KU LEUVEN

ADVANCED TIME SERIES ANALYSIS ASSIGNMENT

UNEMPLOYMENT RATE IN SPAIN (2002-2021)

By Silvia María Goñi Mendia
Academic year 2021-2022
Prof. Christophe Croux

INDEX

MOTIVATION

As a Spanish student, I decided to use a topic that could be related to my country. While I was looking at different datasets and options, I decided to choose the Spanish unemployment rate as the main variable of the study. Currently, and I would say that for many years now, unemployment has been a big issue in Spain. Especially after the 2008 crisis, and now with the crisis generated because of the Covid-19 pandemic, unemployment is an interesting issue. I would like to clarify beforehand that during the first months of the pandemic, the government implemented the "ERTE". It is a mechanism to limit the number of layoffs, where enterprises could maintain their employees and the government would pay for most of their salary, with the condition to not fire them when the company went back in track. This explains why the unemployment rate did not increase that much for the first months of the pandemic. Lastly, I would like to explain the choice of the "tourism" variable. Spain is one of the most visited coutries in the world, so especially in the summer, a lot of jobs are created in hotels in restaurants to keep up with the demand. These jobs are eliminated usually around October.

DATASET

Country: Spain

Time period: 2002 (Q1) to 2021 (Q3) - Quarterly series

Relevant dates:

- March 11, 2004: Terrorist attack in Madrid
- 2008: Start of financial crisis
- March - June 2020: Strict lockdown due to Covid-19
- October 2020 - May 2021: Restrictive measures due to Covid-19

Other variables used in the project:

GDP

GDP at market prices, in millions of euros

TOURISM

Number of foreign tourists in hotels

UNEMPLOYMENT

Main variable

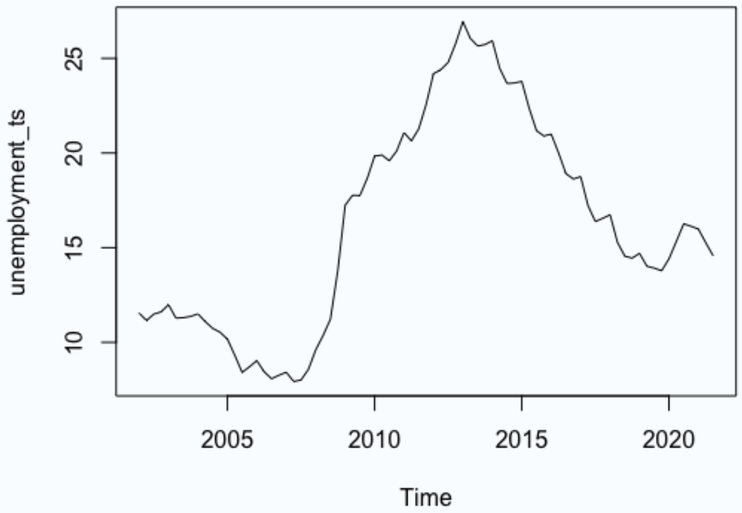
Unemployment rate. Percentage of working population that is unemployed. From age 16, both sexes.

Median: 15.98
Mean: 16.24
Min: 7.93
Max: 26.94

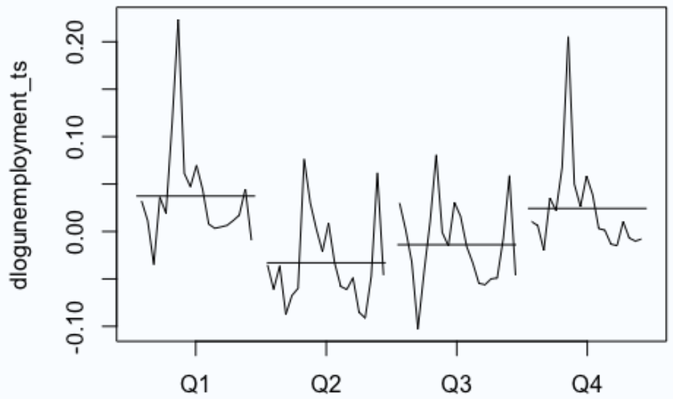
Main source: Banco de España (https://www.bde.es/bde/es/areas/estadis/)
Note: each variable has the source specified in the R Script

UNIVARIATE TIME SERIES ANALYSIS

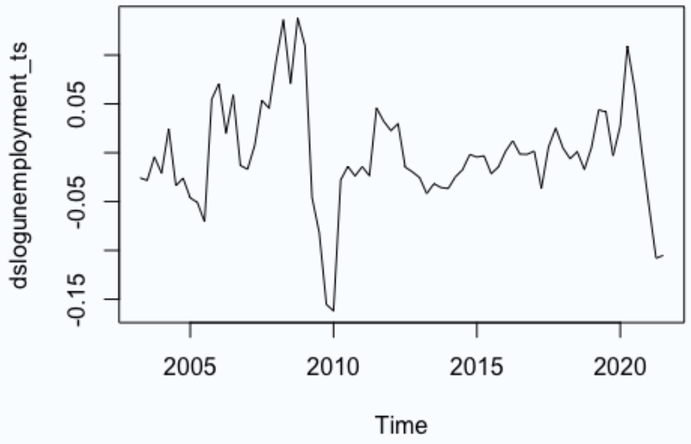
1. **Unemployment time series plot:**



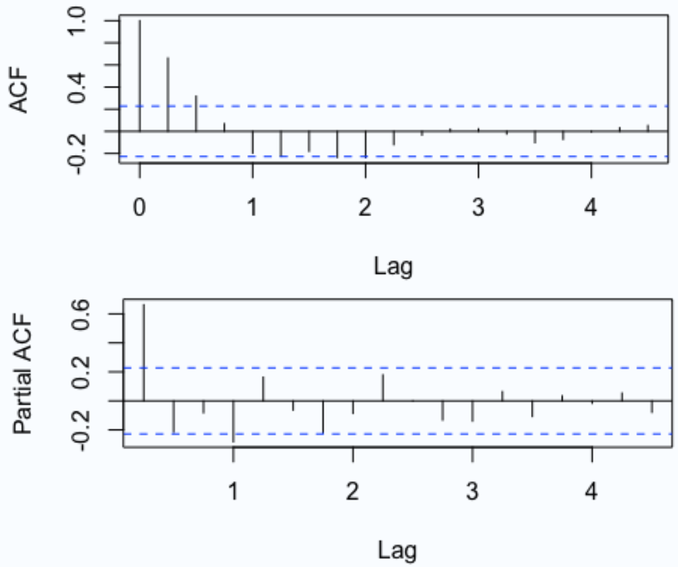
2. This time series is clearly not stationary and some seasonality can be seen.
3. Log-transformed and log-differenced time series are **not stationary**, after checking the plots and testing for **unit root** (p-value = 0.06 > 5%).



4. The **monthplot** shows the strong seasonal effect, as Q1 is the highest, and Q2 the lowest.



5. The time series becomes stationary when taking **seasonal differences.**



6. The Correlogram and Partial Correlogram of the stationary time series will be used later.
7. The **Ljung-Box test** confirms that the time series is not white noise (p-value = 0.00 < 5%).

LINEAR REGRESSION

1. Model:

$\log(population\_ts) = \beta 0 + \beta 1 TREND + \beta 2 Q1 + \beta 3 Q2 + \beta 4 Q3 + \varepsilon$

- As the number of observations –*79*– cannot be divided by three, we will eliminate the first observation –*Q1 2002*– for this model.

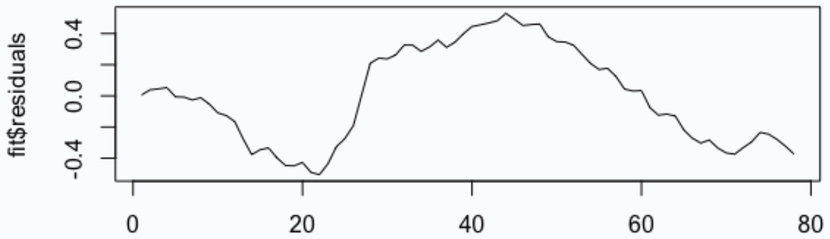
2. Results:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.380922   0.089073  26.730  < 2e-16 ***
TREND       0.008607   0.001585   5.431 6.83e-07 ***
Q1          0.015921   0.087401   0.182    0.856
Q2          0.004862   0.087357   0.056    0.956
Q3                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3149 on 74 degrees of freedom
Multiple R-squared:  0.285,     Adjusted R-squared:  0.256
F-statistic: 9.833 on 3 and 74 DF,  p-value: 1.553e-05
```

3. Interpretation:

- R2 = 0.285, that is 28.5% of the variance of log(unemployment) is explained by the regressors.
- F-statistic is associated to p-value = 0.00 < 5%, thus we reject H0 and conclude that the regressors are jointly significant.
- The p-value = 0.00 < 5%, thus we reject H0 and conclude that TREND is significant.
- Due to TREND, every year "unemployment" decreases 0.81% on average.

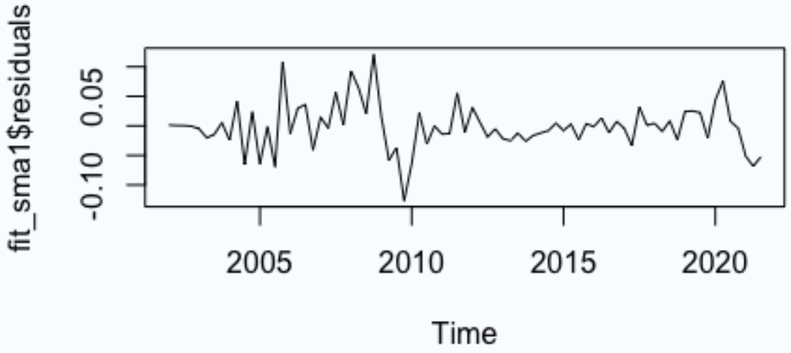


4. Validation: Ljung-Box test's p-value = 0.00 < 5%. We reject H0 and conclude that the residuals are not white noise. Therefore, this model is not valid.

REJECTED MODELS

After looking at both the **correlogram** and **partial correlogram**, some SARIMA models will be tried. Their validity will be checked through the **plot**, **correlogram**, and Q-test of the **residuals**.

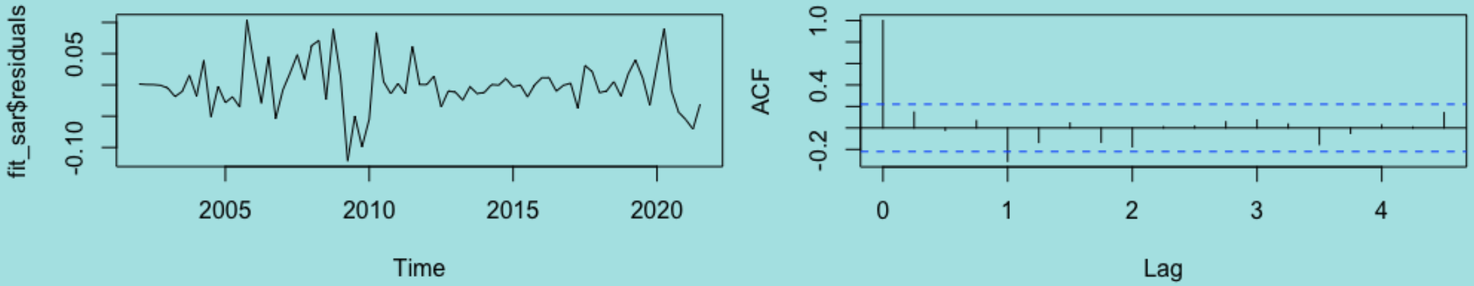
1. **Seasonal ARIMA(0,1,4)(0,1,0)**: the 4th term is not significant, so we try a new model removing this term.
2. **Seasonal ARIMA(0,1,1)(0,1,0)**: this model is not valid as the residuals are not white noise. The Q-test's p-value = 0.03 < 5%, therefore we reject it.



VALID MODELS (I)

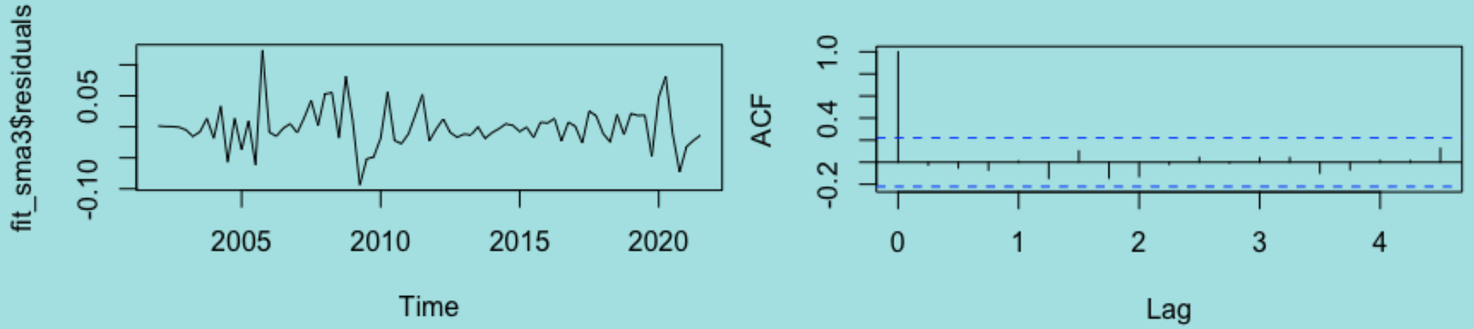
1. Seasonal ARIMA(1,1,0)(0,1,0)

This model is valid, as it has been confirmed by the Ljung-Box test, its p-value is equal to 0.06, higher than 5%.



2. Seasonal ARIMA(0,1,3)(0,1,0)

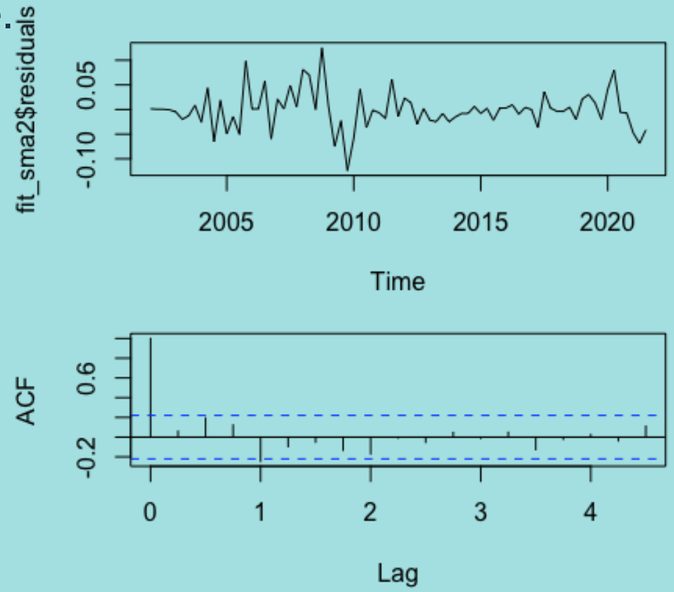
Taking into account the conclusion of the first rejected model, it is important to try this new one, with one less parameter. All the parameters are relevant, and its p-value = 0.066 > 5%. Therefore, this model is valid.



VALID MODELS (II)

3. Seasonal ARIMA(0,1,2)(0,1,0)

Last but not least, this model is tested to try to limit the number of parameters. It is the same as the second valid model, but eliminating one MA parameter in order to simplify it. After seeing the residuals' plot and correlogram, and testing the Ljung-Box test for the residuals, the conclusion is that the model is valid. The obtained p-value = 0.09 > 5%, meaning that the residuals are white noise.



COMPARING MODELS

- As there are three valid models, it is important to find the order of preference.
- Therefore, the two methods that are used to choose a model are Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC). The obtained values are displayed in the following table.

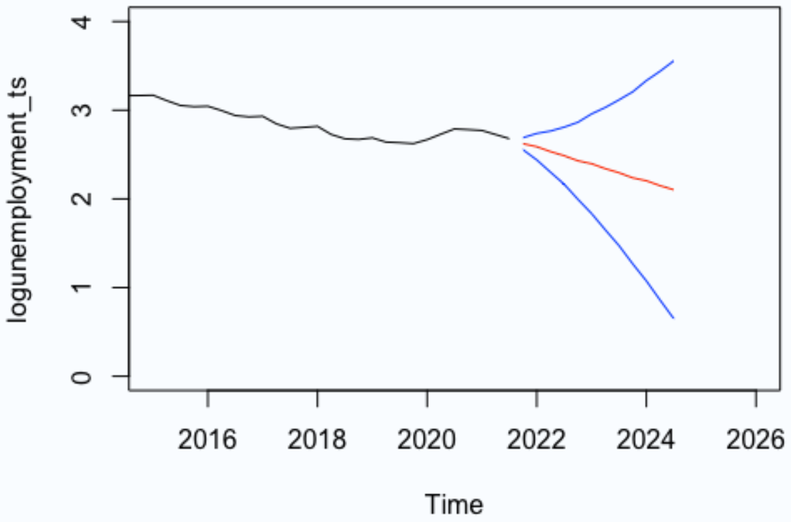
	AIC	SIC
SARIMA(0,1,3)(0,1,0)	-275.84	-266.36
SARIMA(0,1,2)(0,1,0)	-259.87	-252.76
SARIMA(1,1,0)(0,1,0)	-262.54	-257.80

- As the lowest values have been found in the SARIMA(0,1,3)(0,1,0) model, this will be the preferred one, followed by the SARIMA(1,1,0)(0,1,0) model.

FORECAST

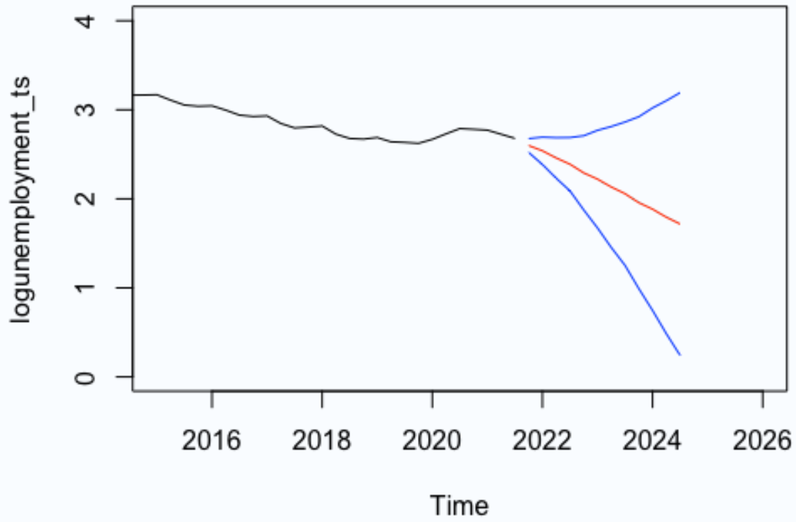
After selecting the best two models, it is possible to make **predict** how the variable will continue in the future. The models are SARIMA(0,1,3)(0,1,0) and SARIMA(1,1,0)(0,1,0). We will be forecasting the next **12 quarters**.

Model 1: SARIMA(0,1,3)(0,1,0)



The **expected** value of model 1 is higher than the one in model 2, as well as both the lower and upper predictions. In this case, as we are talking about unemployment, model 2 is more "optimistic" than model 1.

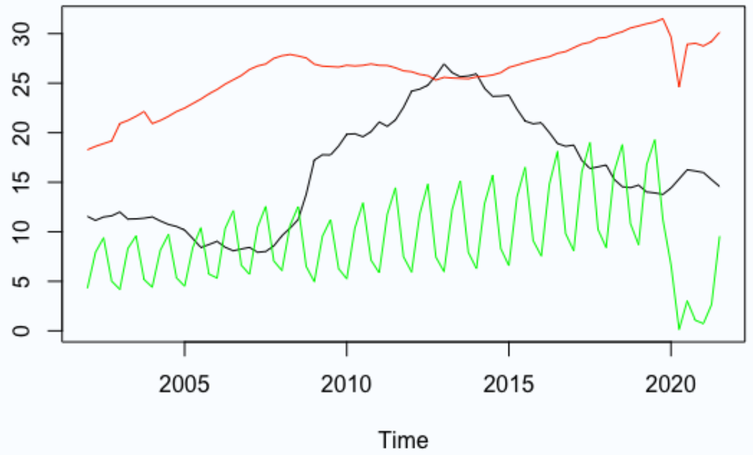
Model 2: SARIMA(1,1,0)(0,1,0)



To select a forecast, we compute the **Mean Absolute Forecast Error** and the **Mean Squared Forecast Error**. The **SARIMA(0,1,3)(0,1,0)** model has the lowest MAE in both tests, meaning that this is the better model. However, the **Diebold-Mariano** test concludes that the forecast performance of the two models, using both the absolute and squared value loss, is **not significantly different**.

MULTIVARIATE TIME SERIES ANALYSIS

ADDITIONAL VARIABLES: "GDP" AND "TOURISM"



1. Plot of the three variables:
It is very hard to see any type of relationship between the variables, except in 2020, where both GDP and tourism dramatically fall, and unemployment slightly increases.

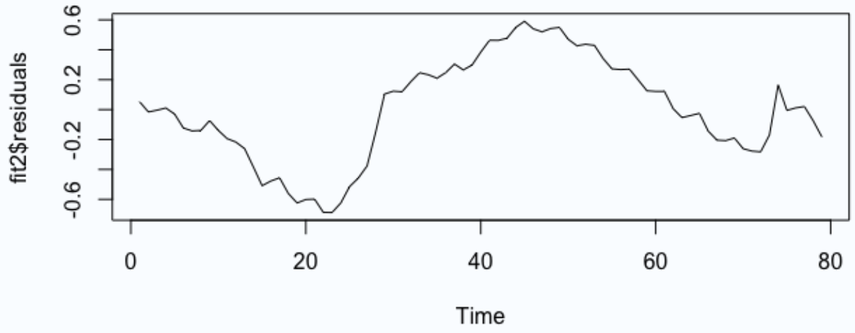
2. Model:
$\log(population\_ts) = \beta 0 + \beta 1 \log(gdp\_ts) + \beta 2 \log(tourism\_ts) + \varepsilon$

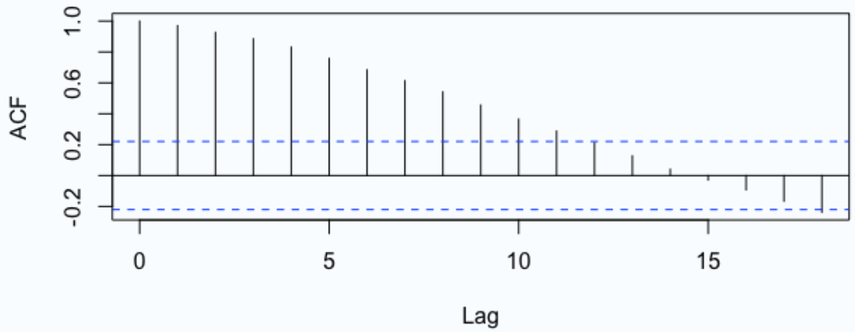
3. Results:

- $\underline{\beta 1}$ = 0.89: if GDP increases by 1%, then unemployment increases by 0.89% on average.
- $\underline{\beta 2}$ = 0.03: if tourism increases by 1%, then unemployment increases by 0.03% on average.

4. Validation:

- We check the plot and correlogram of the residuals (on the right), as well as doing a Q-test to see if they are white noise.
- The p-value of the test is 0.00 < 5%. We conclude that the residuals are not white noise, so the model is not valid.





CONCLUSIONS

- Analyzing and predicting variables, especially in time series where extreme changes are present, is more complicated.
- Both the SARIMA(0,1,3)(0,1,0) and SARIMA(1,1,0)(0,1,0) are valid to analyze and predict the behavior of unemployment.
- In the near future, the unemployment rate is most likely going to decline. However, there is still a chance it will increase, as the most "pessimistic" models predict.
- At least with the multivariate model that has been tried, unemployment cannot be tied to neither Gross Domestic product, nor the number of tourists that go to Spain.

SOURCES

Banco de España. (1965–2021). Indicadores de consumo privado. España y zona del euro. [Dataset]. Banco de España. https://www.bde.es/webbde/es/estadis/infoest/series/ie0301.csv

Banco de España. (2002–2021). Población de 16 y más años por grupos de edad y sexo [Dataset]. Banco de España. https://www.bde.es/webbde/es/estadis/infoest/series/be2402.csv

Expansión/Datosmacro.com. (2002–2021, October 29). PIB de España [Dataset]. Expansión. https://datosmacro.expansion.com/pib/espana?anio=2021