Universidad Pontificia de Comillas
ICAI

Assignment II

Machine Learning I

Authors:
Diego Sanz-Gadea, Sergio Mingot y Agustín Rodríguez Agudo

December 2022

Contents

1 Introduction

This report is part of the second assignment for the Machine Learning I subject. In it, the dataset **UnemploymentSpain.dat** will be evaluated using several forecasting models. The main and final goal is to obtain the best forecast for the unemployment rate of November 2022.

In the figure 1 we can see how the number of unemployed has evolved over the last few years. To a first approximation, we see two major jumps. The first in 2008 when the great crisis occurred, on the other hand, in 2020 caused by the COVID.

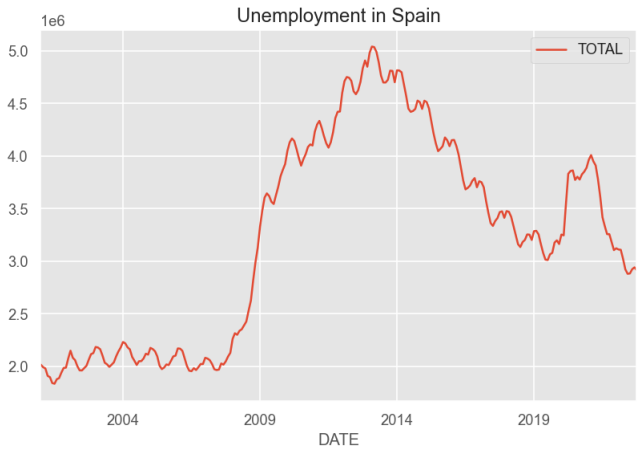


Figure 1: Unemployment data over the last 20 years

2 SARIMA: Seasonal ARIMA

2.1 Identification process

First of all we want to see how the dataset is distributed. To do this, in addition to the time series that we have already seen, we are going to show the ACF (Autocorrelation Function) graph and the PACF (Partial Autocorrelation Function) graph.
The *ACF* presents the correlation coefficient between the series and its previous values. Likewise, the *PACF* measures the partial correlations between the instants y[t] and y[t-k] after having eliminated in both variables the effects of y[t-1],...,y[t-k+1].

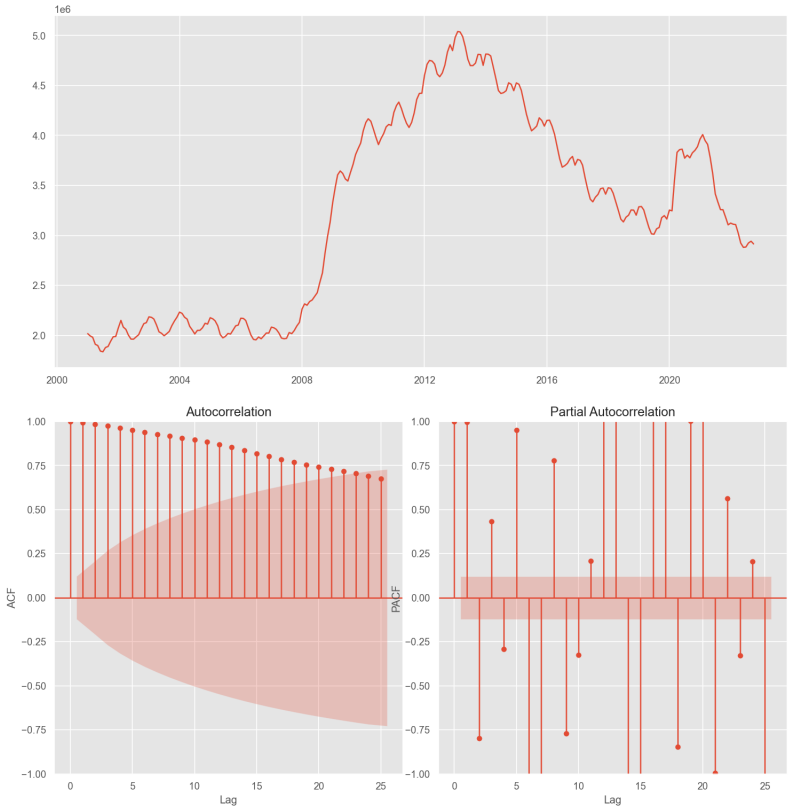


Figure 2: Time series, ACF and PACF

2.2 Stabilize the variance and the mean

In the own graph of the unemployment we can see that our time series is not mean stationary because of the huge jump from 80 to 100 and increasing it. Regarding to the variance we can also see that is not stationary.
Another proof of the non stationary for the mean it is in the ACF graph. Here we view an slowly decreasing curve that implies that non stationary.

Next, it will be shown that the variance is not stationary by means of the graph 3

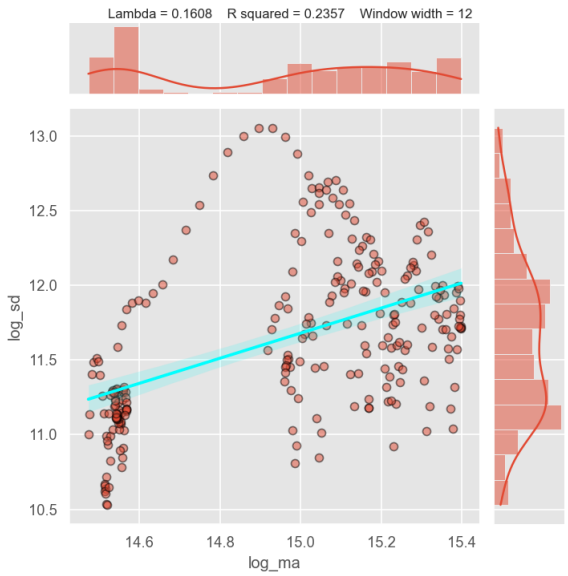


Figure 3: Non stationary variance

In the representation we have several proofs for conclude that the variance is not stationary. The first proof is the λ which is clearly far from 1. The other one is the graph where the straight line shows that as the mean increase, the variance have the same behaviour.

After applying *BOX-COX*, the variance is stationary. In the following graph 4 it is seen that λ is practically one and the straight line is parallel to the horizontal axis.

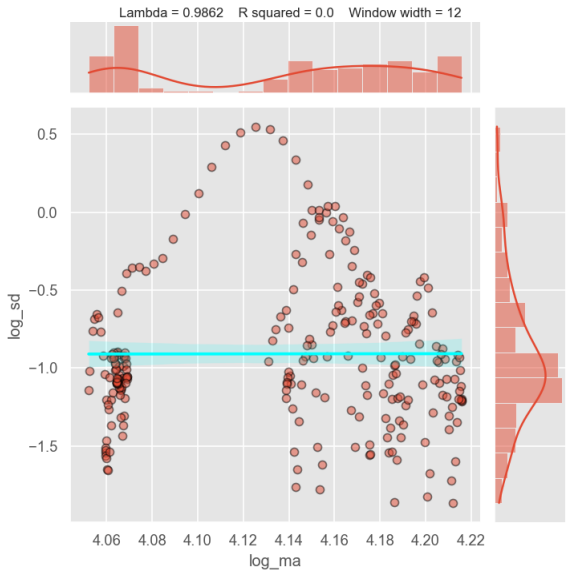


Figure 4: BOX-COX

Through a series of metrics 5, the stationary of the series that had been previously advanced is confirmed. Differentiation is used to get the time series to be stationary in the mean. It will be tested with a first degree differentiation.

```
ADF Statistic: -1.644424
p-value: 0.459950
Critical Values:
        1%: -3.457
        5%: -2.873
        10%: -2.573
The series is not stationary
```

Figure 5: metrics

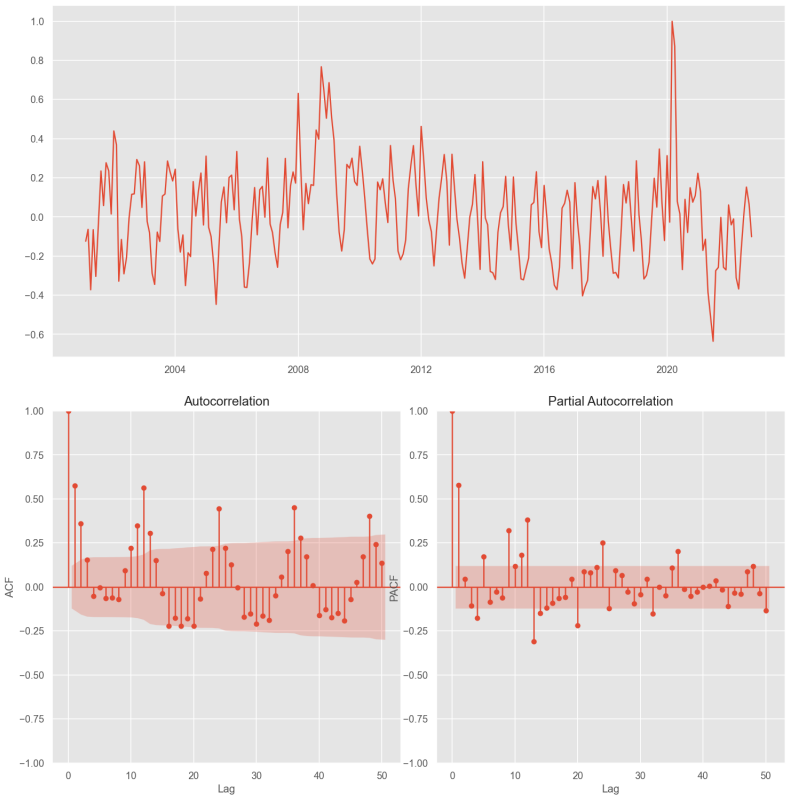


Figure 6: First differentiation

As we see in the graphs 6, the stationary in the mean has been eliminated since there is no slow decrease in the ACF, in turn, in the time series itself it can be seen. Now, another effect can be seen, which is seasonality, as can be clearly seen in the ACF every 12 values, seasonality is repeated. To avoid it, a differentiation will be made again

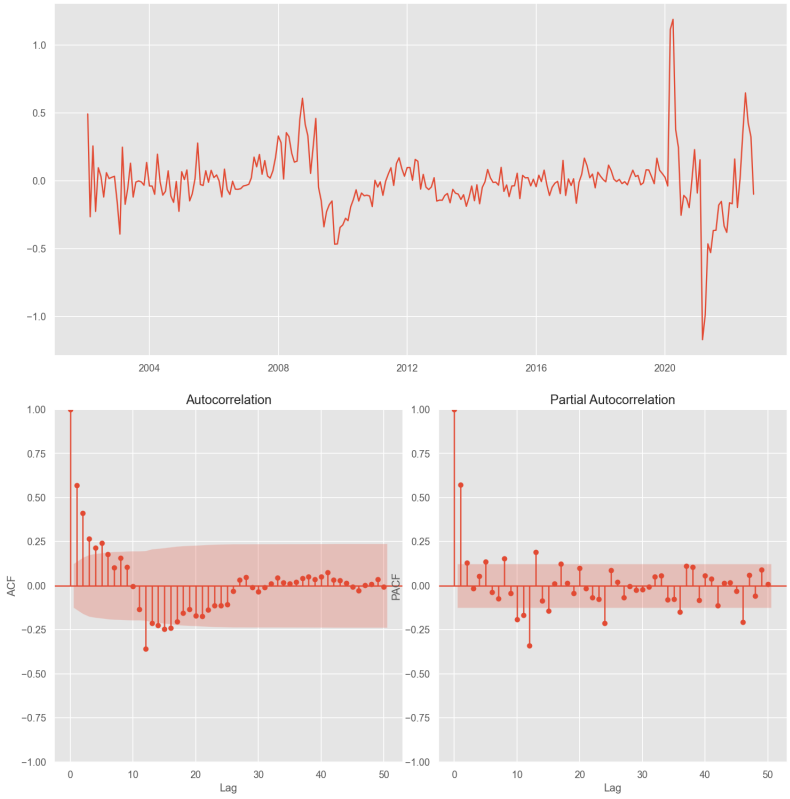


Figure 7: Season differentiation

After the 7 graph, the time series has already become stationary, so we will proceed to identify a model.

2.3 ARIMA Model

First of all, we must obtain which ARIMA parameters are the most optimal to predict our time series. To do this, we have iterated with a series of parameters and the ones with the best results regarding the AIC.

	AIC	order	season	trend
0	-252.317698	(1, 1, 0)	(0, 1, 1, 12)	n
0	-251.843024	(1, 1, 0)	(0, 1, 1, 12)	c
0	-246.092126	(1, 1, 2)	(0, 1, 1, 12)	n

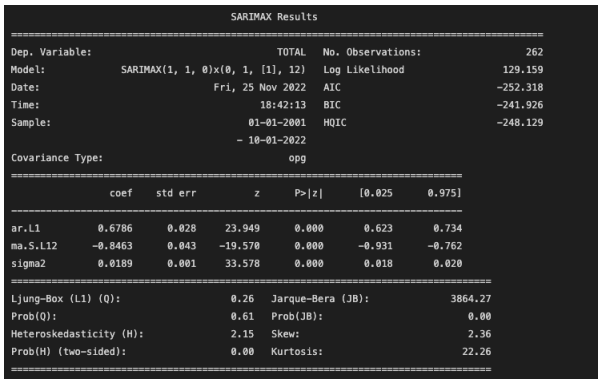
Figure 8: Best parameters

Regarding the image 8, the most optimal is to use an Auto-Regressive of order 1 for the regular part. Regarding the seasonal part, the most optimal is to use a Moving Average of order 1.

This choice could also have been made by observing the graph 7:

- If we look at the regular part in the ACF there are about four significant correlations and in the PACF there is only one, so according to the criterion of taking the least complex, we take the PACF which is an Auto Regressive
- In the seasonal part, that is, every 12, it can be seen that in the ACF there is a single significant correlation, unlike the PACF has three of them. For this reason, a moving average of order one has been chosen for the seasonal part.

With all this we obtain the following summary of the model



SARIMAX Results

Dep. Variable:	TOTAL	No. Observations:	262
Model:	SARIMAX(1, 1, 0)x(0, 1, [1], 12)	Log Likelihood	129.159
Date:	Fri, 25 Nov 2022	AIC	-252.318
Time:	18:42:13	BIC	-241.926
Sample:	01-01-2001	HQIC	-248.129
	- 10-01-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6786	0.028	23.949	0.000	0.623	0.734
ma.S.L12	-0.8463	0.043	-19.570	0.000	-0.931	-0.762
sigma2	0.0189	0.001	33.578	0.000	0.018	0.020

Ljung-Box (L1) (Q):	0.26	Jarque-Bera (JB):	3864.27
Prob(Q):	0.61	Prob(JB):	0.00
Heteroskedasticity (H):	2.15	Skew:	2.36
Prob(H) (two-sided):	0.00	Kurtosis:	22.26

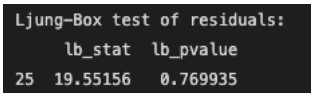
Figure 9: Summary of the model

From this summary, several conclusions can be drawn:

- We can see that the coefficients are adequate since their p-values are practically zero, so they are significant.
- AIC [1] is -252.318 which is clearly very small since you want the more negative the better.
- BIC [2] is -241.926 As in AIC, it is sought to be very small.

2.4 Analyze residuals

Regarding the analysis of residuals, both the Ljung-Box test and the time series and their correlation graphs are used.
The Ljung-Box test measures whether there is auto correlation in a time series. The null hypothesis ($\mathcal{H}_0$) is: the residuals are independently distributed.



Ljung-Box test of residuals:

	lb_stat	lb_pvalue
25	19.55156	0.769935

Figure 10: Ljung-Box test

As can be seen in the figure 10, the p-value is 0.77, clearly higher than 0.05, so it is not rejected $\mathcal{H}_0$ and it can be concluded that the residuals are independently distributed.

[1] determines the relative information value of the model using the maximum likelihood estimate and the number of parameters (independent variables) in the model

[2] is an increasing function of the error variance

In the graph 11 we want to verify that it is white noise.

- The time series can indicate that there is no correlation and a lot of randomness
- The density curve has a Gaussian shape although it is not completely, especially its narrowness and its right tail. Therefore, it can be concluded that the model is improvable.
- Finally, in the ACF and PACF we can say that since there are no correlations we are faced with whitenoise, which is what was intended

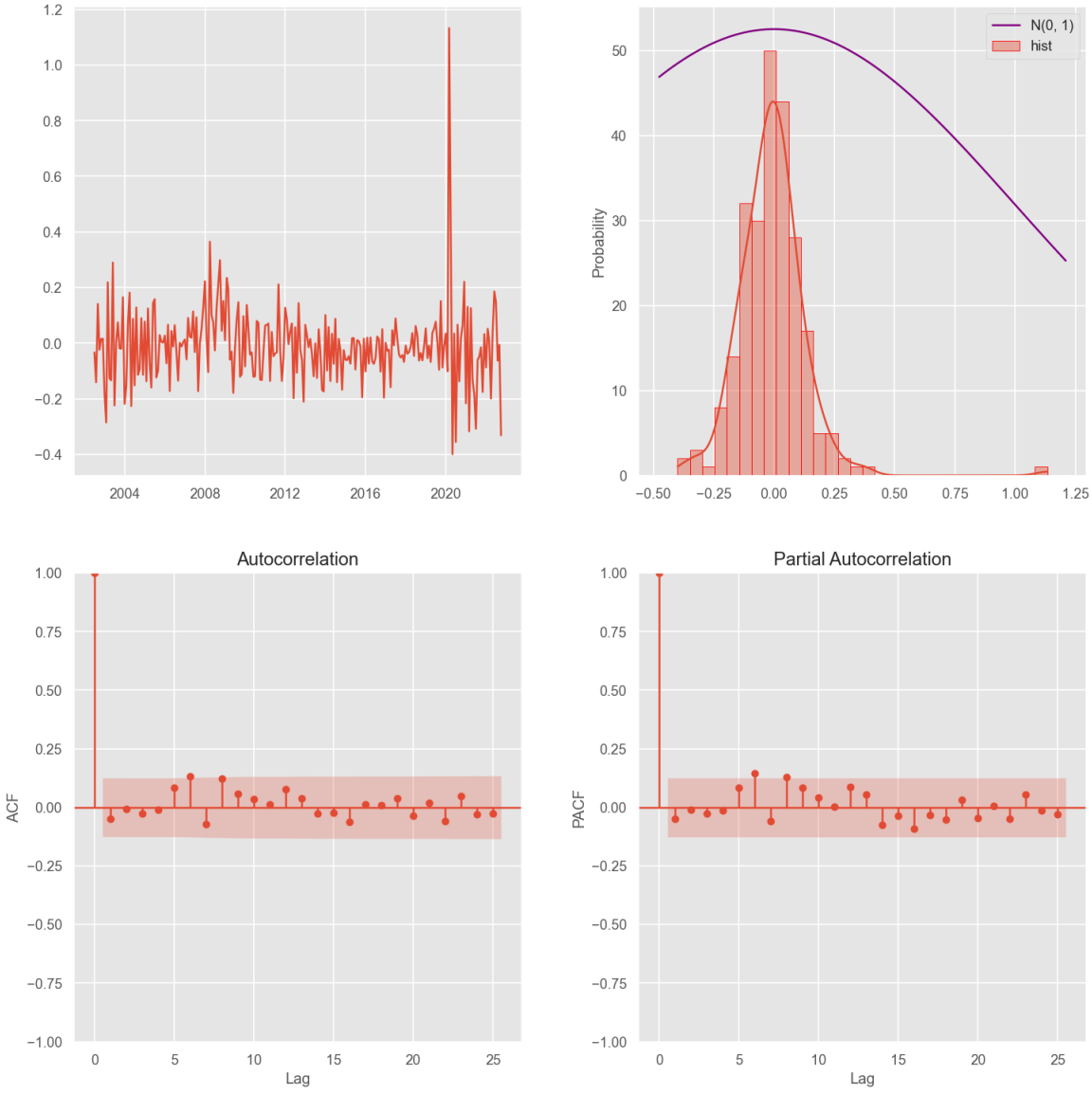
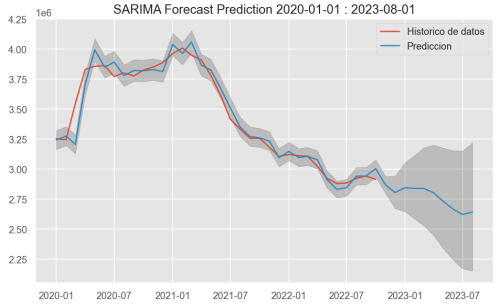


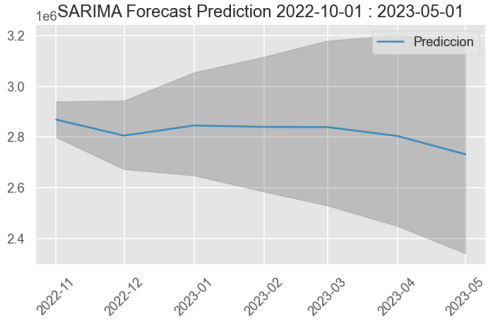
Figure 11: Residulas

2.5 Forecast

In the graph on the left we see the time series of the data and the one generated by our model. On the other hand, on the right you can see the model's prediction for the following months as well as the 95% confidence interval in the shaded area.



(a) Full SARIMA forecast



(b) SARIMA forecast

3 SARIMAX: Seasonal Auto-Regressive Integrated Moving Average with exogenous factors

For training the SARIMAX model we needed to create a new variable for indicate all the period we had to face COVID 19. Pandemic changed the tendency and provoke a decrease in the number of employees. This particular situation could spoil the data for our model, because it breaks all the possible patterns we had.

Nevertheless, all the information we have in the data set is relevant for predicting the future, and that is the reason why we have to consider to create the new variable. This variable will help to explain in which situation we are at the start point, and gives us information about the course the unemployment follows in such a special situation.

Having the SARIMA model trained, the thing we have to do for training the SARIMAX model is introducing Covid variable as an exogenous variable, where covid takes the value 1 if the sample is in Covid period, and 0 in any other situation.

We have selected a model with the same parameters as in the SARIMA, that is we made 1 regular differentiation and use an Auto-Regressive of order 1 method for the regular prediction, and we made 1 differentiation and used a Moving Average of order 1 method for the seasonal prediction, with a seasonal component of 12. For explaining why we have selected this parameters, we go back to the 7 graph.The reason why we used this parameters is the same as in the Sarima model:

- The seasonal component of 12 is because we see a 12 month repetition of periodic patterns(in winter there is usually a lower number of employees than in summer)
- Regarding the ACF and the PACF graphs, we get that in the PACF we have only 1 significant correlation bar, and in the ACF we have multiple significant correlation bars, so the AR(1) is the simplest model for the regular component.
- Regarding the seasonal component, we get that there is only 1 significant correlation bar in the ACF graph whereas in the PACF we have at least 3 significant correlative bars.

For training the model, we set the Covid variable that we have created as an exogenous variable. After training the model, we get the following results:

SARIMAX Results

Dep. Variable:	TOTAL	No. Observations:	262
Model:	SARIMAX(1, 1, 0)x(0, 1, [1], 12)	Log Likelihood	146.402
Date:	Wed, 30 Nov 2022	AIC	-280.804
Time:	00:52:46	BIC	-260.021
Sample:	01-01-2001	HQIC	-272.427
	- 10-01-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
COVID	-0.3905	0.023	-17.217	0.000	-0.435	-0.346
OUTLIERS	-0.3981	0.248	-1.607	0.108	-0.884	0.087
OUTLIERS2	0.2107	0.187	1.128	0.260	-0.156	0.577
ar.L1	0.6910	0.029	23.757	0.000	0.634	0.748
ma.S.L12	-0.8011	0.043	-18.437	0.000	-0.886	-0.716
sigma2	0.0165	0.001	18.255	0.000	0.015	0.018

Ljung-Box (L1) (Q):	0.61	Jarque-Bera (JB):	415.97
Prob(Q):	0.43	Prob(JB):	0.00
Heteroskedasticity (H):	2.10	Skew:	1.08
Prob(H) (two-sided):	0.00	Kurtosis:	9.14

Figure 13: Sarimax results

Regarding the results, we can assert the SARIMAX model we get is a good model because of the low p-values (under 0.05), and we achived a decrease in the AIC value, but we will need to analyze it in next sections because it is not high enough to draw conlusions.

3.1 Analyze residuals

We want to get again that the graph of the residuals is white noise:

- Again we an assert that there is not correlation regarding the time series
- The density curve is slightly right skewed and is not as Gaussian as Sarima's density curve, but is is not so far from having Gaussian shape.
- Regarding PACF and ACF graphs we can observe that there is not a high correletion, which means that we have white noise.

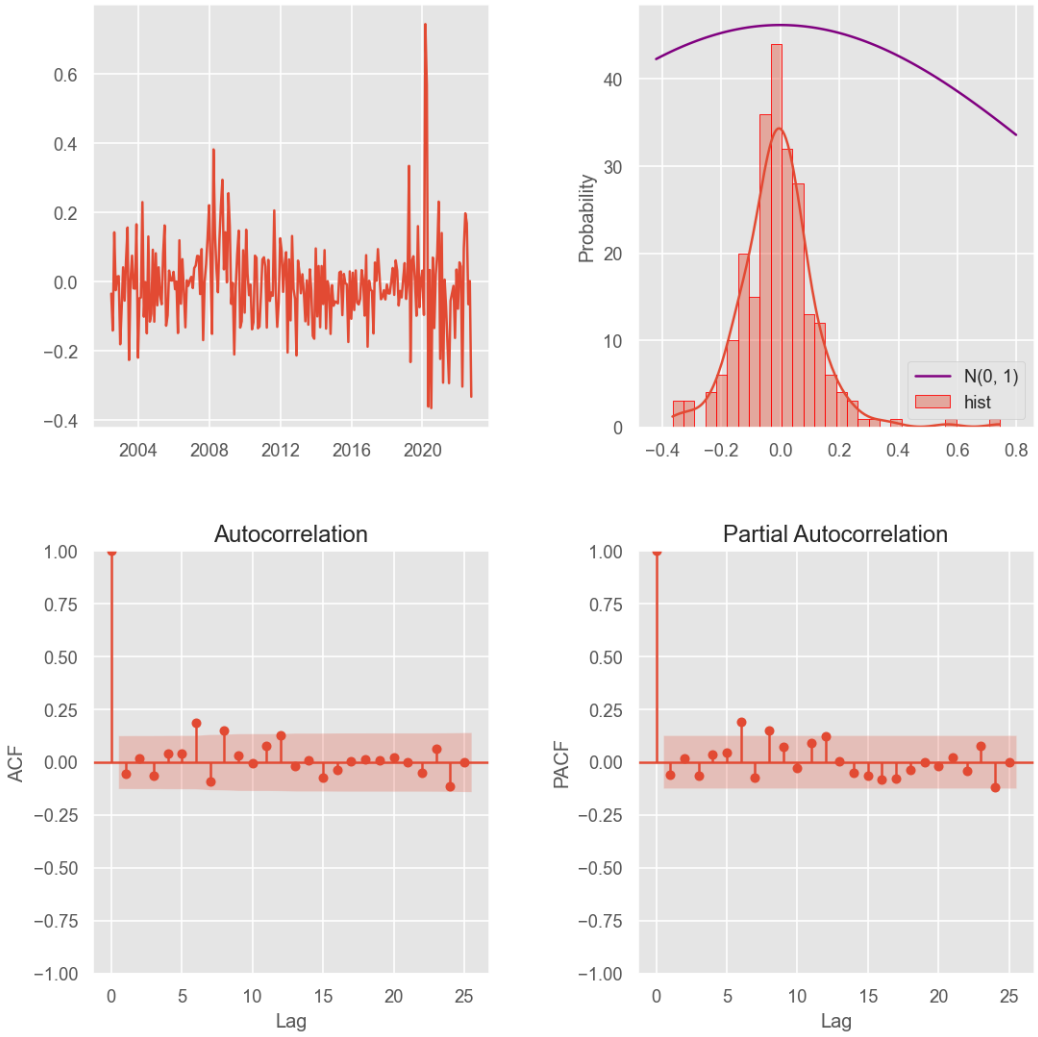
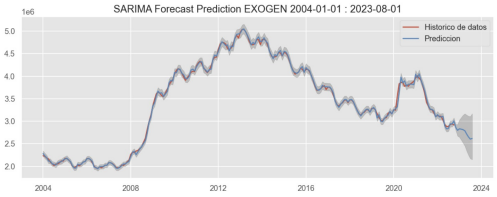


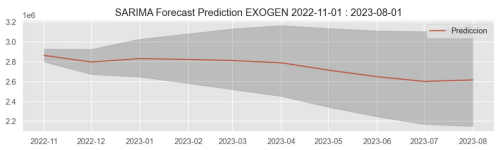
Figure 14: Sarimax residuals

3.2 Forecast

After realizing that this Sarimax selection is a good model for predicting future values, we used it for predicting unemployment, getting the following results:



(a) Full SARIMAX forecast



(b) SARIMAX forecast

In the first picture we can see the entire time series data whereas in the second one we

particularize in just the prediction for the following months and the 95% confidence interval. By the prediction, it seems that the unemployment may stabilize in the following months, which is a sign that indicates that our model interpreted Covid period as the special situation it is.

4 Non-Linear Model

In this section, a non-linear model will be used to make the prediction of the time series. The chosen non-linear model is Multi Layer Perceptron (MLP).

For the prediction using this model it is necessary to create a series of variables that function as *INPUT*. These variables will be made through 'shifts' in a range from one to fourteen.
First, we want to determine the distribution of the training and test set, so we will test with values between 0.7 and 0.99. At the same time, you will want to determine the most optimal hyperparameters, for this you will try the following:

- *hidden layers* = [(4,2),(2,2),(3,3),(6,3),(15,),(9,),(8,),(6),(5,),(3,)]
- α =[0.0001,0.001,0.01,0.1]

To get the best selection of both lags and hyperparameters, the performance is measured by the RMSE scored of the model and the mean and standard deviation of the model for a given split of the relative percent difference between TRAIN and TEST.

$$\frac{(RMSETRAIN - RMSETEST) \cdot 100}{RMSETRAIN}$$

The following image 16 shows the result of the comparisons that have been made between the different possibilities that have been presented before. On the left you can see that the train grows slowly, unlike the test that drops off suddenly. This is caused by COVID since for those sections this event is already taken into account when training. Once we are in these sections, with the image on the right we select that the distribution will be 0.93 since it is the one with a nearby zero mean and a narrow standard deviation of the relative difference bewteen RMSE train and test.

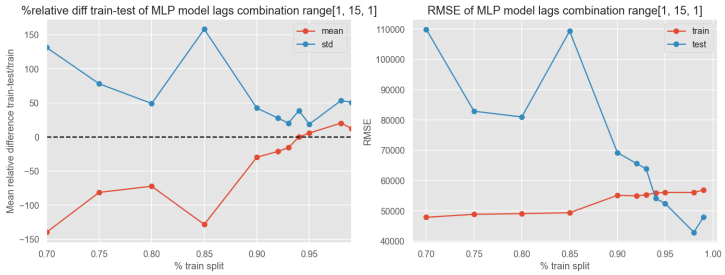


Figure 16: Split

4.1 Model

Once it has been decided to take 0.93 as the distribution of the train/test, we want to know which are the most optimal hyperparameters. The following image shows a summary of the best, ordered by Test RMSE.

MODEL/LAGS	RMSE_TRAIN	RMSE_TEST	R2_TRAIN	R2_TEST	BEST_PARAMS	LEN_DATA	SPLIT train/test	%rmse test-train
MLP [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]	44735.488074	47684.971715	0.997975	0.915718	{'MLP__alpha': 0.01, 'MLP__hidden_layer_sizes'...	248	0.93	-6.593163
MLP [1, 2]	58326.736383	54113.608222	0.996705	0.928231	{'MLP__alpha': 0.0001, 'MLP__hidden_layer_size...	260	0.93	7.223322
MLP [1, 2, 3, 4]	58105.080106	54998.053271	0.996711	0.925866	{'MLP__alpha': 0.1, 'MLP__hidden_layer_sizes':...	258	0.93	5.347255
MLP [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]	54959.245355	55443.745308	0.996976	0.92466	{'MLP__alpha': 0.01, 'MLP__hidden_layer_sizes'...	251	0.93	-0.881562
MLP [1, 2, 3, 4, 5, 6, 7, 8]	57007.79539	56844.442734	0.996782	0.920805	{'MLP__alpha': 0.01, 'MLP__hidden_layer_sizes'...	254	0.93	0.286544
MLP [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	51531.074382	58243.819904	0.997333	0.916858	{'MLP__alpha': 0.1, 'MLP__hidden_layer_sizes':...	250	0.93	-13.026597
MLP [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]	47028.773594	58548.40053	0.997766	0.872943	{'MLP__alpha': 0.0001, 'MLP__hidden_layer_size...	249	0.93	-24.494849
MLP [1, 2, 3, 4, 5]	57572.549751	58650.967013	0.99676	0.915691	{'MLP__alpha': 0.1, 'MLP__hidden_layer_sizes':...	257	0.93	-1.873145
MLP [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	54527.799713	60625.261499	0.997033	0.90992	{'MLP__alpha': 0.0001, 'MLP__hidden_layer_size...	252	0.93	-11.182299

Figure 17: Hiperparameters selection

After this it is seen that the best model is:

- hidden layers: (3,)
- α: 0.01
- solver: lbfgs
- activation function: relu

Although in the image 17 you can see that the best results are given by using fourteen lags, really most of them are useless. So the ones that give the most information are taken, which are: 1,2,12,13 and 14, that is, we do shifts of those rows.

The following figure 18 shows through the sensitivity plots the importance for the model of these variables since all of them are far from (0,0), that is, they have above all a mean and a variance far from 0.

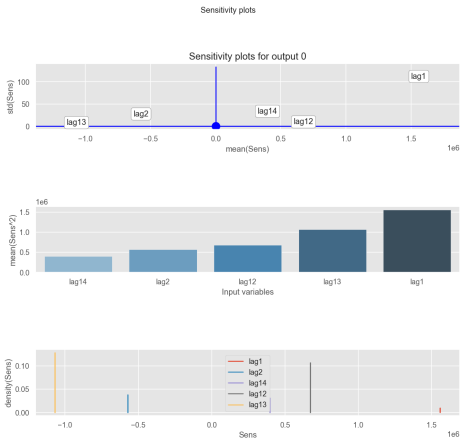
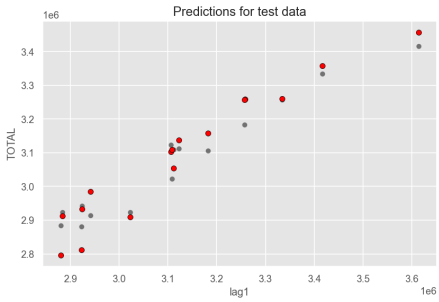


Figure 18: Sensitivity plots

Below is the behavior of our model where the predictions of *lag1* versus *Total* have been plotted. In red the prediction and in black the actual data. It is seen that in the image 19a the prediction fits almost perfectly. On the other hand, in 19b there is a greater difference between the prediction and the real.

(a) Predictions for training



(b) Predictions for test

To measure how good our model is, we are going to analyze a series of metrics and their residuals. Regarding the metrics, it has been decided to take: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R^2.
In the image 5 it can be seen that both the MAE and RMSE are quite high both in train and test, it was expected to have these high values. If we look at the R^2, this gives us very significant values since in the test it is 0.91, which is practically 1, this being the best value.

```
Training MAE: 28851.97898462178
Test MAE: 40268.65531943354
Training RMSE: 45621.918283201594
Test RMSE: 49798.79559204936
Training R2: 0.9978938980049025
Test R2: 0.9080806312769827
```

Figure 20: metrics

Regarding residuals, they can be seen in the image 22. The expected would be a homogeneous band around the x axis. In our graph, you can see this band for all cases. It should be noted that around half, there are clear scattered values. Finally, in the histogram it can be seen that it follows a normal distribution.

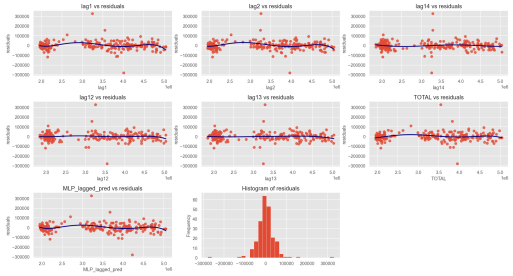


Figure 21: Residuals

Finally, the graph **??**is shown where the curve of the data and the prediction of the model can be seen. As can be seen, the prediction fits the data very well.

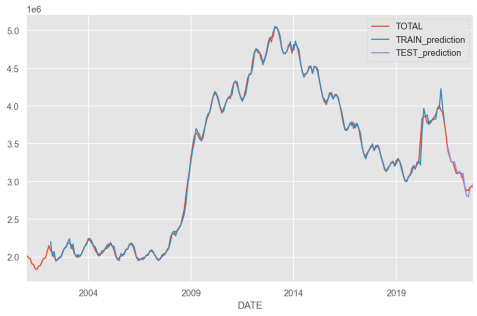


Figure 22: Residuals

5 Prophet

We will us prophet in attempt to get a better model for predicting unemployement. We think in this case prophet will be a good model because it splits the series in holidays, trend and periodic changes. This is the reason why we think it will understand well covid period, which is the thing that can spoil our predictions in other model, because of its strangeness.
So the first thing we have done for getting predictions with prophet is defining a Covid Period where we had not have expected values. We get the following graph of the results for the following year:

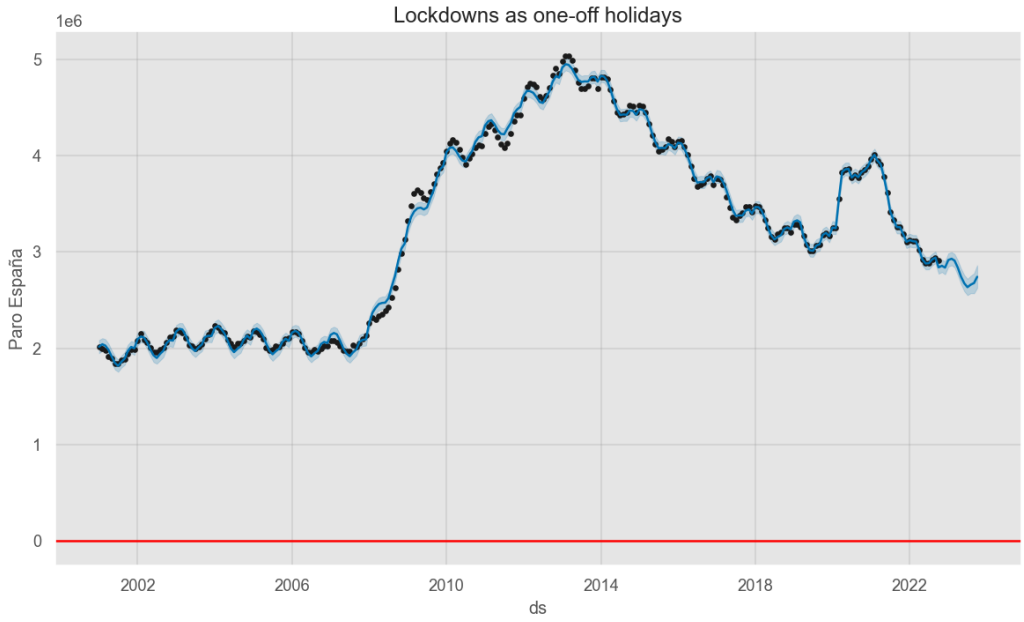


Figure 23: Prophet

As we can observe, there is a fast risen next to the covid period in 2020, so we have to make the model aware of this special change as we said before. Then, separating in 3 graphs: the trend, the covid period and the periodic changes, we get to the following graphs:

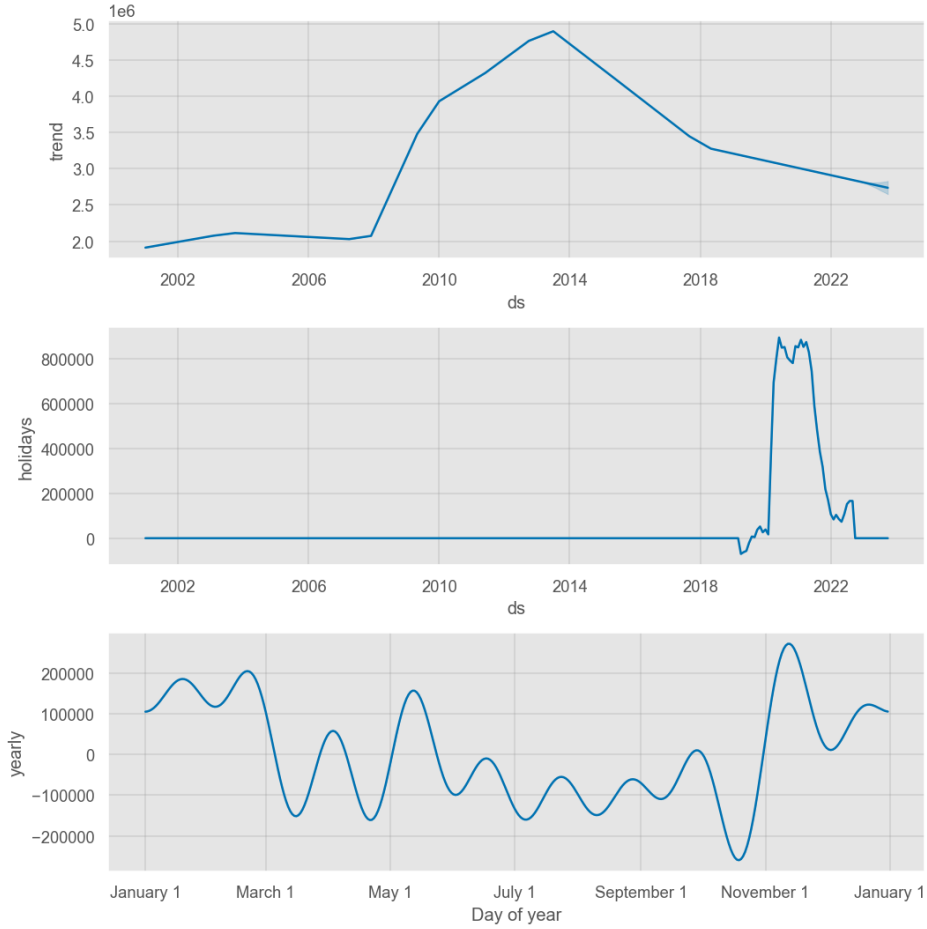


Figure 24: Componentes

As we see, the covid period made the number of unemployment rise against the trend, which means that in a standard situation it may decrease. The model will value this and decide which is the best estimation. We end up concluding that the prediction may follow the trend and will decrease unemployment.

6 Comparing Models

After getting all the models, the next step we want to take is comparing the results off all models for getting conlusions abut which model will work better for predicting Unemployement for the next winter.

We start comparing which model is better between SARIMA and SARIMAX, and we get to the conclusion that SARIMAX has a better performance when predicting model.

The fact that we can introduce the covid period as an exogenuous variable makes the model predicts better. Both models have low enough p-values for being the best SARIMA and SARIMAX respectively, and the principal reason we would predict with SARIMAX instead of SARIMA is the lower AIC value; Sarimax has less than -280 AIC whereas Sarima has over -252.

Then,we have to select which model we would use for predicting between Sarimax, Prophet and MLP. For selecting it, we will observe which one has better RMSE error:

	RMSE	MSE	MAE	R2	AIC	lower_yhat2022-11	yhat2022-11	upper_yhat2022-11
ProphetCovid	50082.595299	2508266351.851996	35003.904569	0.997415	NaN	2790283.750714	2855990.155672	2924011.09251
sarimaEXOGEN	39406.889886	1552902970.484412	27079.467043	0.998262	-280.804424	2798051.039214	2863394.214068	2930013.21244
sarima_fit	42032.255735	1766710522.140876	28101.626288	0.998023	-252.377129	2798511.848217	2868528.526995	2940009.425571
MLP_lag[1:14]	TRAIN: 45257.1 TEST: 49880.3	TRAIN: 2048203781.8 TEST: 2488047158.5	TRAIN: 29086.1 TEST: 40725.4	TRAIN: 0.998 TEST: 0.908	NaN	NaN	2874236.630436	NaN

Figure 25: Compare

The MLP was expected to perform worse than other models because we did not indicate that pandemic period that seriously affected the data set, and as we see, it has higher RMSE than SARIMAX. Before training it, we could have expected that prophet was going to be the best model, because it usually performs well with special situations as pandemics. . But in this case, we have that all models have a similar R2 error, but the one which has lower RMSE is Sarimax, and the one which has higher RMSE is prophet. For this reason, if we had to predict the next month unemployment we would go for predicting with a Sarimax model.

6.1 Selecting Models

Although we have assert that Prophet would be best option in normal conditions, we are going to forecast with the MLP model. The reason why we choose MLP is because we want to get the highest values of unemployment, and we think that the Ukraine War and the current crisis affects negativelye to our data.

To sum up, our unemployment forecast for November 2022 is 2874237