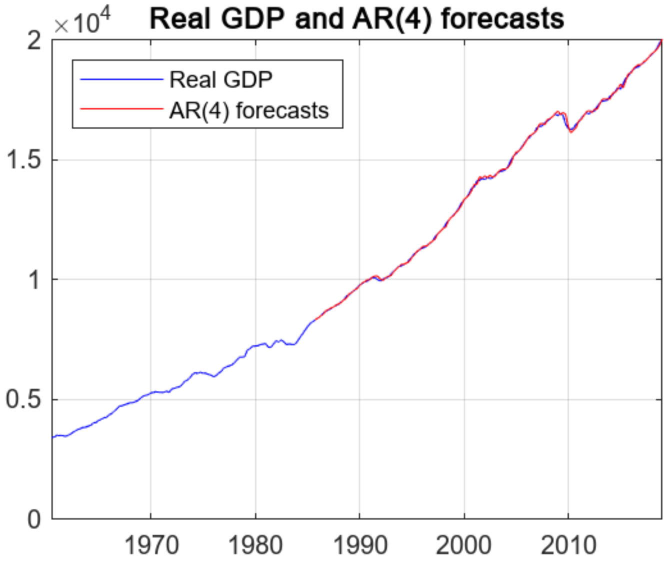

Figure 26: AR(4) forecasts

4.3 Proposed model

To sum up, the best model we obtained was the previous one, the AR(4). We thought that we could improve our forecasting accuracy by again, as we did for the VAR models, using AIC for estimating p. We computed it for all the models up until $p = 8$. We decided to choose p according only to the first 100 observations, as we did for obtaining the best VAR model. The AIC is presented in Figure 27.

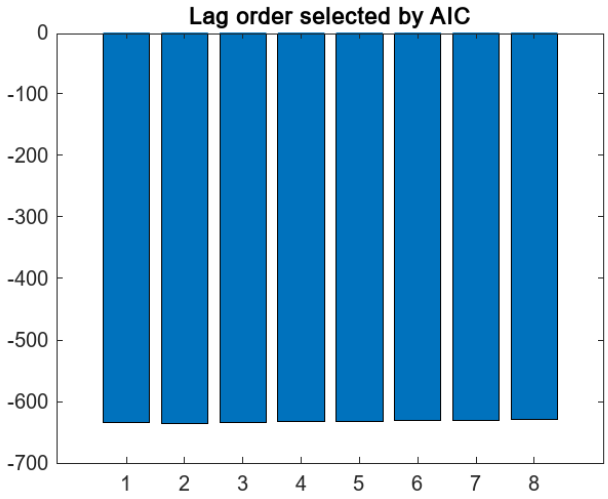

Figure 27: AIC value for each p for AR models

The minimum value for the AIC, -635.67, is obtained with $p = 2$, so we chose to estimate an AR(2) whose formula is the following:

$$\Delta ln(GDP)_t = a + \sum_{j=1}^{2} b_j \Delta ln(GDP)_{t-j} + u_t \tag{5}$$

with $u_t \sim WN(0, \sigma^2)$.

As expected, this model is the best, indeed $\boldsymbol{RMSE = 77.05}$ and the plot is:

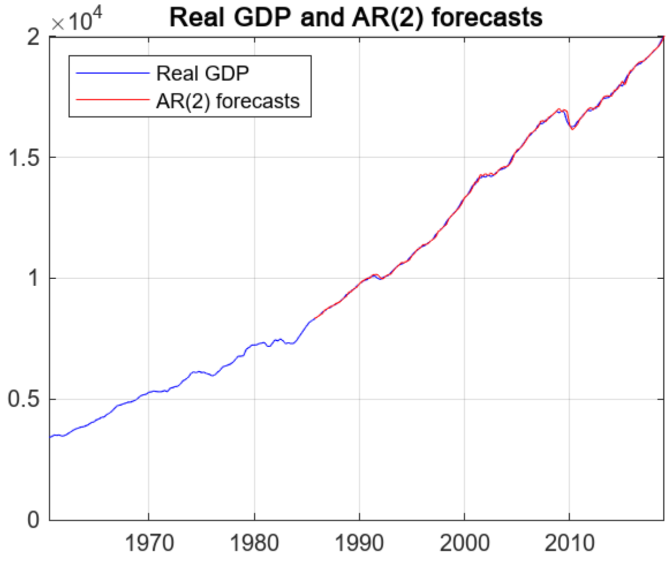

Figure 28: AR(2) forecasts

Moreover, we checked the mean of the residuals and we found that it is extremely close to zero with no significant spikes in either the ACF or the PACF.

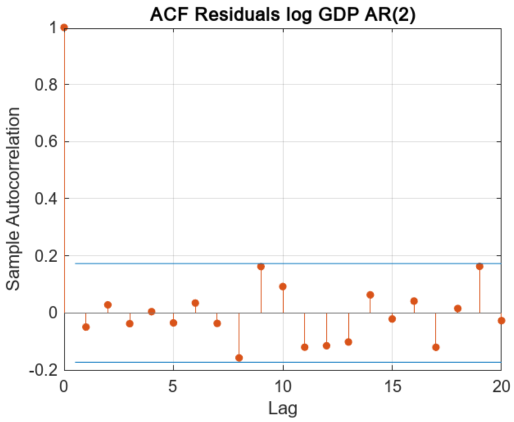

Figure 29: ACF Residuals AR(2)

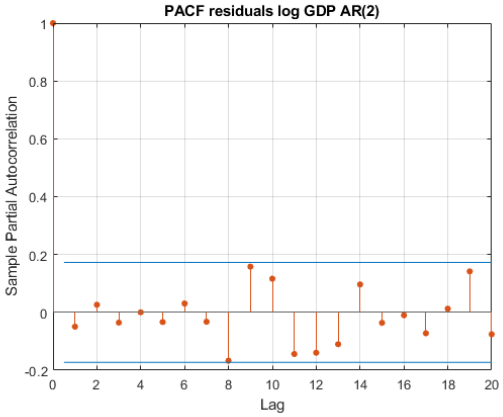

Figure 30: PACF Residuals AR(2)

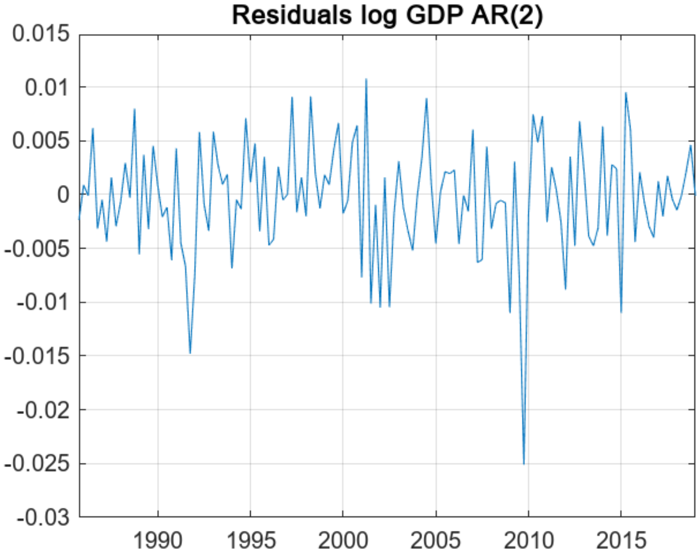

Figure 31: Residuals AR(2)

5 Conclusion

Now, we will summarise all the results in a table in order to compare the models.

Model	**RMSE**
VAR(4)	93.35
VAR(1)	84.99
VAR(p)	85.84
AR(4)-X	84.60
RW	120.06
AR(4)	79.34
AR(2)	77.05

Table 1: RMSE for each model

According to the RMSE, the random walk is the worst model and, since this is often taken as the benchmark to check other models' performance, we can say we have found good models. VAR models perfomed quite good but the AR models performed better with our model, the AR(2), which was the best. Moreover our model, if we exclude the random walk, is the least computationally intensive and, since it has also the lowest RMSE, we can confidentially say it is the best model we have found.