Antonio López Rosell, 2019

Bivariate Analysis and Forecasting of Suicide and Unemployment in Spain (1998 - 2017)

Summary

The aim of this project is to use the analysis of bivariate time series to understand how suicide and unemployment in Spain behaves in the period between 1998 and 2017 and to know if there is a relationship between both time series, understanding unemployment as an independent variable and suicide as a dependent variable.

In the period and data analyzed, the results of this study show that there is no correlation between both series. This result leads us to other conclusions that we will discuss in this report.

Background

Economic instability and unemployment are factors that in social psychology are associated (in many cases) with mental illness.
Durkheim, in his work on suicide (1976), noted that economic crises have an aggravating effect on suicidal behavior. Platt (1984) affirmed that there are positive associations between unemployment and suicide. Pritchard (1988) observed that, in the CEE countries, there was a significant increase in the suicide rate of unemployed men (1974-1987). In July 2009, The Lancet published a study, *The public health effect of economic crisis*, which concluded that a 1% increase in unemployment was associated with a 0.79% growth in suicide risk among those under 65 years of age.

There is enough literature to investigate a little more about this possible relationship and, thus, know a little more about the risk factors of suicide.

Suicide in Spain

According to INE, very year between 3,600 and 3,700 people commit suicide in Spain, that means that they commit around 10 suicides a day.

Different sources such as the Confederation of Mental Health in Spain, The Ecologist, El País, RTVE and the Spanish Foundation for Suicide Prevention talk about this problematic issue that is often treated as tabu and whose data aren't usually well recorded or, directly, don't exist for the public eye.

The lack of information on the characteristics of those people who commit suicide has led us to try to understand this phenomenon in Spain from a broader perspective. That's why, in a first approach, we want to understand how suicide works in its monthly numbers at national level in relation to the synchronous numbers of unemployment in Spain.

Antonio López Rosell, 2019

Method and Results

To complete our objectives we have divided the study into 2 phases:

- Univariate phase. In this phase we analyze the time series of unemployment and suicide registered in Spain for the period between 1998 and 2017.

- Bivariate phase. In this phase we calculate the existing correlation between both series based on the methodology proposed by Box and Jenkins (1970) for the calculation of cross correlation.

With the intention of explaining the project in a simple and understandable way for someone without specific knowledge about the subject, I will make some clarifications about the main concepts.

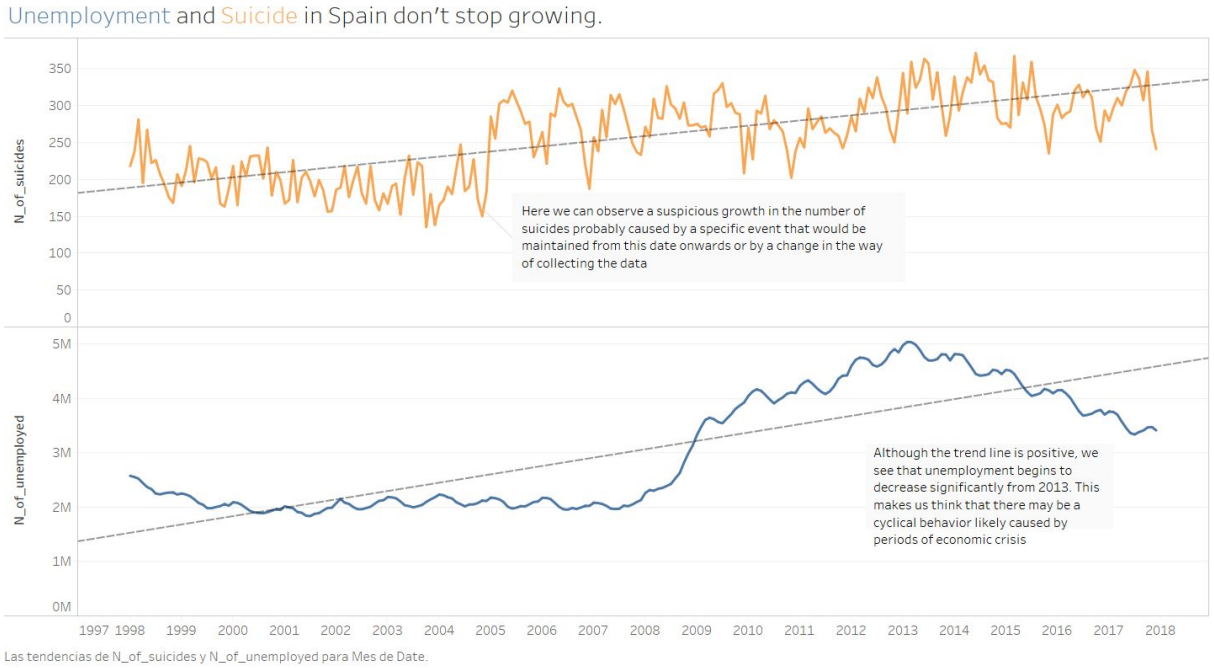
Univariate phase

First, we must know, that a time series is a sequence of data, observations or values, measured at certain times and arranged chronologically.

1. Principal Components Analysis and Stationarity

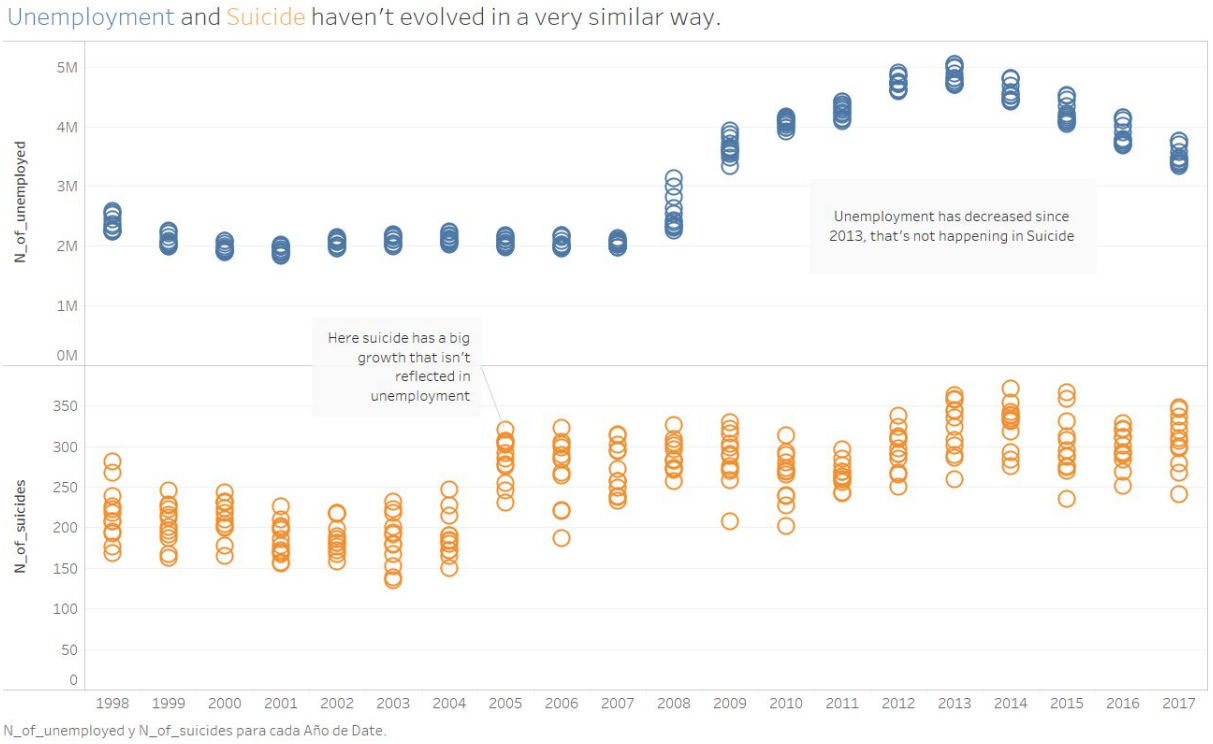
If we want to understand the behavior of our time series we have to identify its components and characteristics:

- **Trend**: It can be defined as a long-term change that occurs in relationship to the average level, or the long-term change of the mean. Plotting the data in a line plot often helps us to understand how the values of the phenomenon have evolved over time. In our case:



Plot 1. Suicide and Unemployment over the years. Line Plots

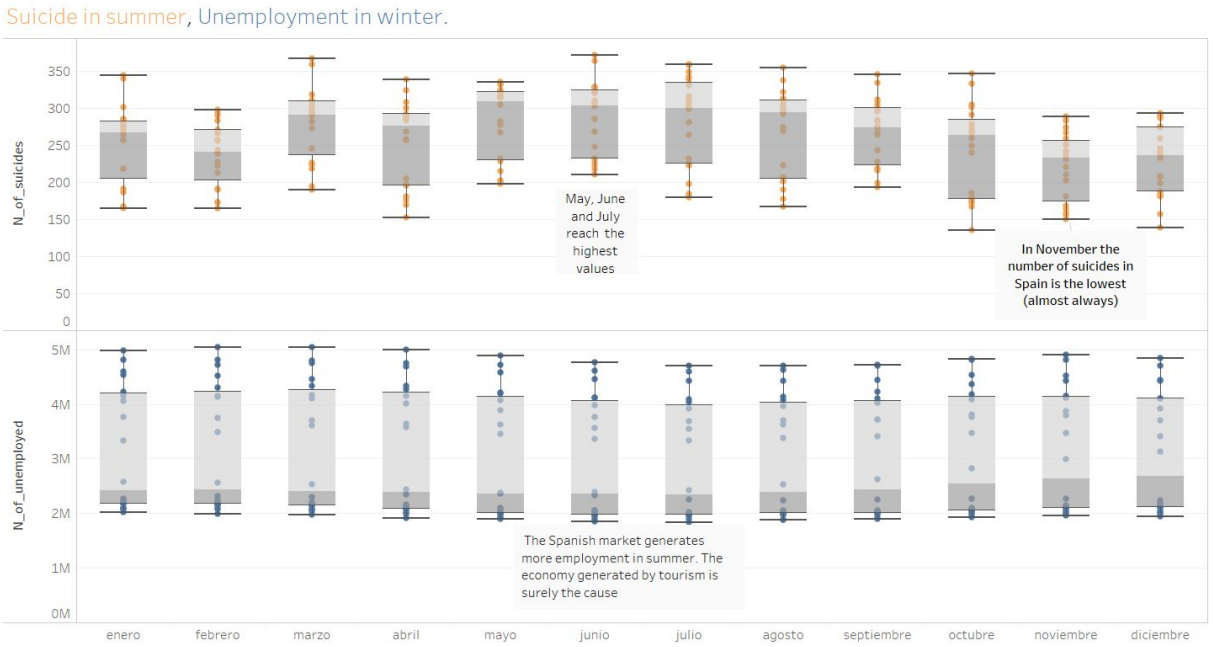
We can see in the plot that both time series show an increasing trend marked by the line. That the trend doesn't remain constant indicates that the variable is undergoing a change over time. This gives us useful information when considering making statistical models to estimate future values. Even though both series have a positive trend, we can see that they don't evolve in the same way.



Plot 2. Suicide and Unemployment over the years. Scatter Plots

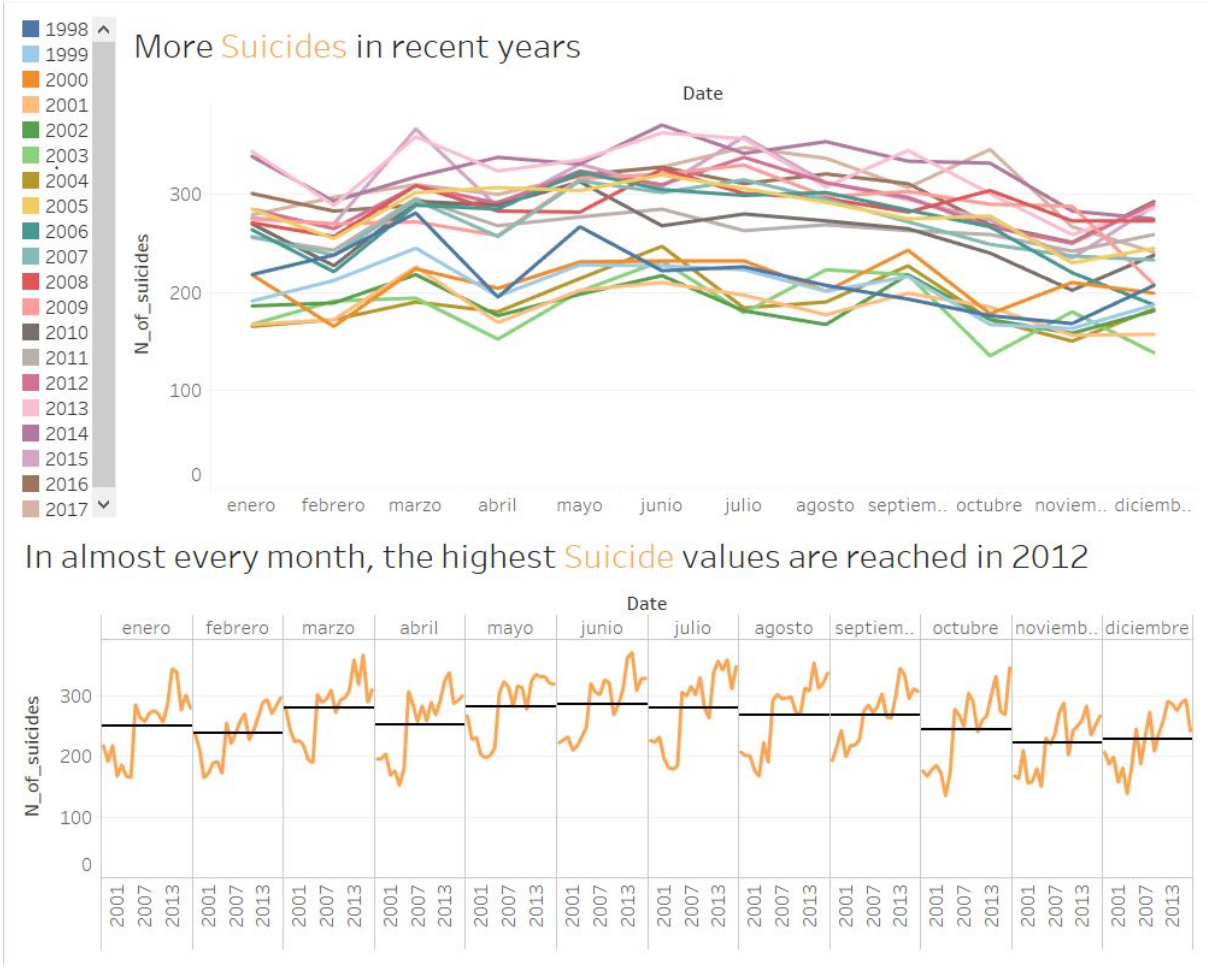
With the trend we can see the variations that the time series has in a long period of time. Another component that we have to be aware of is the seasonal one, which shows us the variations that follow a pattern for more specific periods of time.

- **Seasonality**: Many time series have variation of a certain period (annual, monthly ...). In our case, we have seen through various tests and graphs that there is a monthly seasonal pattern in both time series.



Plot 3. Suicide and Unemployment per month. Box Plots

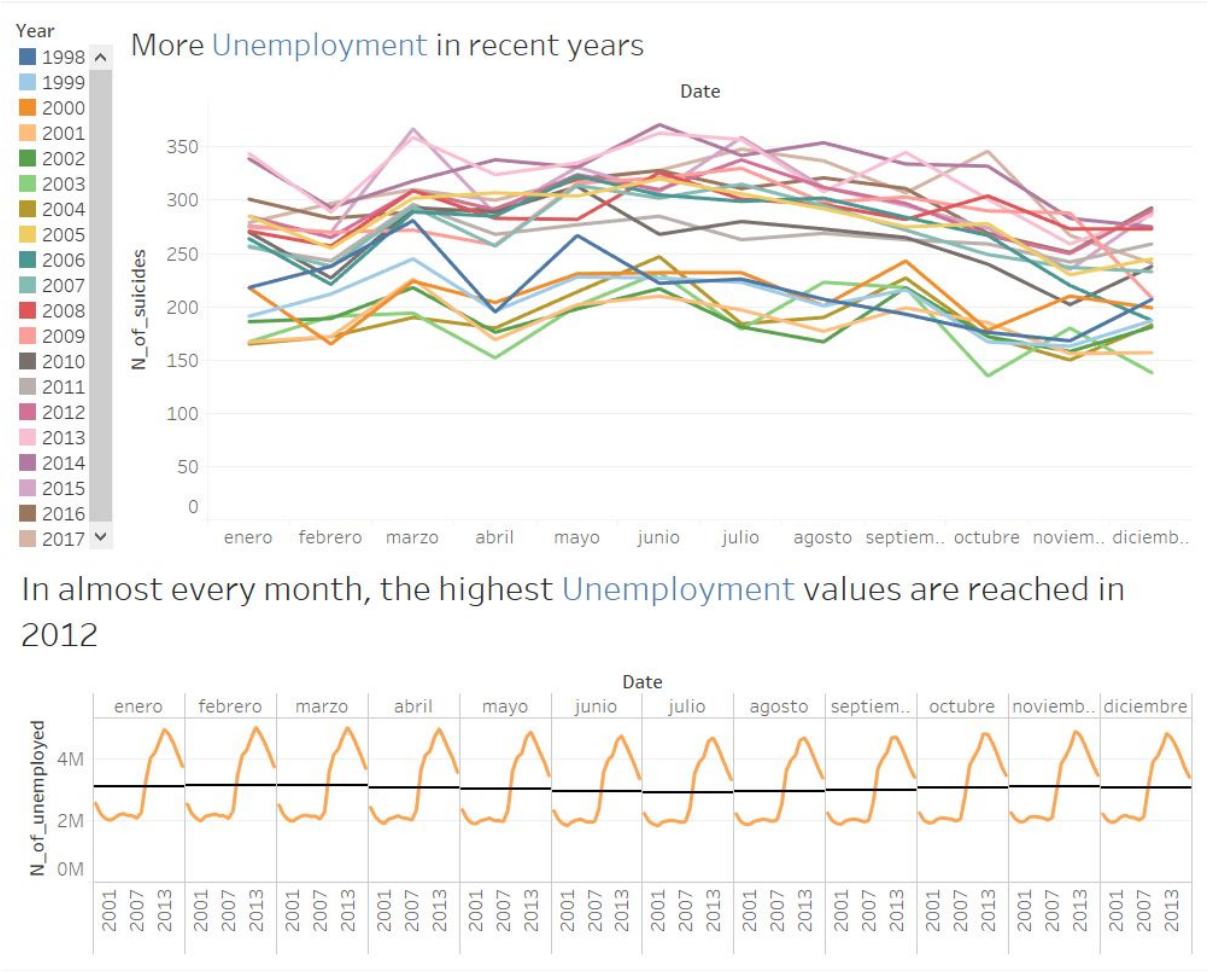
Thus, the trend component and the seasonal one give us information that helps us explain how the values of our time series have been affected over time. If we zoom in a bit we can come up with some interesting conclusions, as we showed in Plots 4 and 5.



Plot 4. Suicide values comparison by month and year

We can see how values increase every month as time goes by, which indicates the strength of the trend component when explaining the dynamics of our time series.

In addition, we can see that the lines of the monthly plot don't behave in a very smoothed way, which indicates that the values are a little scattered and that it is possible that the seasonal component varies over time.



Plot 5. Unemployment values comparison by month and year. Line Plots

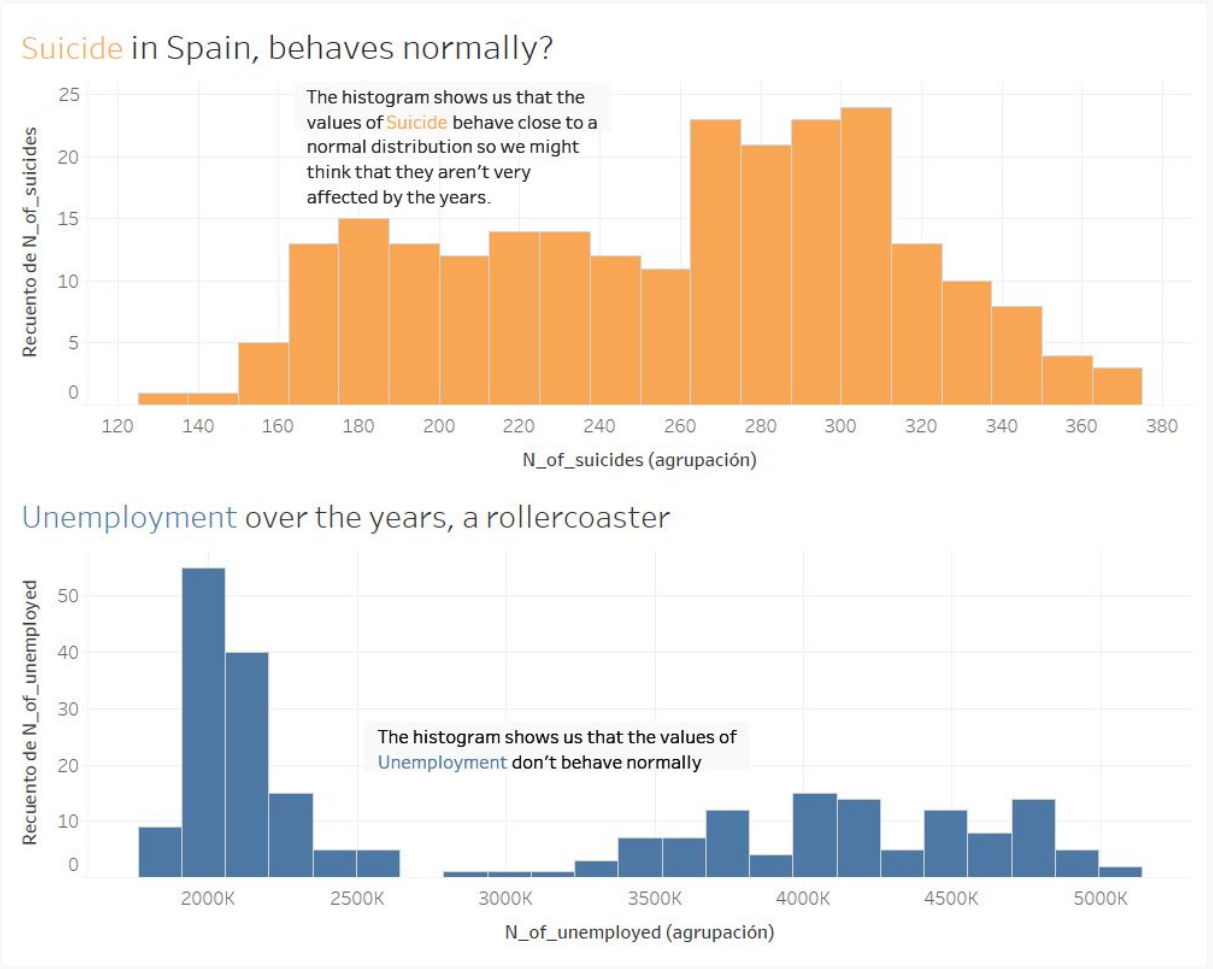
In Plot 4 for Unemployment, we can see how values increase every month as time goes by until 2013, since then, there has been a significant decrease.

In this case, we can see that the lines of the monthly plot behave in a smoothed way, which indicates that the seasonal component remains more or less constant over time and that there isn't much dispersion in its values. In Plot 1 we can see that Unemployment time series has a very marked seasonal component and surely is important when explaining this time series.

In both plots (4 and 5) we have seen that the highest values for all months were reached in 2012 and in the first months of 2013, moments in which the economic crisis in Spain was at its highest

peak. In the absence of a more exhaustive analysis, we could say that the economic factor of a country at critical moments could explain, to some extent, a growth in our dependent variable, the number of suicides.

When working with these time series and try to make estimates we must observe how they are distributed and if they behave in a 'stationary' way. Let's see why.



Plot 6. Unemployment and Suicide values distributions. Histograms

The values of our time series are not distributed normally or their normality is very weak, as we see in plot 6 and as we have verified doing normality tests in the notebook of the project. That they aren't distributed in a very normal way will make it difficult to forecast accurately.

If we want to apply other models like 'ARIMA' to make forecasts we have to have in mind that most of the time series models work on the assumption that the time series is stationary. Also, the theories related to stationary series are more mature and easier to implement as compared to non-stationary series.

But, what is stationarity?...

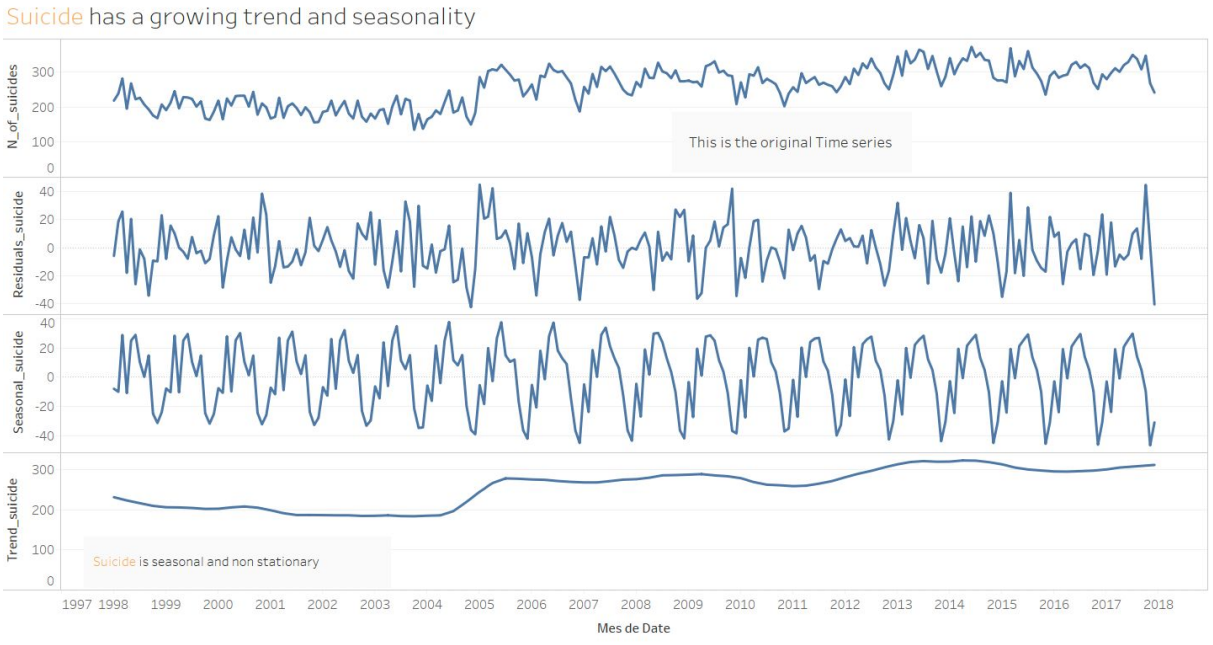
A series is stationary when it is stable, that is, when the mean and variability are constant over time. This is reflected graphically when the values of the series tend to oscillate around a constant average and the variability with respect to that mean also remains constant over time.

We have seen that these conditions are not met in our time series and are necessary to apply the chosen methodology in the realization of forecasts and in the study of the cross-correlation between both series.

To work with the Box-Jenkins (1970) approach, we have to eliminate the trend and the seasonal part (through transformations or filters) leaving only the probabilistic part (residuals, our third component). Parametric models are adjusted to this last part.

- **Residuals**: Once the previous components have been identified and after if they have been eliminated, values that are random remain (residuals).

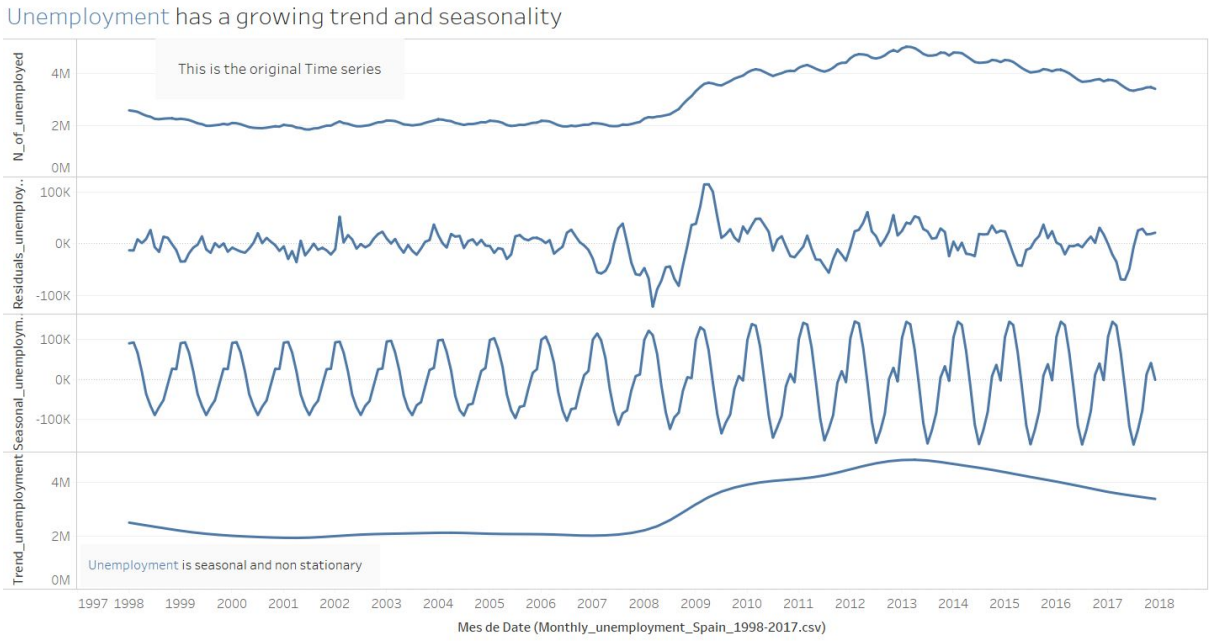
To identify these components well, we can decompose our series. We use the STL decomposition as we explained in the notebook *'Time_Series_Analysis_Forecasting_and_Cross_Correlation_Suicide_and_Unemployment_in_Spain'* because is the method that best fits what we want.



Plot 7. Decomposed Suicide time series. Line Plots

The plot above shows the original time series (top), the estimated residuals component (second from top), the estimated seasonal component (third from top), and the estimated trend component (bottom).

We have seen that this series is non-stationary and seasonal through various tests and graphics that we have developed in the notebook of the project.



Plot 8. Decomposed Unemployment time series. Line Plots

The plot above shows the original time series (top), the estimated residuals component (second from top), the estimated seasonal component (third from top), and the estimated trend component (bottom).

We also have seen that this series is non-stationary and seasonal through various tests (like ADF test and KPSS test) and graphics that we have developed in the notebook of the project.

To convert our series in stationary and non-seasonal and meet the conditions to apply in Box and Jenkins approach we have used logarithmic transformations, Box Cox transformations and the Differencing method. All these procedures are explained in the reference notebook.

Having made our series suitable for the Box and Jenkins approach, we can use the ARIMA model to make predictions.

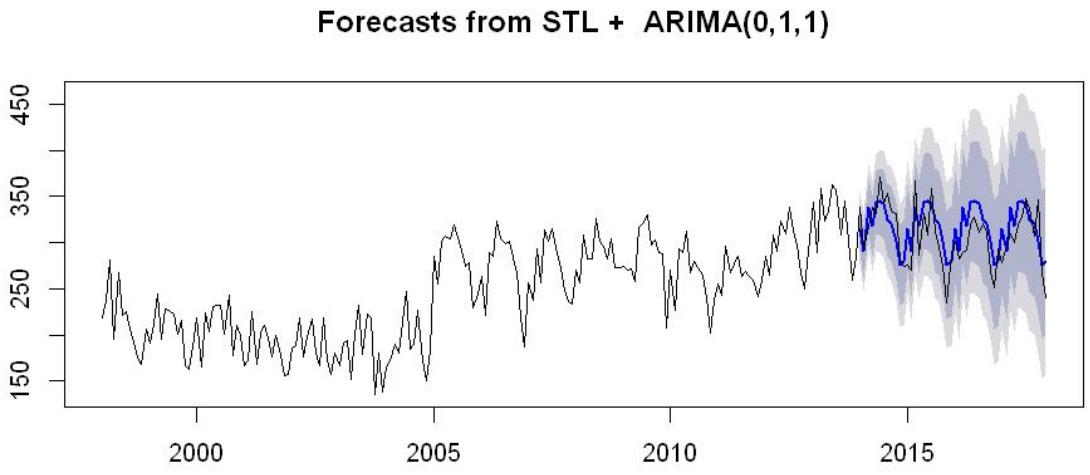
2. Forecasting

ARIMA (autoregressive integrated moving average) is a model that uses variations and regressions of data in order to find patterns to make forecasts. It is a dynamic model of time series where future estimates are explained by data from the past and not by independent variables.

To know if we are doing a good estimation we have to observe the metrics of accuracy of the model and the behavior of the residuals (to make sure that we have been able to generate information in the future that is not biased by the time variable).

To compare the estimates of the model with the real data we have separated the sample into one part for its train and another part for its test (in the test part we have data with which the model has not trained). In this way we are able to see if the training and fit of the model correspond to the real values in their forecasts (always with an error).

Throughout the reference notebook we have adjusted 3 models for each time series, below we show the one that best fits for each case.



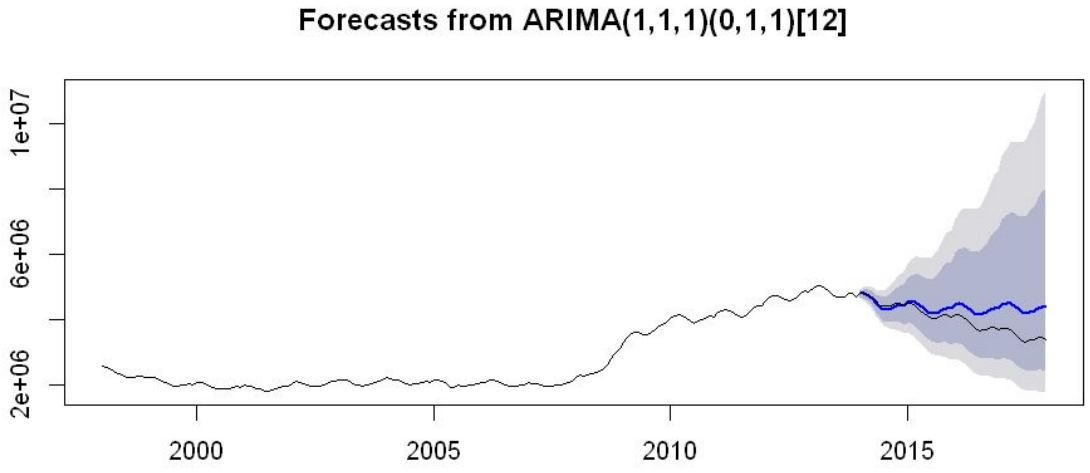
Plot 9. ARIMA forecasts for the Suicide time series. Forecast Plot

The real values are the black line, the forecasts are shown as a blue line, and the blue and grey shaded areas show 80% and 95% prediction intervals, respectively.

In the Plot 9 we can see the estimates of the model that best fits for the temporary series of suicide and It gives metrics that indicate that it isn't a bad model and that it'd work better than a naive. In addition, its residuals behave correctly, which indicates that it's a generalizable model; which we can observe comparing the train and test, and, since they do not differ much, we can assume that there wouldn't be an overfitting.

For more details and values you can check the reference notebook, 'Time_Series_Analysis_Forecasting_and_Cross_Correlation_Suicide_and_Unemployment_in_Spain'

However, we face a stochastic process and the characteristics of the object of study require much more information for more accurate predictions. Suicide is a complex social fact and this small analysis just wanted to get a little closer to this problem that needs much more dedication.



Plot 10. ARIMA forecasts for the Unemployment time series. Forecast Plot

In Plot 10 we can see the best ARIMA forecasts for the Unemployment time series. However, isn't a good model. As we can see, after 2015 our time series suffers another change of trend that we couldn't predict in our model. This means that our model is overfitted and hasn't been able to generalize despite the good behavior of its residuals. We would have to study how to include the cycles in our project and see if we need more data to estimate them.

Like suicide, unemployment is a stochastic process that has undergone many changes in Spain in recent years, difficult to predict with the information we had.

Bivariate phase

1. Cross Correlation Analysis

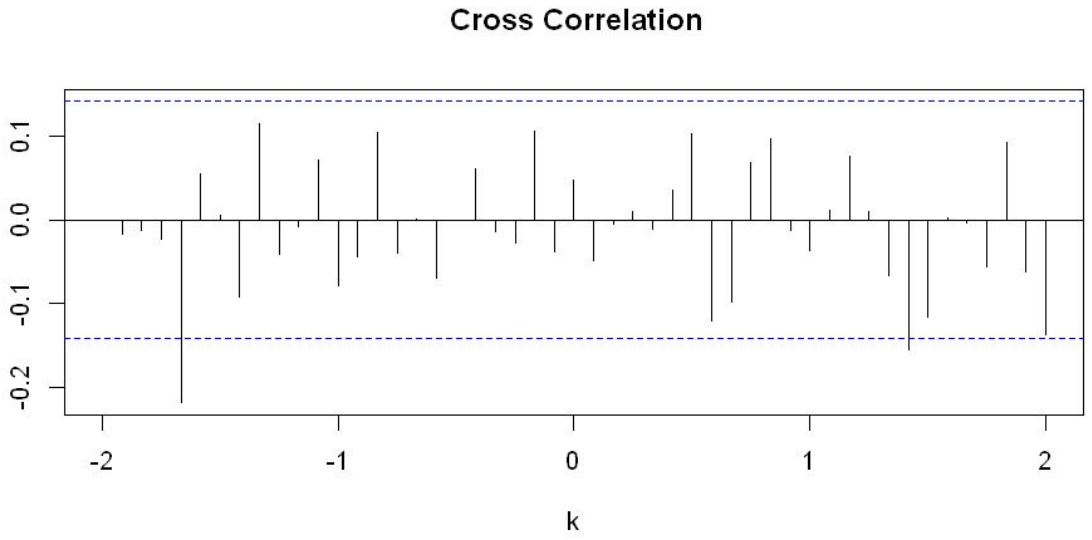
In this section, we are going to analyze the statistical relationship of the time series of Suicide and Unemployment in Spain (1998-2017). For this we are going to use the proposals of Box and Jenkins (1970) for the calculation of the cross-correlation function and the subsequent identification of the potential transfer between two systems (two time series).

In this approach the relationship between two time series is studied; relationship that does not pretend to be explanatory, but informs us about the strength and direction of the possible impact of a series 'input' to another series 'output'.

The point of this methodology and not calculate the cross correlation between the original time series is that when an input series is autocorrelated, the direct cross-correlation function between the input and response series gives a misleading indication of the relation between the input and response series. This is usually called prewhitening. First we have to fit an ARIMA model for the input series sufficient to reduce the residuals to white noise; then, filter the input series with this model to get the white noise residual series. Then, we filter the

output series with the same model and cross-correlate the filtered response with the filtered input series.

This is the result of the cross-correlation function having followed the Prewhitening methodology:



Plot 11. Cross - Correlation between Unemployment and Suicide time series.

The cross correlation function measures not only the strength of the relationship, but also its direction. This last property is useful to identify causal variables.

For this reason, it is important to examine the cross correlation function for both the positive values of k and for the negatives. For negative values of k, the cross correlation function describes the linear influence of the past values of Y(time) over X(time). For positive values of k, the cross correlation function indicates the linear influence of the past values of X(time) on Y(time).

In our case (Plot 11) the cross correlation function shows the most significant correlation in a negative value of k, so it could be said that there is an impact from one series to another but in the opposite direction to what we thought. Suicide 'would explain', in some aspect, unemployment in Spain.

A priori this doesn't make much sense, but it does lead us to other conclusions that bring us closer to understanding the phenomenon of suicide in Spain. With all this information, we would try to respond to the following possibilities:

- We are right and there is no relationship between unemployment and suicide in Spain
- There is a problem with the data collection or we need a bigger sample
- We haven't found a good fit for the unemployment series and that makes it difficult for us to study the cross-correlation
- There are one or several exogenous variables that explain the behavior of both time series
- Random events explain the trend changes in our time series. We will investigate what could have been

- This methodology doesn't fit with the problem we want to solve

So suicide and unemployment are difficult to forecast and they aren't correlated.

Conclusions

A problem with the data collection in the Suicide TS

Until 2004, in the annual record, suicide is combined with self-inflicted injuries, isn't like this in the monthly record, where only suicides are recorded, so the numbers are not the same. After 2004 this process is repeated but the numbers coincide, so the total annual of the annual record of suicides and self-inflicted injuries is the same as the total annual of the monthly record of only suicides.

We have investigated more in this possible error and, in the INE web, we haven't seen any changes in the collection methodology; apart from that after 2013 INE has access to data from the Forensic Anatomical Institute of Madrid and introduces a methodological improvement has allowed them to assign more accurately the cause of death in deaths with judicial intervention (deaths that were assigned to causes wrong defined have been reassigned to specific external causes).

In other articles they usually use annual data, where the INE joins death by suicide with death by self-inflicted injuries. So it seems there may be a problem there and, with this data collection there is no significant trend in suicide in 2005, and stay relatively constant. Although we have coincided with the majority of studies in their conclusions,, let's dig a little more to see if we can find data that match.

In a second round, browsing in the INE website, we have found the data corresponding to the monthly record of deaths for suicides and self-inflicted injuries until 2004. The total of the records coincides with the annual record of deaths due to suicides and self-inflicted injuries and this is why we will replicate the methodology with these data to work on them in this repo.

In addition to all this, several articles make reference to irregularities in the record of the number of suicides in the INE, irregularities where most of the time the causes of death are not classified very clearly.

Other exogenous variables as explanation

This project discards the existence of a cross-correlation between unemployment and suicide, rejecting the conclusions of some older studies and coinciding with more current ones.

We have also observed an event that leads us to think about the influence of other macroeconomic variables on unemployment and suicide. The year in which the economic crisis in Spain reached its highest point (2012) recorded the highest values of unemployment and suicide. This leads us to think about the usefulness of using this methodology to try to find relationships between suicide and other macroeconomic variables such as GDP, as Ramon Martín did in this article (2014), in which he also investigates the variable sex with important conclusions.

Antonio López Rosell, 2019

Lack of data on the characteristics of those who commit suicide.

We have missed the provision of monthly public data on the characteristics of those who commit suicide. Sex, work situation, marital status or country of birth would have been very interesting variables to analyze to get a little closer to understanding the risk factors of suicide.