# Lab Assignment 2: Unemployment Forecasting

Pablo Rodríguez, Ángel Visedo and José Carlos Riego

Machine Learning I - MSc in Big Data

November 2024

**Abstract**

This study explores the application of statistical and Machine Learning methods to forecast Spain's unemployment rate for November 2024. By utilizing a dataset comprising monthly unemployment rates from 2001 onward, the research evaluates the performance of three models: SARIMA, SARIMAX, and a Multi-Layer Perceptron (MLP) neural network. Additionally, the analysis incorporates the potential influence of the recent DANA event in Valencia on unemployment trends. Among the models tested, SARIMAX emerged as the most effective and simplest, demonstrating a superior capacity to capture intricate patterns in the data.

## 1 Time series analysis

If we observe the total unemployment time series for Spain (Figure 1), we can identify several key patterns:

1. From 2008 to 2014, there was a significant increase in unemployment, due to the financial crisis.

2. In 2020, we observe a clear rise in unemployment, this time caused by the COVID-19 pandemic. This effect did not dissipate until 2022.

On the other hand, in Figure 2, which shows the seasonal component of the series (having a period of 12 months), we can observe the following:

1. In the summer months, we can clearly see an increase in unemployment, due to the appearance of temporary work contracts, mostly related to tourism jobs, which significantly increase during the summer.

2. Secondly, in the month of December, we also observe a slight decrease, due to seasonal hiring in response to increased demand during the holiday season, especially in sectors such as hospitality and courier services.

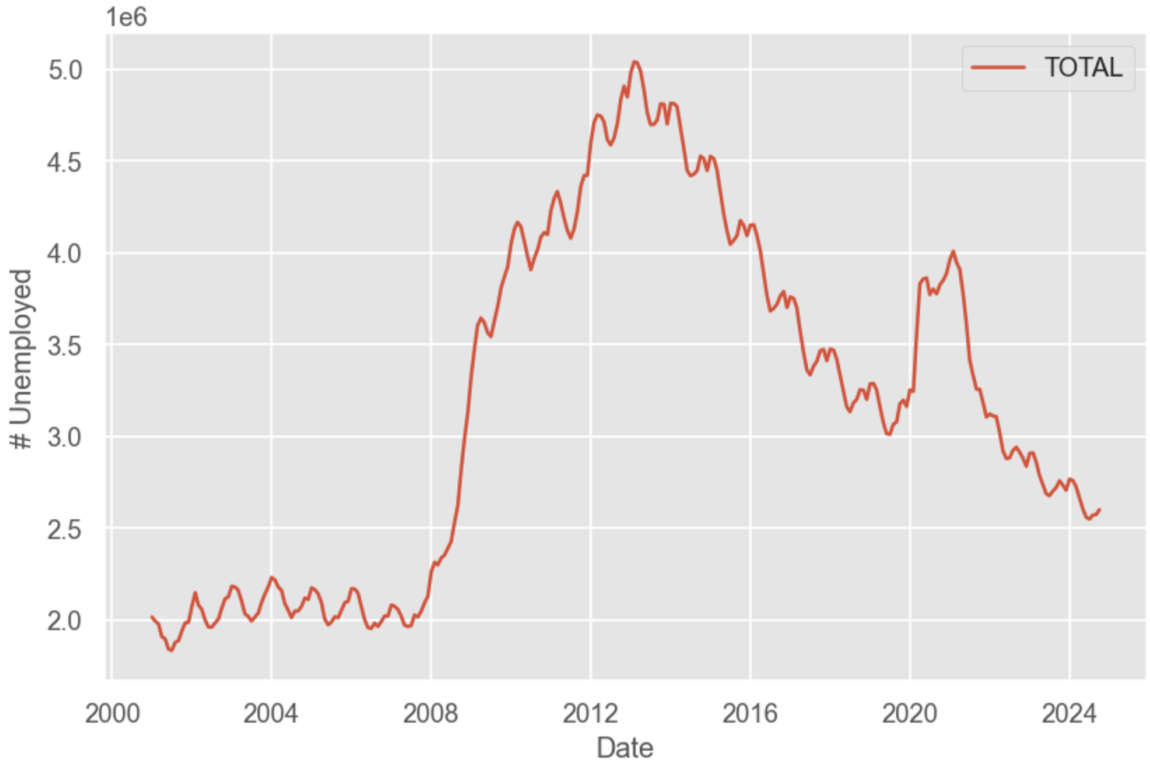


Figure 1: Unemployment data in Spain from January 2001 to October 2024

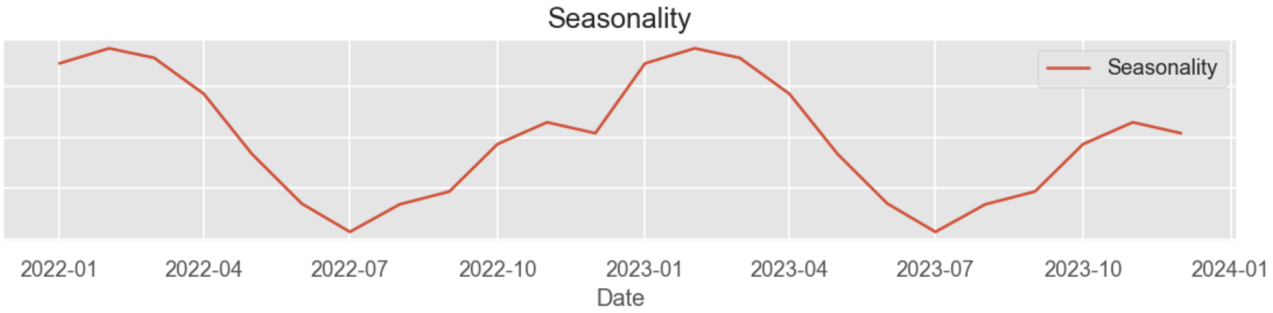


Figure 2: Seasonal trend in unemployment

Additionally, throughout this analysis, we scaled the unemployment data by a factor of $10^{-6}$, as the values are reported in millions. This adjustment was necessary to ensure that the scale of the data did not adversely affect the performance of the models. A slight difference in results was observed after applying this transformation.

## 2 Forecast with SARIMA model

Before determining the AR (p) and MA (q) orders of the SARIMA model, we aimed to visualize the evolution of the standard deviation versus the mean of the series using 12-month windows (Figure 3). What we observed is that, although there is a slight positive correlation between these two variables, this relationship is primarily driven by the onset periods of the 2008 financial crisis and the COVID-19 crisis in 2020. During these times, there was a temporary increase in the variance of the series relative to its mean.

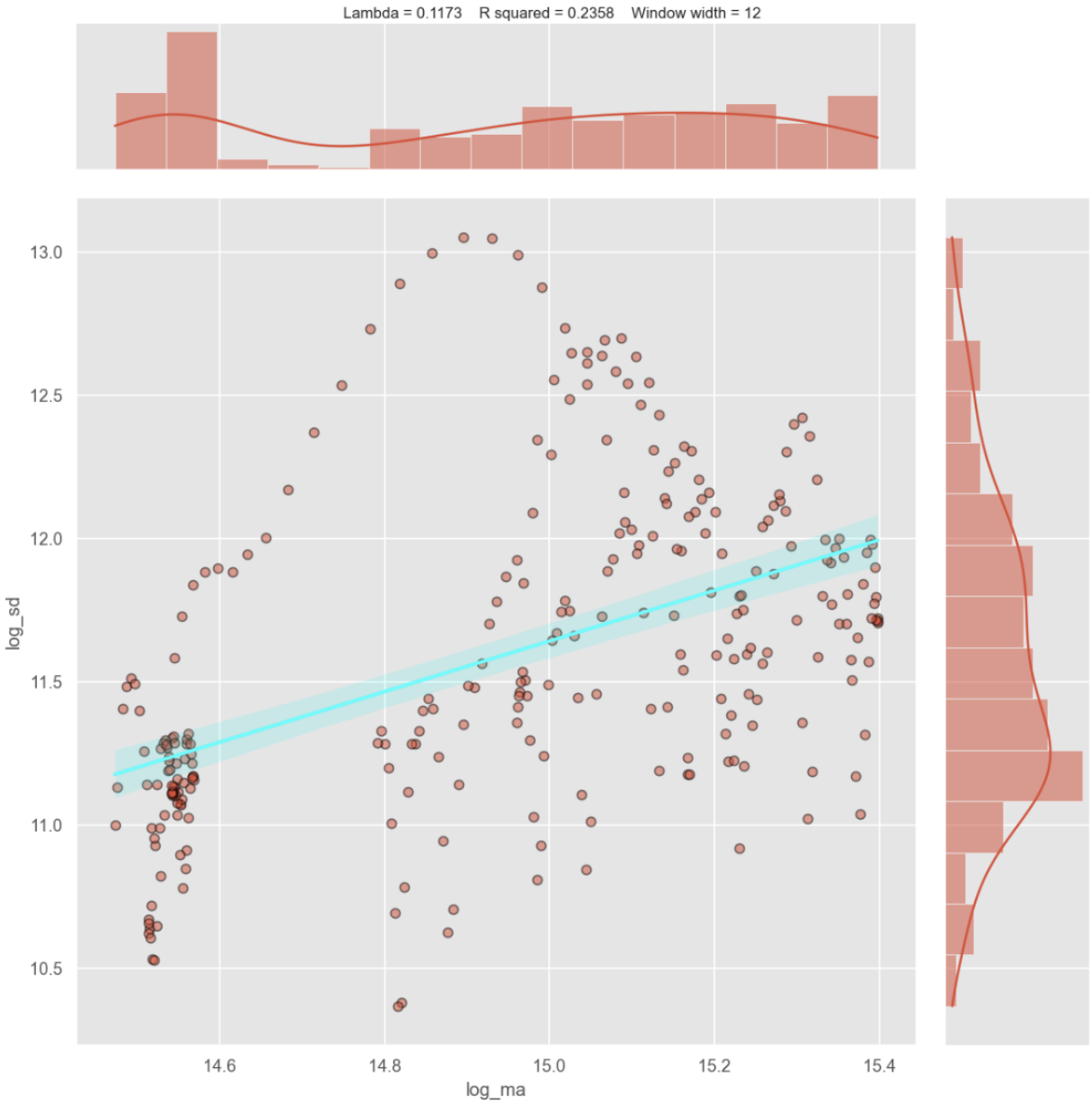


Figure 3: Relationship between log(variance) and log(mean)

Additionally, since the $R^2$ value is less than 0.4, we can rule out applying a Box-Cox transformation to the data. This is because the series can generally be considered stationary in variance, and applying Box-Cox in this case would unnecessarily alter the selection of ARMA orders.

Subsequently, given the slow decay in the coefficients of the ACF shown in the upper part of fig. 4, we differenced one time in the regular component of the series. After confirming the seasonal period of 12 months ($S = 12$), we applied an additional differencing to the seasonal component, finally achieving stationarity in the mean.

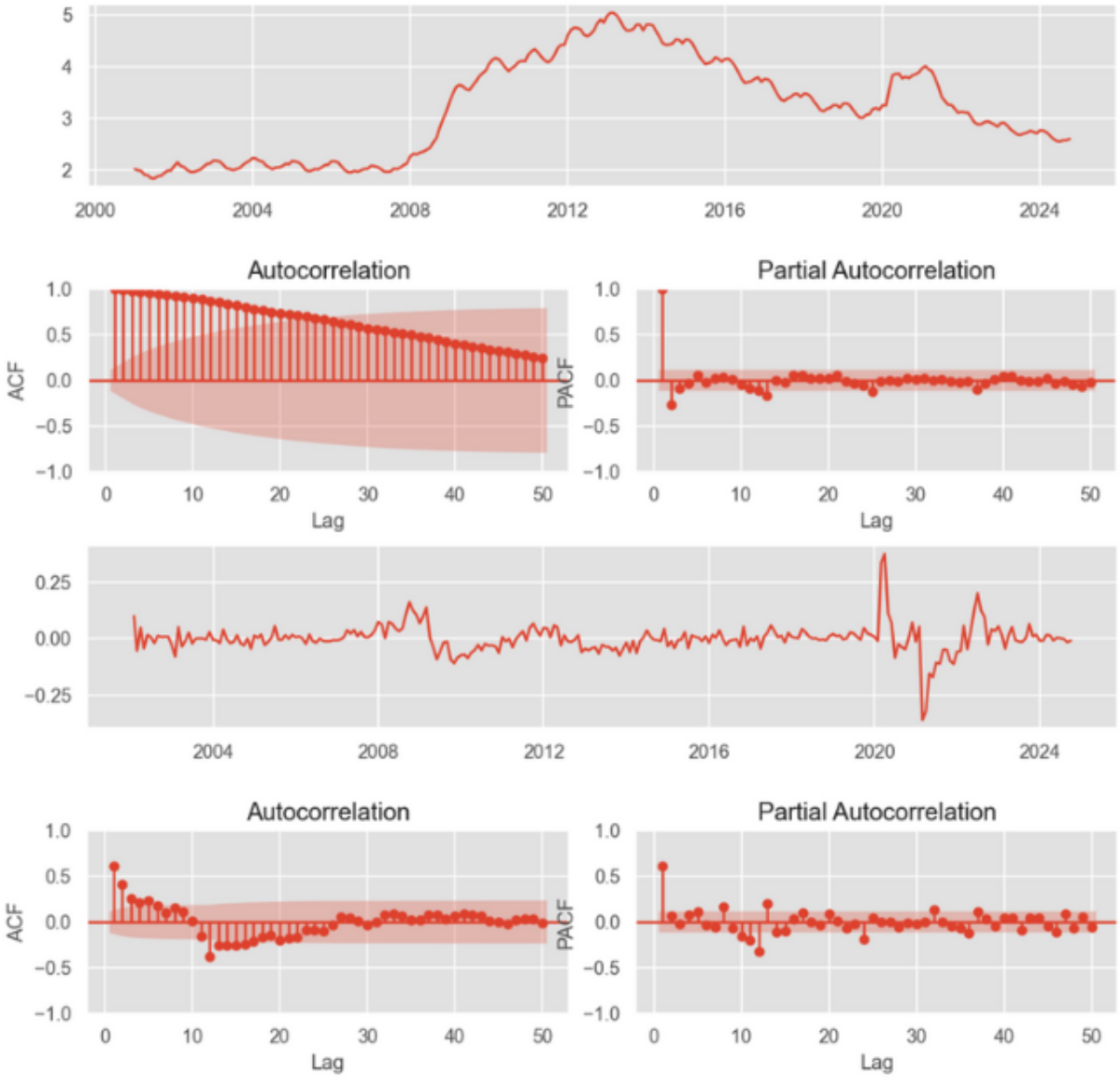


Figure 4: Time series, ACF and PACF before and after differentiation (d = D = 1)

After that, when analyzing the ACF and PACF coefficients of the differenced series, we observe several points:

1. As we can see, there is a highly significant coefficient in the PACF and an exponential decay in the first period of the ACF. Thus, we will include at least one $p$ order for the AR in the regular component of the series.

2. Additionally, we observe a significant coefficient at lag 12 in the ACF and an exponential decay at lags 12, 24, and 36 in the PACF. Therefore, we will include at least one $Q$ order for the MA in the seasonal component of the series.

Finally, after several iterations, we settled on the orders specified in the following table, where the term `n` indicates that no additional trend has been included in the series, as it is already centered in 0:

| **Regular** | | | **Seasonal** | | | | |
|---|---|---|---|---|---|---|---|
| **p** | **d** | **q** | **P** | **D** | **Q** | **S** | **Trend** |
| 1 | 1 | 0 | 0 | 1 | 1 | 12 | n |

Table 1: Orders of SARIMA model and trend

The chosen model has been diagnosed according to the Box-Jenkins methodology, yielding the following results:

1. The included ARMA coefficients are all significant ($p\_value = 0$), thus rejecting the null hypothesis.

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | 0.6471 | 0.023 | 28.509 | 0.000 | 0.603 | 0.692 |
| ma.S.L12 | −0.7866 | 0.033 | −23.812 | 0.000 | −0.851 | −0.722 |
| sigma2 | 0.0017 | 4.71e−05 | 35.448 | 0.000 | 0.002 | 0.002 |

Figure 5: Significance of SARIMA's model coefficients

2. According to the histogram in Figure 6, the residuals are normally distributed with a mean of 0, although some outliers are observed, which are also visible in the residual series itself, corresponding to the COVID period.

3. Furthermore, it is worth mentioning that we decided to plot and analyze the residuals starting from 2006 to gain a broader perspective on the model's performance. Through this analysis, we can confirm the earlier observation: the model performs well, with residuals centered around zero, except during the COVID period, where notable outliers are observed.

4. According to Figure 6, since there are no significant coefficients in the ACF or PACF, and the *p_value* of the Ljung-Box test is $0.67 > 0.20$, we can conclude that the model's residuals are white noise.

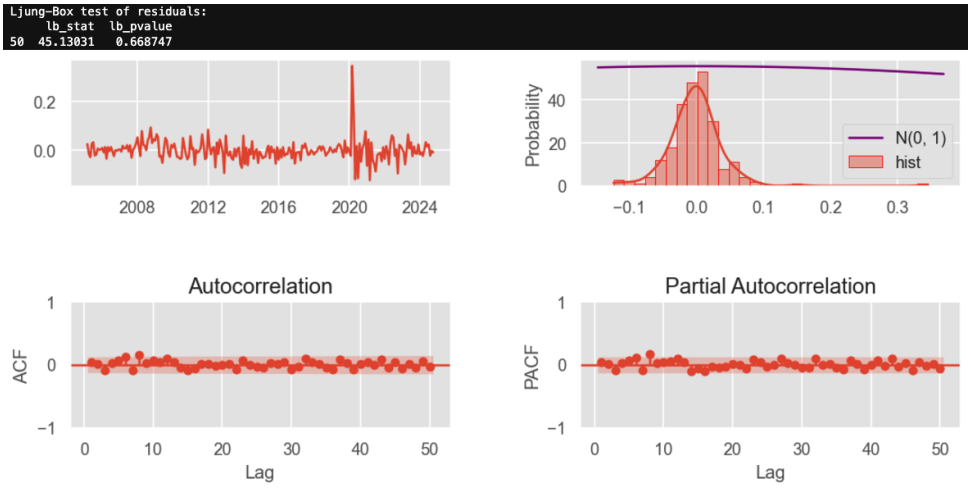


Figure 6: ACF and PACF of the residuals of the SARIMA model

Finally, we obtained the prediction for unemployment over the entire period for which we have data (Figure 7), observing that, except for the COVID period, the SARIMA model's predictions closely approximate the available data in a very acceptable manner.

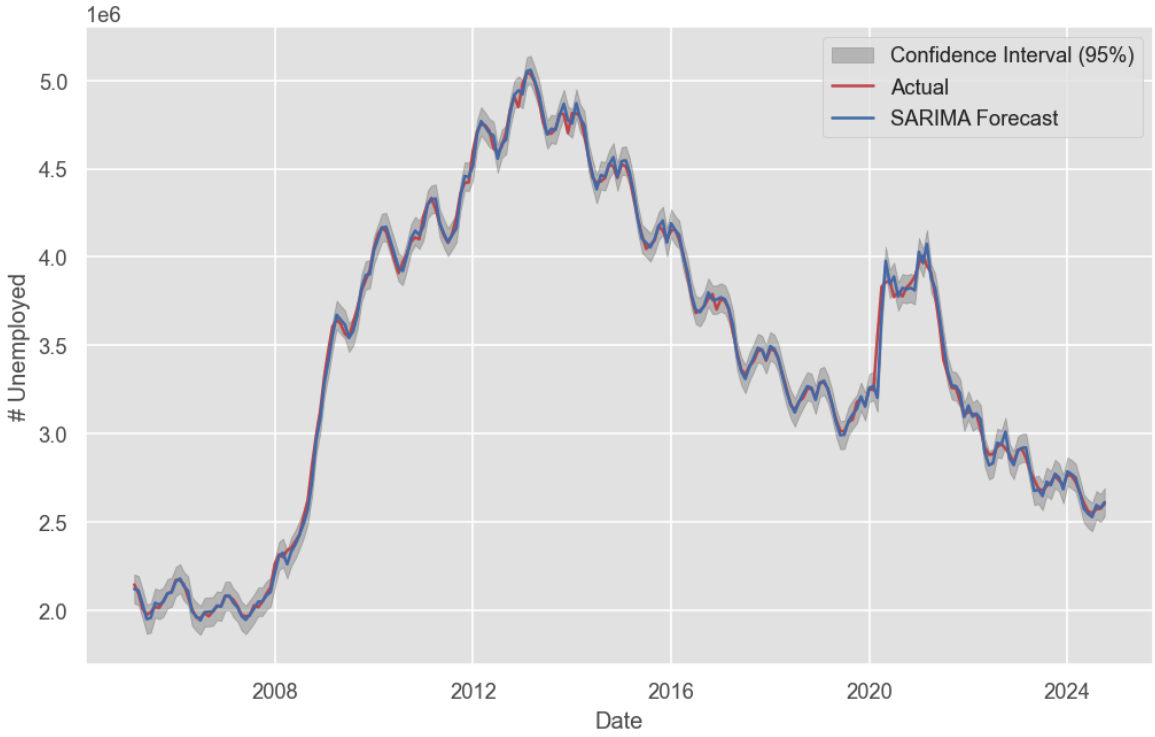


Figure 7: SARIMA forecast for unemployment in Spain from January 2001 to November 2024

# 3 Forecast with SARIMAX model

In order to account for the influence of the COVID crisis on the unemployment data between 2020 and 2022, we included an intervention variable, which, after testing with a step function and a Gaussian function centered in 2021, we chose as a chi-squared distribution of three degrees of freedom, as it minimizes some metrics like MAPE (*Mean Absolute Percentaje Error* and AIC (*Akaike information criterion*)).

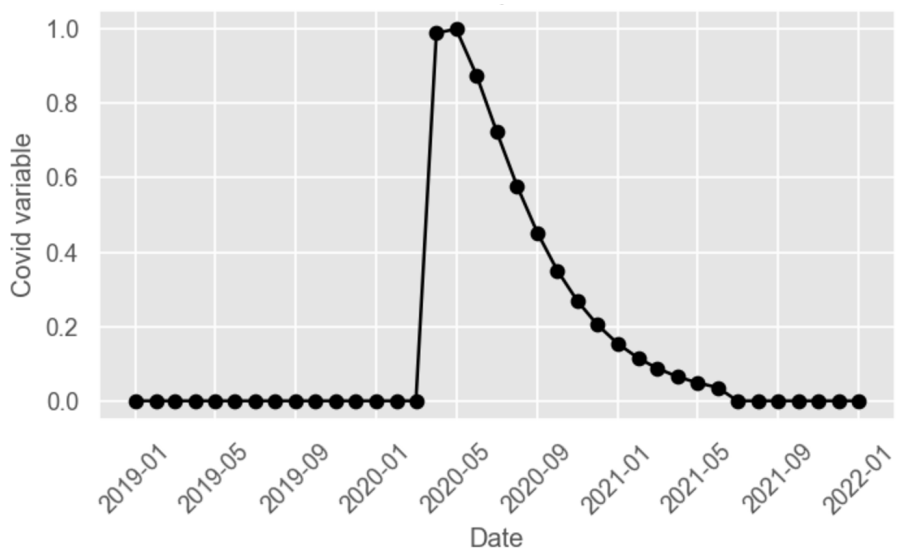


Figure 8: Intervention variable for COVID, modeled as a chi-square distribution

After including this exogenous variable in the model, the obtained coefficients remain the same as those of the previous SARIMA model (table 1), once again being significant, as shown in the figure just below.

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| COVID | 0.1574 | 0.064 | 2.475 | 0.013 | 0.033 | 0.282 |
| ar.L1 | 0.6273 | 0.048 | 12.967 | 0.000 | 0.532 | 0.722 |
| ma.S.L12 | −0.7666 | 0.031 | −24.467 | 0.000 | −0.828 | −0.705 |
| sigma2 | 0.0015 | 4.1e−05 | 37.530 | 0.000 | 0.001 | 0.002 |

Figure 9: Significance of SARIMAX's model coefficients

Furthermore, by examining the ACF, PACF, and histogram (Figure 10), we observe that the residuals are once again white noise, with a $p\_value = 0.57 > 0.20$. These results are very similar to those obtained in the previous section, Figure 6, leading to the same conclusion.

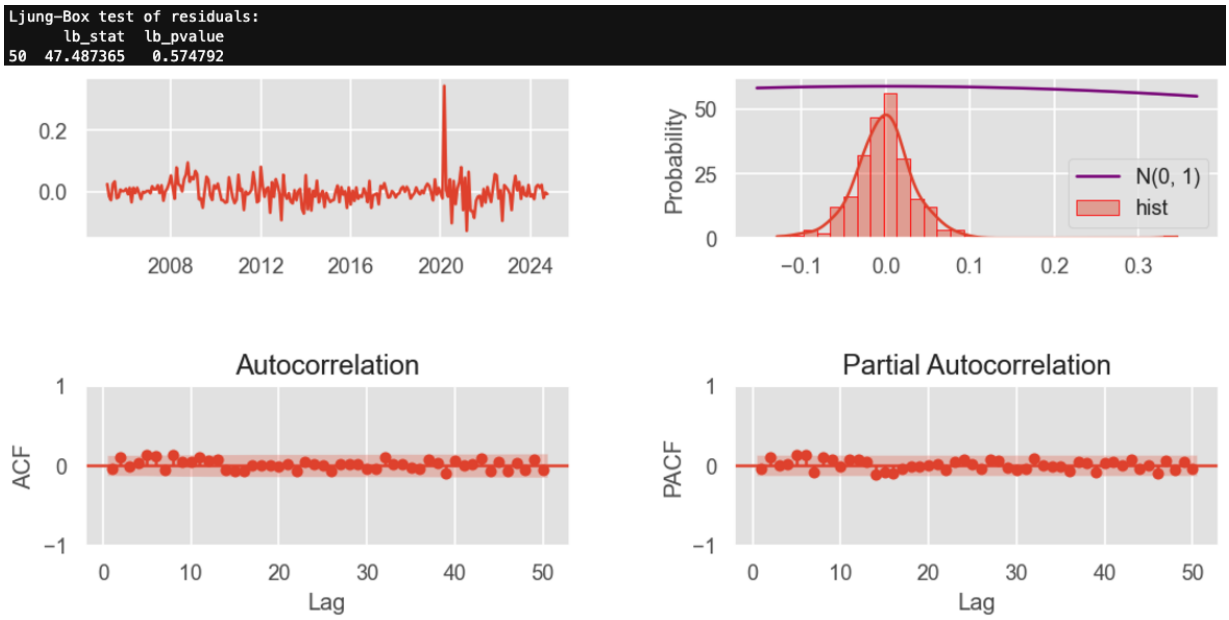


Figure 10: ACF and PACF of the residuals of the SARIMAX model

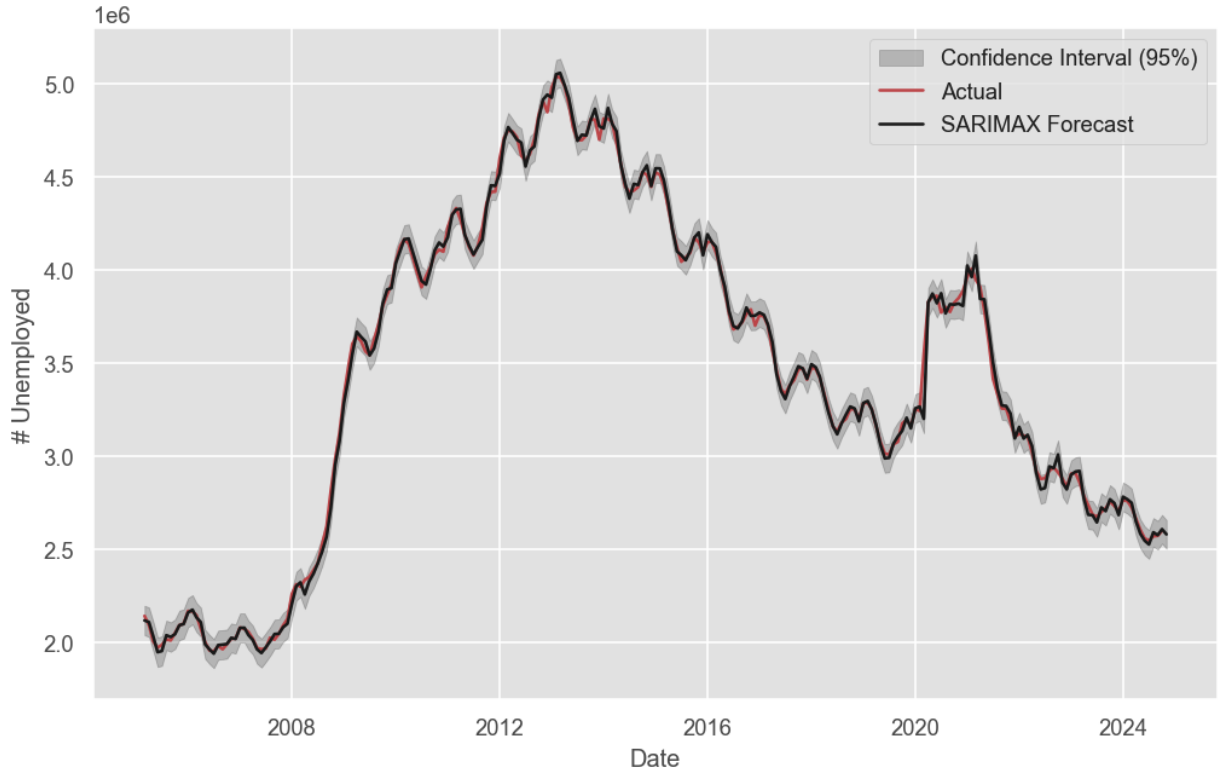


Figure 11: SARIMAX forecast for unemployment in Spain from January 2001 to November 2024

As we can see in the previous figure, the prediction appears to be the same between the SARIMA and SARIMAX models. However, if we focus on the first half of 2020 in Figure 12, the


Figure 12: SARIMA vs SARIMAX's forecast between 2019 and 2024

prediction of the SARIMAX (black) fits slightly better to the actual series (red) compared to the SARIMA (blue). Therefore, we can conclude that the intervention variable has contributed to improving the prediction during the COVID period.

# 4 Forecast with NARX model (MLP)

To forecast unemployment, we have decided to train a Multi-Layer Perceptron (MLP) model using the NARX (Nonlinear AutoRegressive model with exogenous inputs) approach. For this purpose, we utilized the previously mentioned COVID variable as an exogenous input, along with several lags of the unemployment series: `"TOTAL_lag1"`, `"TOTAL_lag2"`, `"TOTAL_lag12"`, and `"TOTAL_lag24"`, corresponding to one month, two months, one year, and two years, respectively.

The model was trained using the following hyperparameter grid (with bold indicating the parameters selected through Cross Validation, which minimize the *MAPE*, chosen as the primary metric for model comparison due to its interpretability as a percentage error):

- **Regularization values (`alpha`):**

$$[\text{1e-7}, \text{1e-6}, \text{1e-5}, \text{1e-4}, \text{1e-3}, \text{1e-2}, \mathbf{\text{1e-1}}, 1.0, 10.0]$$

- **Hidden layer sizes:**
  - Single-layer configurations: (1,), (2,), (3,), (4,), (5,), (10,)
  - Two-layer configurations: (1, 1), (2, 2), (3, 3), (5, 5)
  - Decreasing-layer configurations: (2, 1), (3, 2)
  - Three-layer configurations: (2, 2, 2), (3, 3, 3), **(5, 5, 5)**, (10, 10, 10), (20, 20, 20), (50, 50, 50)
- **Activation functions (`activation`):**['identity', 'tanh', **'relu'**]
- **Optimization method (`solver`):** ['adam', 'sgd', **'lbfgs'**]
- **Tolerance for convergence (`tol`):**

  [**1e-5**, 1e-4, 1e-3, 1e-2]

- **Maximum iterations (`max_iter`):** [**500**]

Furthermore, a sensitivity analysis (Figure 13) reveals that the most important variables are `TOTAL_lag1`, followed by `TOTAL_lag2`. The remaining lags, as well as the COVID variable, appear to have negligible relevance. However, when we attempted to train the model without these less significant variables, we observed a deterioration in performance. For this reason, and to provide a complete overview, we have chosen to retain these variables in this brief analysis, even though this will not represent our final model.

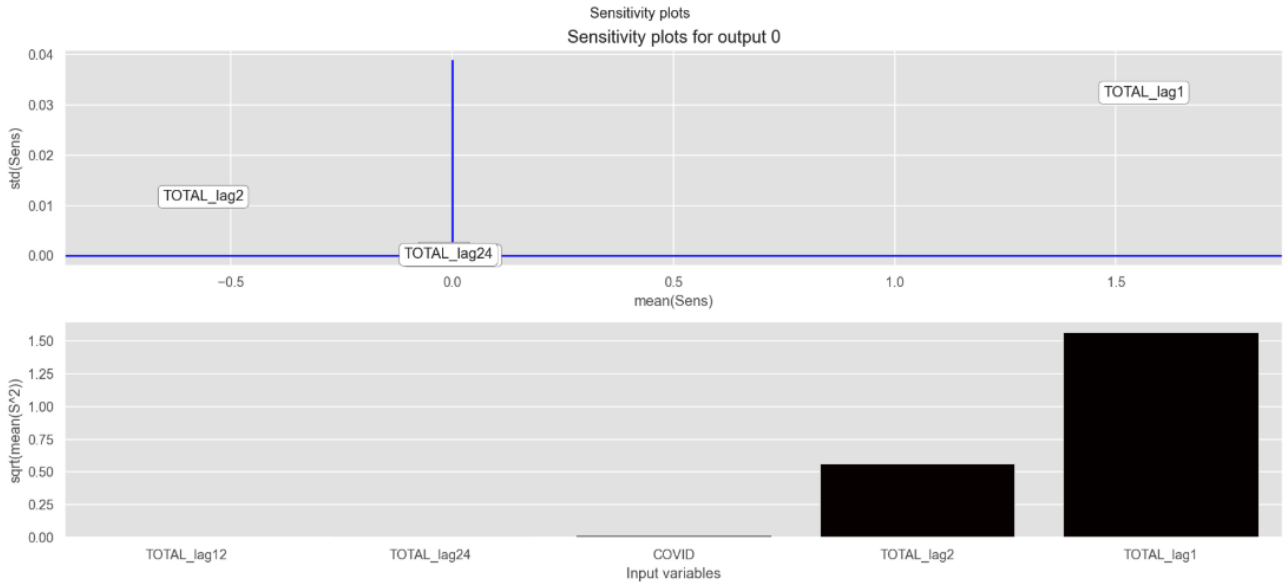


Figure 13: Feature importance of MLP inputs using Neuralsens

After training the MLP, we decided to model the error of the model (actual data - MLP forecast) using a SARIMA, aiming to improve the results. As shown in Figure 14, there is a slow decay in the ACF coefficients corresponding to the periods (lags 12, 24, 36), leading us to differentiate once in the seasonal component ($D = 1$). Additionally, in the lower part of the figure, the PACF of the differentiated residuals shows an exponential decay at lags 12, 24, and 36, as well as a significant coefficient at lag 12 in the ACF. Therefore, we considered an MA order ($Q = 1$) in the seasonal component.

| **Regular** | | | **Seasonal** | | | | |
|---|---|---|---|---|---|---|---|
| **p** | **d** | **q** | **P** | **D** | **Q** | **S** | **Trend** |
| 0 | 0 | 0 | 0 | 1 | 1 | 12 | n |

Table 2: AR and MA orders of the SARIMA model that captures the MLP residuals.

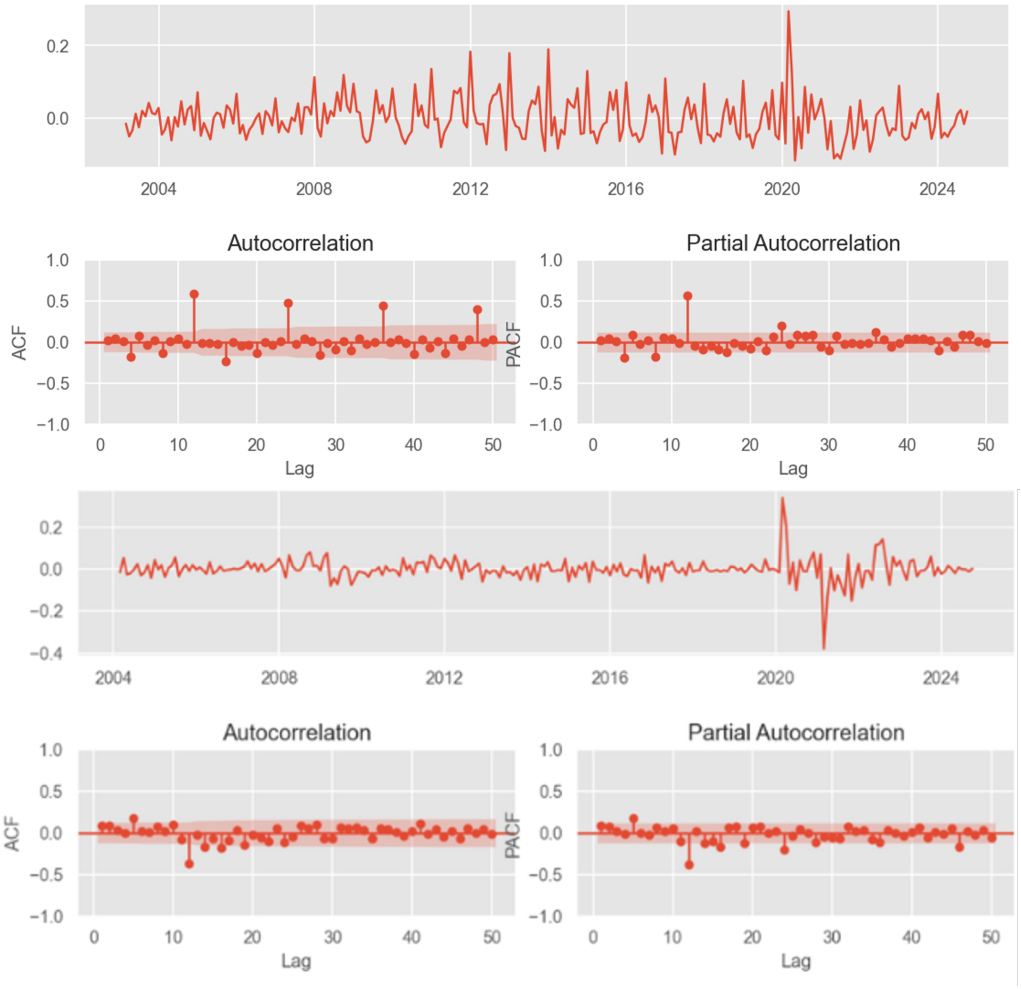


Figure 14: ACF and PACF of the residuals of the MLP model

Thus, we can see that all the coefficients are significant with a p-value = 0 (Fig. 15), and once again, we can also verify that the residual is white noise $p\_value = 0.47 > 0.20$.

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ma.S.L12 | -0.5753 | 0.031 | -18.302 | 0.000 | -0.637 | -0.514 |
| sigma2 | 0.0021 | 7.1e-05 | 29.322 | 0.000 | 0.002 | 0.002 |

Figure 15: Significance of the coefficients in the SARIMA model of the MLP errors

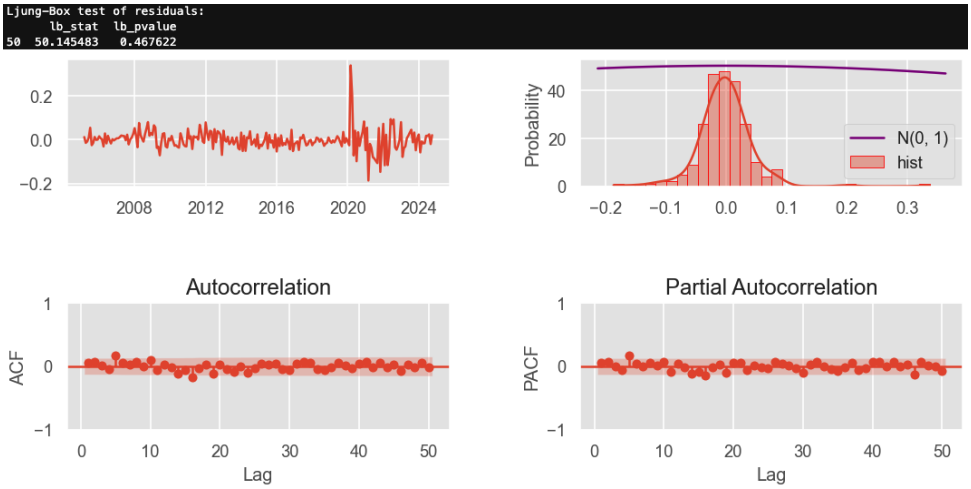


Figure 16: ACF and PACF of the residuals from the SARIMA model of the MLP errors

Finally, we obtain the model's forecast, again noticing a significant similarity between the predictions and the actual data:

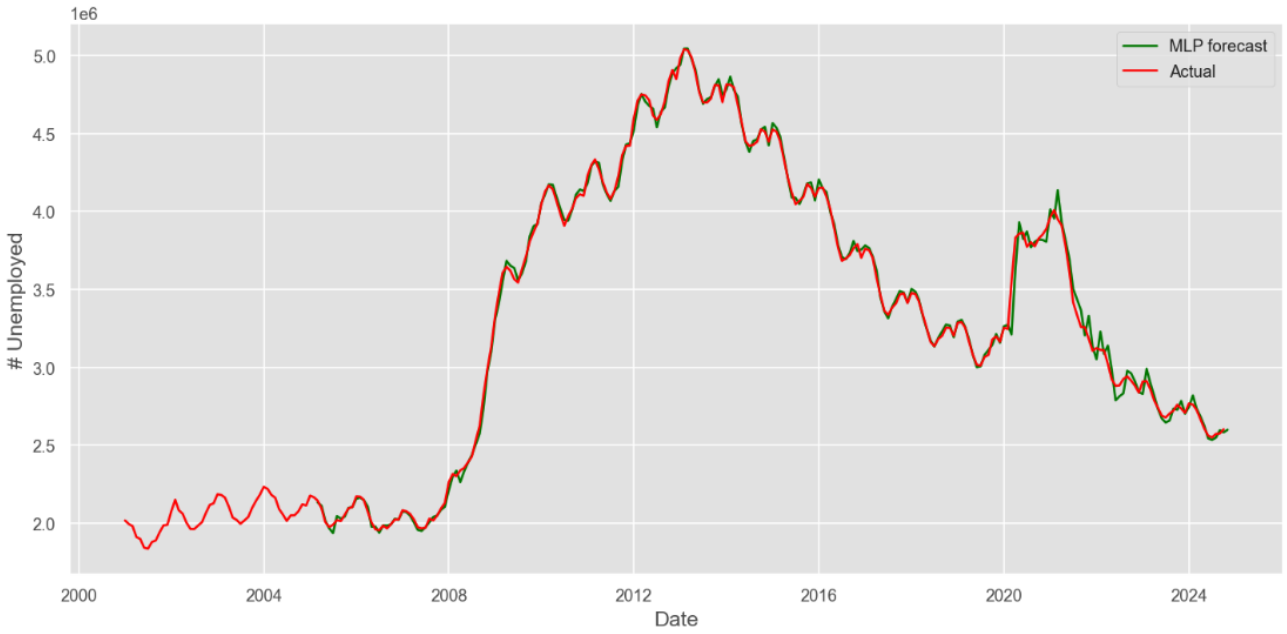


Figure 17: MLP forecast for unemployment in Spain from January 2001 to November 2024

# 5 Model comparaison

In order to decide which model is the best for predicting unemployment, we focused on comparing the MAPE during the test data period of the MLP model (starting from 2019-05-01), as this seems the most reasonable approach to compare the models' performance towards the end of the time series. This is because our main goal is to improve predictions at the end of the series.

Furthermore, as mentioned earlier, this metric provides greater interpretability, being a percentage error rather than an absolute error. Thus, in Figure 18, we can observe that the model that minimizes this error is the SARIMAX, with performance being quite similar across models.

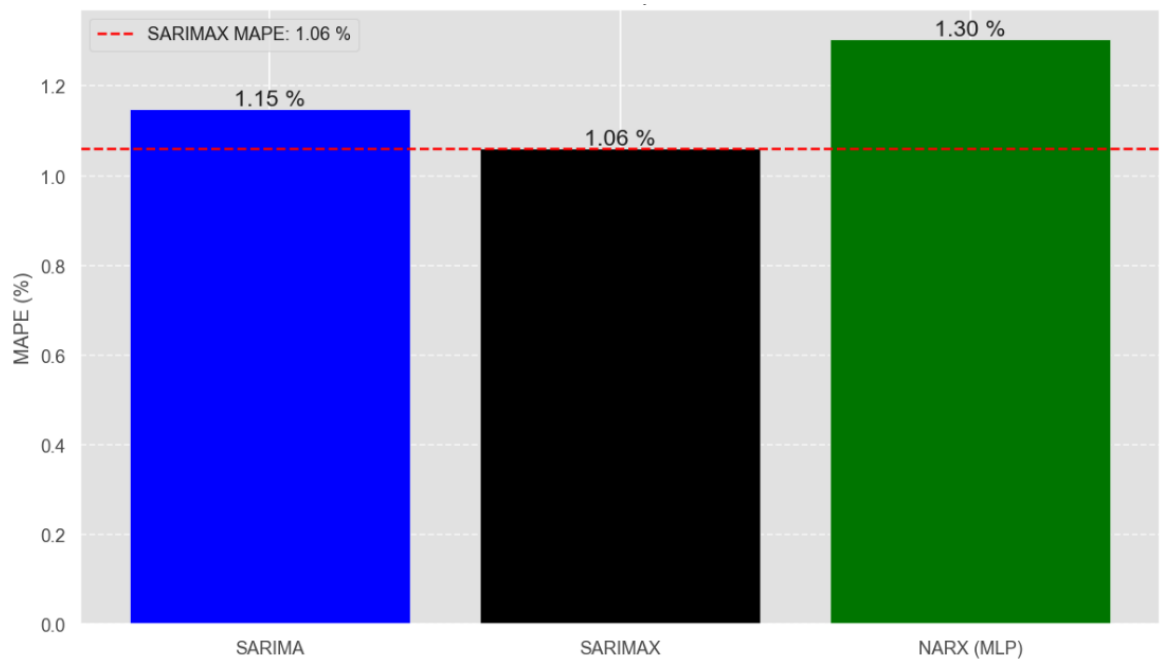


Figure 18: MAPE comparaison between models

# 6 Forecasting unemployment in November 2024

In this final part of the assignment, we will calculate the unemployment forecast for November 2024 using the three models employed. As shown Figure 20, both the SARIMA and SARIMAX models exhibit a downward trend, indicating that the number of unemployed individuals is expected to decrease. If we look at the other figure 19 displaying data for October and November, we observe that this downward trend has been consistent in recent years. However, the MLP model shows an upward trend, predicting an increase in the number of unemployed individuals, which, according to what we have stated, would not make complete sense and, as a result, leads to a higher MAPE value.

Therefore, based on the analysis in Section 5, we will rely on the prediction made by the SARIMAX model, resulting in a forecast of 2,582,958 unemployed individuals.

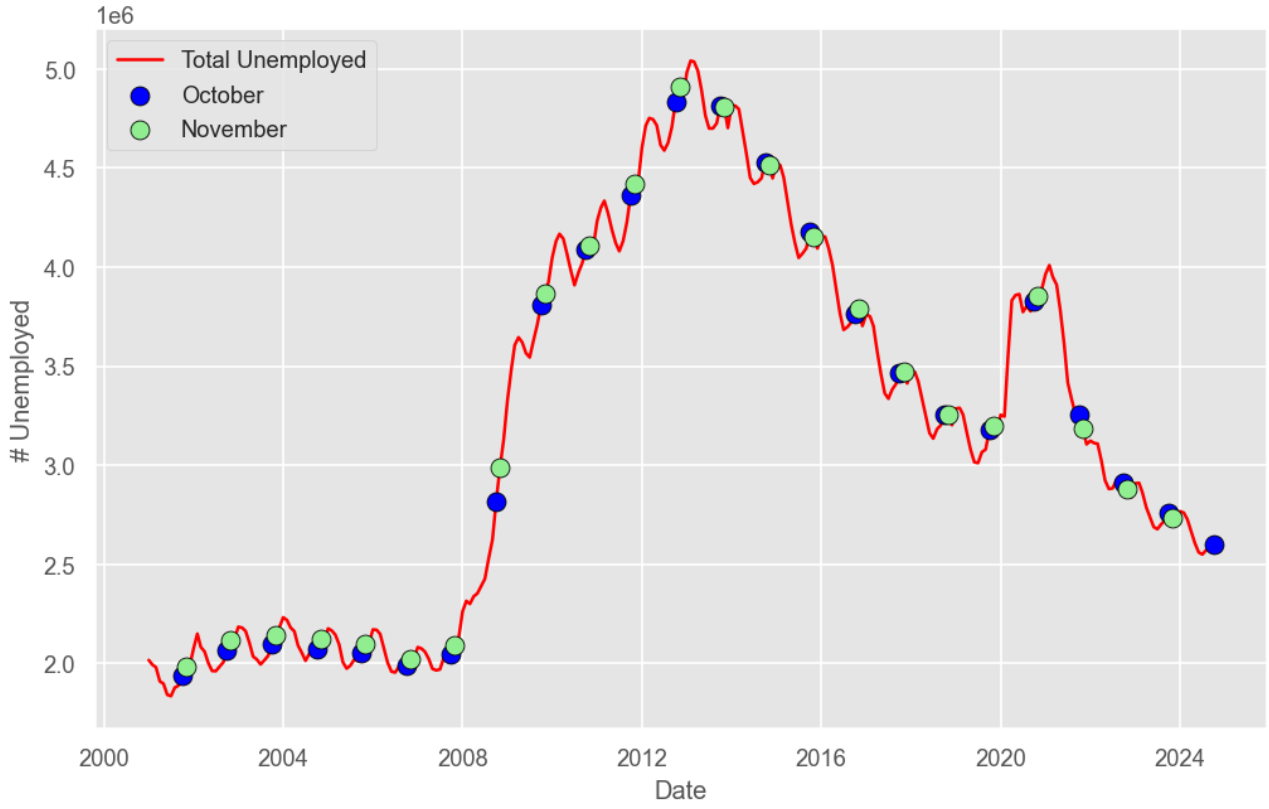


Figure 19: Unemployment data in Spain marking the months of October and November

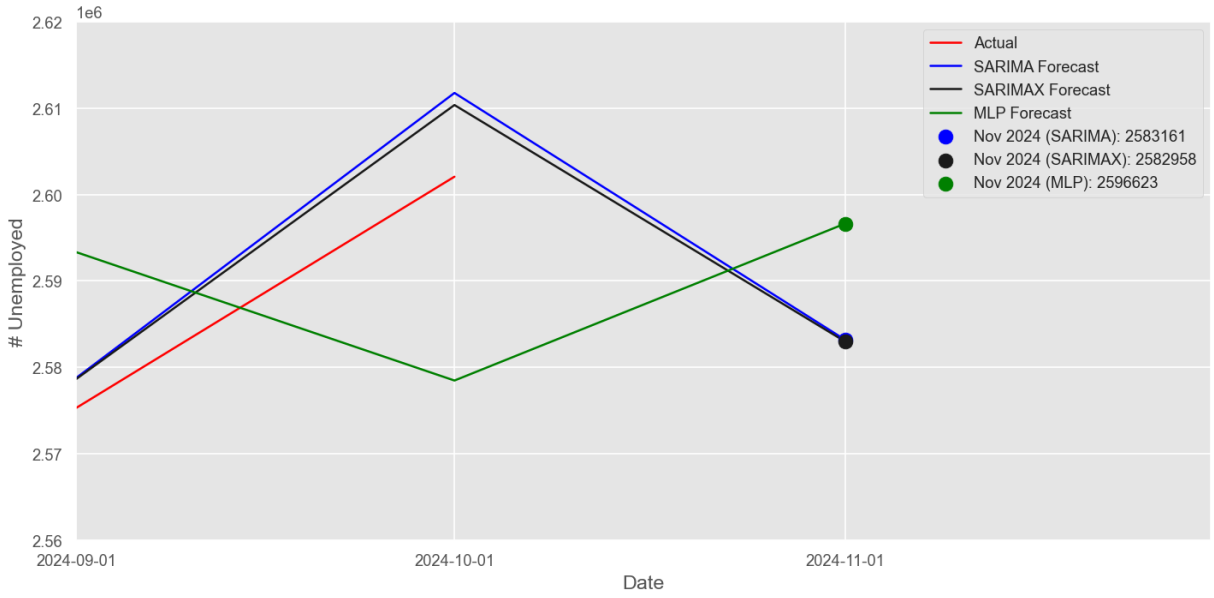


Figure 20: Forecast for November 2024 for the different models

Additionally, it is worth considering the potential impact of the DANA (isolated high-altitude depression) that affected Valencia on October 29th on the unemployment figures for November. As of today, November 30th, the information available indicates that due to the special economic aid provided to freelancers, business owners, and workers affected by this natural disaster, although the number of unemployed individuals may increase, they should not be officially counted as unemployed.

For this reason, we have ultimately not factored in any increase in unemployment, resulting in a final predicted figure of **2,582,958 unemployed individuals** in Spain for November 2024.

# 7 Conclusions

Firsty, the analysis of unemployment data in Spain shows a significant shift in trends following the COVID-19 pandemic, particularly between October and November. This emphasizes how crises can profoundly impact unemployment behavior, altering both the variance of the series and its seasonal patterns.

Additionally, the SARIMAX model proves highly effective despite its simplicity compared to the MLP model. SARIMAX achieves smaller errors and successfully captures the temporal dynamics of the series, making it a strong candidate for forecasting.

Lastly, specific events like the DANA in Valencia can temporarily disrupt the series' dynamics. While the model itself cannot inherently account for such phenomena, these events should be considered to ensure accurate predictions.