UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Applied Statistics and Operational Research, and Quality

Interconnectedness between financial markets and the Real Economies: Modeling and Forecasting with univariate and multivariate strategies

Master's Thesis

Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

AUTHOR: Matossian , Robin

Tutor: García Díaz, Juan Carlos

Cotutor: Ferrer Riquelme, Alberto José

ACADEMIC YEAR: 2021/2022

Table of content

1. Introduction

As societies grow, they get more complex, and forecasts are becoming more relevant because many of them are the basis of contrast between the theory and the practice of the monetary and fiscal policies carried out by the governments. Economic predictions can be interpreted as a national and an international benchmark with which a government can support its analysis and propose stimuli according to the economic situations.

Although the importance of predictions is clear enough and could be defined as an inevitable accessory to guide a government decision. It is therefore equally important to identify that the quality of these predictions must be excellent, otherwise, the estimate made by the statistical model in charge could lead to erroneous policies. Note that it is not uncommon to observe fiscal and monetary impulses which have been analysed incorrectly and which have slowed down the rate of growth of one's nation. The cases are numerous and can be notably illustrated during a period of crisis (late recovery of Spain during the 2008 crisis), or of high inflation (common in the countries of South America). All these decisions are taken in a very volatile context and involve an increase in mistakes made by governments. This is why statistical analysis of economic data is fundamental, firstly to be able to interpret the world around us and derive essential information on how it is made up. And secondly, to use this information and this wealth of data to predict fundamental macroeconomic indicators such as the gross domestic product (GDP), the inflation rate and the unemployment rate (Mankiw, 2016). Therefore, throughout this work, the objective will be to analyse and predict these indicators as the basis for macroeconomic decision making.

Once the context in which companies develop and what the main indicators of their development are understood, a problem arises: the availability of quality economic information. And this is no less important, because if the available information is not reliable, the use of statistical methods is not adequate, which generates an opportunity cost for government institutions.

This problem is not trivial in economics and differs significantly from the data available in industrial processes (Boumans, 2012; Einav et al., 2008; Gunderman & Chen, 2015; Morgenstern, 1974). As a result, many factors influence the quality of economical information, including political interests, social interests, personal interests, the gathering process of the data and the media. There is a limitation directly linked to the complexity of the societies around us and human nature itself. An estimate of the error between the predicted the real value of indicators is around 10% (Kuznets, 2006).

It is therefore in a contentious context, where forecasts are a fundamental tool to guide macroeconomic policies, but their use is limited by the available economic information.

Drawing inspiration from various works such as Kitov (2011) or Stern (1993), the aim of this work will be to propose an alternative in the prediction process of the main economic indicators, using an alternative source of information trying to get around the problem of the quality of economic data.

The source of information will be data available in the financial markets. Historically financial markets were created as a source of investment for the financial agents having an economic surplus, thus,

accelerating the growth of the companies and indirect growth of nations. Although societies have become much more complex, dynamic and volatile, there may still be some evidence that ties financial markets and real economies.

But why using the financial markets? The financial markets offer a consistent quantity of information. Economic data is usually gathered quarterly or annually. Although the financial markets are perceived as volatile, the information is quantitative since all developed economies do not allow any opacity of the stock market information available. This is the reason why there is a strict regulatory framework that ensures the veracity of the information available.

It is therefore clear that if it is possible to determine a statistical link between these two data spaces it will be possible to offer a new alternative in the process of predicting macroeconomic indices, enhancing both the reliability and the quality of the information when applying government policies.

The objective of this work is, based on the information collected in the financial markets, an using statistical models, to forecast the GDP, the inflation rate and the unemployment rate, as the main indices for evaluating the economic health of the United States. This will allow decision makers to describe and predict macroeconomic variations to improve governmental policies.

The route map to fulfil this objective follows:

1. Determine the relationships between the various selected financial market indices.
2. Determine the relationships between the macroeconomic indices.
3. Determine the relationships between financial markets and macroeconomic indices as representatives of the economic health of the United States.
4. Forecast macroeconomic indices using econometric and multivariate statistical models.

2. Material and methodology

2.1 DATABASE AND SOURCES

The database is made up of financial markets indices (X space) and macroeconomic indices (Y space) of the United States.

Regarding X space, as financial markets are diverse and there is too much information to be processed, this work aims at gathering sufficient information to detect economic movements.

As a representation of the economic activity in the United States, the next three indices are selected:

- Dow Jones Industrial
- Nasdaq
- S&P500

To represent commercial relationships and how currencies fluctuations affect the United States economy the next Forex pairs are chosen:

- Euro / Dollar
- GBP / Dollar
- Canadian Dollar / Dollar
- Dollar / Yuan
- Dollar / Indian Rupee
- Dollar / Yen
- Dollar / Peso

Given the potential impact that main commodities and precious metals prices may have on economic stability, the following have been selected:

- Brent oil
- Crude oil
- Natural gas
- Gold
- Silver
- Copper

And finally, as a reference of the interest rates in the United States, the 10- and 30-years bond are incorporated:

- 10 years American bond
- 30 years American bond

All those variables will be the predictors (i.e., the X´s variables) of the different models built in this work. The data for all regressors will be gathered from "Yahoo finance" (https://finance.yahoo.com/, last access July, 2022)

With respect to macroeconomic indices, governments usually tend to predict the GDP (as the most common and used indicator for economic activity and growth), the inflation rate (as an indicator for prices stability), and the unemployment rate. Most fiscal policies are intended to affect economic growth and reduce the unemployment rate or at least control those values depending on the economic cycle. And, above all, monetary policies are aimed at controlling inflation rates, using the control and issuance of money as the main lever.

Therefore, the Y space will be constructed using these three main macroeconomic indicators. Data has been retrieved from the following official US institutions to guarantee its validity:

- GDP → Bureau of Economic Analysis, U.S. Department of Commerce, (https://www.bea.gov/, last access April 2022).
- Inflation rates → U.S. Bureau of Labor Statistics, (https://www.bls.gov/, last access April 2022).

- Unemployment rates → U.S. Bureau of Labor Statistics , (https://www.bls.gov/, last access April 2022).

Quarterly data has been collected from 1990 – Q2 to 2019-Q4 yielding a total of 119 quarters.

In the next section we will discuss the methodology of the different model we will use along this paper.

2.2 METHODOLOGY OF STATISTICAL MODELS TO FORECAST TIME SERIES

2.2.1 Linear regression

One of the most used predictive models is the regression model, although its use for time series is limited. Therefore, we will quickly treat how to validate a regression model and how it would be possible to adapt them to forecast the inflation rate, the GDP, and the unemployment rate.

Linear regression suffers from four major issues. The first and most important one is the multicollinearity, which has a direct effect on regression coefficients (high variance of the coefficient, estimates are highly dependent on each other, a small variation in the data can cause substantial variations in the estimation of the coefficients and it may appear that the model is statistically significant but that no coefficient is) (Montgomery et al, 2006). The major problem of the multicollinearity is parameter interpretation and multiplicity of potential models with similar predictive performance.

The second major issue of regression models is the lack of normality because it affects the property of the estimators and therefore, they are not efficient. The main causes are usually the appearance of anomalous data and/or the influence caused by asymmetry in the data. To correctly detect the problem of lack of normality, graphic methods are used mainly that will allow the analyst to quickly identify the existence or not of this problem, always considering that normality is relative and is not fulfilled but must approximate the same. The normal probabilistic paper will be efficient to detect normality although the skewness and kurtosis coefficients can be used. To correct a lack of normality, the variable Y can be transformed and/or the anomalous data detected can be corrected.

The third major issue of regression models is the heteroscedasticity, because estimators are no longer representative, and the formulas derived from the variance are incorrect. Eventually the last issue of regression models is the autocorrelation because it makes regression estimators inefficient.

Furthermore, once the 4 main problems in building and validating a regression model have been identified, special attention must be paid to influential data. At first, the concept of leverage must be defined, being how far from an observation *j* is from the rest of the observations, in the space of the explanatory variables. In other words, how far an observation is from the centre of gravity, measured by the Mahalanobis distance, which is a multivariate statistical measure, which, unlike the Euclidean, considers the variance-covariance matrix to consider the units of measurement between variables as well as correlation structures. Leverage measures the "a priori" influence of an observation. In addition, to detect influential data, we will use Cook's distance. This index measures both the influence of the

observation on the regression coefficients and the predictions. In this way, it is a measure of "a posteriori" influence of an observation.

As discussed, there are a lot of hypotheses that make the regression analysis complicated to validate for time series. Especially in the context of stock market data, which greatly limits normality, autocorrelation, and multicollinearity. Even though we assume that it would be difficult to validate a regression model we will perform it and try to apply the methodology to time series.

As an alternative to regression model, we could use the Principal Component Analysis (PCA) to generate a matrix of scores which are independent to each other. Thanks to PCA we could avoid multicollinearity and apply a regression model to the matrix of scores. This methodology is called Principal Component Regression (PCR). These techniques will be explained in the following sections.

2.2.2 Principal component analysis and partial least squares

In this paper the exploratory analysis of the dataset will be entirely based on Principal Component Analysis and Partial Least Squares. Notice that both methodologies are based on maximizing some properties of latent structures. PCA use only one space (e.g. the financial indices space or the macroeconomic indicator space) and by exploring its correlation structure finds out the main important sources of variability. On the other hand, PLS connects two spaces (e.g. macroeconomic indicators space and the financial indices space) by finding out those sources of variability in both spaces with maximum covariance. For this reason, we will, at first, explore each space individually with PCA and then try to connect both spaces with PLS.

The Principal Component Analysis known as PCA is a powerful and versatile multivariate method capable of providing an overview of complex multivariate data by analysing the correlation structure and condensing it in a new space of lower dimension (Bro & Smilde, 2014; Wold et al., 2001). Thanks to PCA, outliers, relationships between variables and patterns can be detected.

The PCA takes weighted average of the variables called linear combination or latent variable. A linear combination of the variables can be written as:

$$t = p_1 X_1 + \cdots + p_j X_j$$

Being *J*, the number of columns composing the matrix, and p_j the weighting coefficients.

Taking linear combination permits to obtain a new **t** vector defined in the same space as the X variables. In matrix notation the scores vector for a principal component can be re-written as follows:

$$\boldsymbol{t = Xp}$$

As supposed the most variation a linear combination can gather the better it is to condensate the information of the dataset. And if the variation taken in t is sufficient it might be useful to detect and summarize the X variables. More formally, the variation it is measured by the variance. So as just said

before the objective should be to maximize the variance of each t or linear combination taken. The formal problems become:

$$argmax_{\|\boldsymbol{p}\|=1}\ var(t)$$

As it seen above the problem should be read as finding the **w** of length one that maximizes the variance of t. as the matrix X is mean-centred. The latter problem is a standard problem in which linear algebra and the optimal **p** is the first eigenvector of the covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$/(N-1). Once understood the main objective of the Principal Component Analysis which is maximizing the variance in each new linear combination obtained, the focus will stay the explained variation.

Using as a general concept the linear regression, the columns of X will be projected on t. This is performed by regressing all variables of X on t using the linear regression equation. And then the explained variation will be studied taking advantage of the residuals.

$$\boldsymbol{X} = \boldsymbol{t}\boldsymbol{p}^T + \boldsymbol{E}$$

As observed, this a is traditional linear regression equation where vector **p** stands for the regression coefficients vector, and **E** for the matrix of residuals. So as usual the overall adjusting of the regression can be quantified by calculating the R^2:

$$R^2 = \frac{\|\boldsymbol{X}\|^2 - \|\boldsymbol{E}\|^2}{\|\boldsymbol{X}\|^2}$$

Is it not uncommon that the variation explained by the first component is too small, so the summarized information of the data could be improved. Therefore, taking more components is an efficient way to better the summary of the data. Extending the previous equation, the model can be written:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \cdots + \boldsymbol{t}_R\boldsymbol{p}_R^T = \widehat{\boldsymbol{X}} + \boldsymbol{E}$$

First, it is important to know that the PCA is a powerful mathematical method, but it is widely expanded because of the visualization of its inner parts. The analysis is meant to be based on four principal parts, the data itself, the scores, the loadings, and the residuals. In Figure 1 all PCA's components are shown.

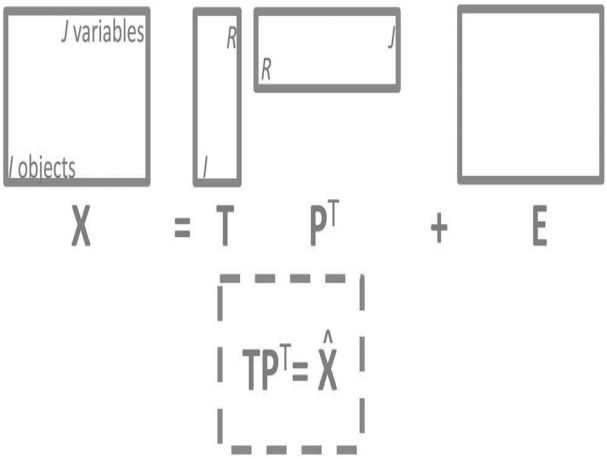

Figure 1. PCA construction

Usually, a pre-processing of the data is needed to perform correctly a PCA. To do it we should centre and scale the data. There are many other types of pre-processing methods available though. The appropriate pre-processing typically depends on the nature of the data that it is investigated.

Notice that the PCA is it not widely used in the analysis of time series though information about how to pre-process time series hasn't been defined perfectly yet. This is the reason why usual centring and unit variance scaling will be performed to the data set.

Now we must discuss how to select the correct number of components. The most common techniques used to select the number of components are the following ones:

- Use as many components as needed for exploratory studies.
- Scree test
- Kaiser´s rule**.** If the data are auto scaled, each variable has a variance of one. If all variables are orthogonal to each other, then every component in a PCA model would have an eigenvalue of one since the pre-processed cross-product matrix (the correlation matrix) is the identity matrix. It is then fair to say, that if a component has an eigenvalue larger than one, it explains variation of more than one variable. This has led to the rule of selecting all components with eigenvalues exceeding.
- Broken stick rule. A more realistic cut-off value for the eigenvalues is obtained with the so-called broken stick rule. This line is calculated assuming that random data will follow a so-called broken stick distribution. It can be shown that for auto-scaled data, this theoretical distribution can be calculated as:

$$b_r = \sum_{j=r}^{J} \frac{1}{j}$$

We have discussed how to select the number of components for the PCA model. However, we must remember that the model cannot include outliers. For this, we will have to detect, identify, understand, and eliminate them if necessary.

For PCA and PLS the same methodology will be applied to detect and treat outliers in the dataset. To this end two different statistics will be used: SPE and Hotelling´s T^2. First, the SPE (squared prediction error) measure the Euclidean distance an observation has from the hyperplane defined by the loading vectors. It detects anomalous observations. On the other hand, the Hotelling´s T^2 measures the estimated Mahalanobis distance of the projection of an observation to the centre of gravity of the data set. It detects extreme observations that may influence the fitted PCA model.

Notice that when observations are identified and defined as outliers (either anomalous or extreme observations) the analyst should look for the registered variables contributing to this outlying behaviour. This is done by using contribution plots. This is finally at this time that the PCA and PLS user should decide to include or not a sample in the modelling process.

As explained the PCA is a powerful tool to condensate, detect, and understand a complex multivariate dataset. As it defines a new component maximizing the variation in it, it makes possible a better understanding and visualization of what a system is about and how it is related. As the aim of this paper is exploratory and to forecast, the PCA will be performed on both spaces X and Y to have a first understanding of how both spaces separately work. Then the objective will be to demonstrate with PLS if both spaces can be related, and if yes, the objective will be to forecast from the financial markets the value of the macroeconomic indicators for the next year. Before starting the analysis, some clarifications will be unlighted about PLS and its principal differences to the PCA.

As we anticipated before, PLS is also a latent variable-based multivariate statistical method. The objective of PLS is to generate latent variables in both spaces X and Y that connect space X with the space Y. The objective of PLS is to connect space X with space Y not directly, but through latent variables. This is done by finding directions in X space (**w**) and in Y space (**c**) such that the resulting latent variables **t** = **Xw** and **u** = **Yc** have maximum covariance (Hoskuldsson, 2003; Wold et al., 2001).

The function to maximize is the following:

$$\underset{\boldsymbol{w},\boldsymbol{c}}{argmax}\ Cov(t,u) = \sigma_t \sigma_u Corr(t,u)$$

This way, PLS explains the sources of variability in X and Y that are correlated.

Once the function is maximized, new latent spaces are obtained for each matrix X and Y. The interpretation of the graphs is in some way analogous to PCA, but new graphs appear to complete the analysis and their relationship between both spaces:

- Scores t/u
- Weights wc/wc
- Weights w*c/w*c

In addition, to select the variables that are relevant in the analysis, the VIP is used as a measure of the relative importance of the variable X_k in the definition of the projection space. It is another way of showing the importance of the terms of the model to explain Y and to explain the projection of X.

Once the model has been validated and built, in the same way as in PCA, the prediction is carried out. The approach that will be used in this work is the regression approach whose expression is defined below:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}} \; and \; \boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}^{*}$$
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{W}}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-1}\boldsymbol{C}^{\boldsymbol{T}}$$
$$\widehat{\boldsymbol{B}}_{PLS} = \boldsymbol{W}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-1}\boldsymbol{C}^{T}, scaled \; and \; centered \; coefficients$$
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}$$

Once introduced and explained the PCA and PLS methodologies we will move on ARIMAs models which are the most used and accepted models in econometrics.

2.2.3 Arima

Based on the book Ttime series analysis, forecasting and control" (Box and Jenkins, 1976), we will explain the fundamentals of ARIMA models. It is common for time series to evolve with certain inertia due to the influence that the past event has on the present. Producing a dependency between its present heats and its past values. We can say that when observations of a variable that evolves in time are collected, there is a relationship between observations collected at different instants of time. The analysis, characterization, and exploitation of this dependence between observations at different instants of time make it possible to develop prediction models based on the variable's past.

The true mathematical model for a time series is the concept of a stochastic process. Let's assume that the observed value of the series at time t is a random extraction of a random variable defined at that time. Therefore, a time series of n data will be a sample of n vector of n random variables ordered in time z_1, z_2, z_3,…, z_n. A stochastic process is a succession of random variables Z_t that evolve. The observed series is a realization or trajectory of the said stochastic process.

To model stochastic models, simple linear models, autoregressive models (AR), moving average models (MA), autoregressive moving average models (ARMA), non-stationary ARIMA models, seasonal ARIMA models can be used.

A) AR MODELS

AR models explain the evolution of a variable from its linear dependence with another variable, in the case of time series this variable is time t. When working with auto-regressive models, as many instants of time before the current one as desired could be used, although at the statistical level very distant moments do not usually have statistical relevance.
Compact:

$$\phi_p(B)Z_t = c + a_t$$

Where $\phi_p(B)$ is the self-regressive polynomial of order p, given by:

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$$

And B is the delay operator:

$$BZ_t = Z_{t-1}$$

The parameters c, ϕ_p, are constants to be determined in the model fitting process. For an AR (p) process to be stationary, it must be verified that the roots of the self-regressive polynomial have their modulus greater than unity.

B) MA MODELS

The concept of moving average was previously studied in exponential smoothing methods. Considering that the last q values are used to make the prediction, it could be expressed by:

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

The polynomial of moving averages of order q in B is as follows:

$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$$

The delay operator is as follows:

$$Z_t = c + \theta_q(B) a_t$$

C) ARMA MODELS

Autoregressive models of moving averages are models in whose representation the two models mentioned above intervene.

$$\phi_p(B) Z_t = \theta_q(B) a_t$$

D) NON-STATIONARY ARIMA MODELS

Previously, the delay operator B has been defined, from this we can define the difference operator Nabla given by:

$$\nabla^d = (1 - B)^d$$

This difference operator makes it possible to obtain the variation of a stochastic process concerning the previous instant, thus obtaining the accumulated increases of the series.

$$\nabla Z_t = (1 - B) Z_t = Z_t - Z_{t-1}$$

When applying the Nabla operator to a stochastic process, a stationarisation of its structure occurs, an essential condition for a process to be studied as ARMA (p, q).

E) SEASONAL ARIMA MODELS

It is said that a model presents seasonality when the same behaviour pattern is repeated every certain number of observations. This phenomenon occurs with great frequency in the time series.

To model this seasonality, the seasonal difference operator (∇_s)can be used:

$$\nabla_s{}^D = (1 - B^s)^D$$

Where B is the delay operator and s the number of periods covered by the seasonal cycle of the process so that when applying it, the variation experienced by the process concerning a seasonal period is obtained.

$$\nabla_s \mathrm{Z}_t = (1 - B^s)\mathrm{Z}_t = \mathrm{Z}_t - \mathrm{Z}_{t-s}$$

A non-stationary process can be so by applying the seasonal difference operator.
Finally, the ARIMA models are defined in a general way.

$$\mathrm{Z}_t \sim ARIMA\ (p, d, q)(P, D, Q)_s$$
$$\phi_p(\mathrm{B})\Phi_P(B^s)Z_t\nabla_d\nabla_S^D = c + \theta_q(\mathrm{B})\Theta_Q(B^s)a_t$$

F) BOX-JENKINS METHODOLOGY FOR FITTING ARIMA MODELS

The time series modelling process using ARIMA models consists of three steps according to the Box-Jenkins methodology (Box and Jenkins,1976). The first stage consists of identifying the possible ARIMA model for which the stationarity of the series must first be achieved. The series must then be transformed to stabilize the variance, make the mean constant, and differentiate it if it were seasonal. The tentative ARIMA models are obtained in this way.

In a second place, the model is estimated, and the significance level of its parameters is studied. Thus, obtaining the residuals of the model.
The third and last part is the validation of the model by diagnosing the series of residues. In this phase, it is checked whether the residues are independent and follow a white noise process. If so, the created model could be used to proceed with the prediction process.

Eventually, as it can be assumed, the analysis of a series based on its past values can be very efficient, but the information extracted from the variations of other series could be used to improve the prediction of the ARIMA models. This is the reason why, the Transfer Function Models (TFM) are presented below.

As the objective of this paper is to take advantage of the information contained into the financial markets to predict the macroeconomic indicators, the ARIMA model will be used as a benchmark for this study.

2.2.4 Transfer functions models

In the models studied so far, the prediction of a time series based on its past has been considered, constituting the well-known univariate models. Now it is introduced how to make predictions of a time series when it is known in addition to its past, the past and present of another time series that may be related to it. In this section we study the relationship between an input X_t time series and an output Y_t time series. This relationship is known as a dynamic regression model or simple linear regression models. In many other cases, this relationship may not be instantaneous and present a more complex

nature, transmitting with certain delays in time. It is common to find that the dependence of Y_t can be not with X_t but with X_{t-k}, the challenge k being unknown a priori, or with all the past values of X_t.

As a hypothesis, it is assumed that the present and the past of the input influence the present of the output but not the other way around, that is, there is no feedback or "feedback" effect. Therefore, it will be true that the correlation between Y_t and X_{t+k} is null, and the future heats of X do not depend on the past and present values of Y, and the correlation between X_t and Y_{t+k} can be non-null depends on the future values of Y of the past and present values of X.

To explain the dynamic relationship that exists between a stationary time series impute X_t and a stationary time series output Y_t we can formulate the following model known as Transfer Function Model.

$$Y_t = V(B)X_t + N_t = \left(\nu_0 + \nu_1 B^1 + \nu_2 B^2 + \cdots + \nu_k B^b\right)X_t + N_t$$

Where V (B) is known as a transfer function of order k and its coefficients ν_i, where N_t is a noise process that encompasses the combined effects of all the other factors that are influencing Y_t and are not considered in the model.

Transfer Function Model with a single input:

$$Y_t = c + \frac{\omega_s(B)}{\delta_r(B)} B^b X_t + N_t$$

$$\omega_s(B) = \omega_0 - \omega_1 B^1 - \omega_2 B^2 - \cdots - \omega_s B^s$$
$$\delta_r(B) = 1 - \delta_1 B^1 - \delta_2 B^2 - \cdots - \delta_r B^r$$

Where,
Y_t time series to predict (output)
X_t explanatory time series (input)
N_t noise process and $N_t \sim$ ARIMA
c, r, s, b are constants to be estimated

From this methodology we hope that we could improve the result obtained in the ARIMA modelling. Once the most common econometrics models have been discussed we will briefly explain the methodology behind Neural Networks and Recurrent Neural Networks.

2.2.5 Neural networks

Artificial neurons are modelled in such a way that they imitate the behaviour of a cerebral neuron. The information will be processed in what so-called a "node" and through processing this information the output will be generated to other neurons. We can think of the connections between artificial neurons and in the synapses of the brain neurons (Hill et al., 1996). The image of a typical artificial neuron is as follows. In this project we will work with simple neural network as shown in the figure below.

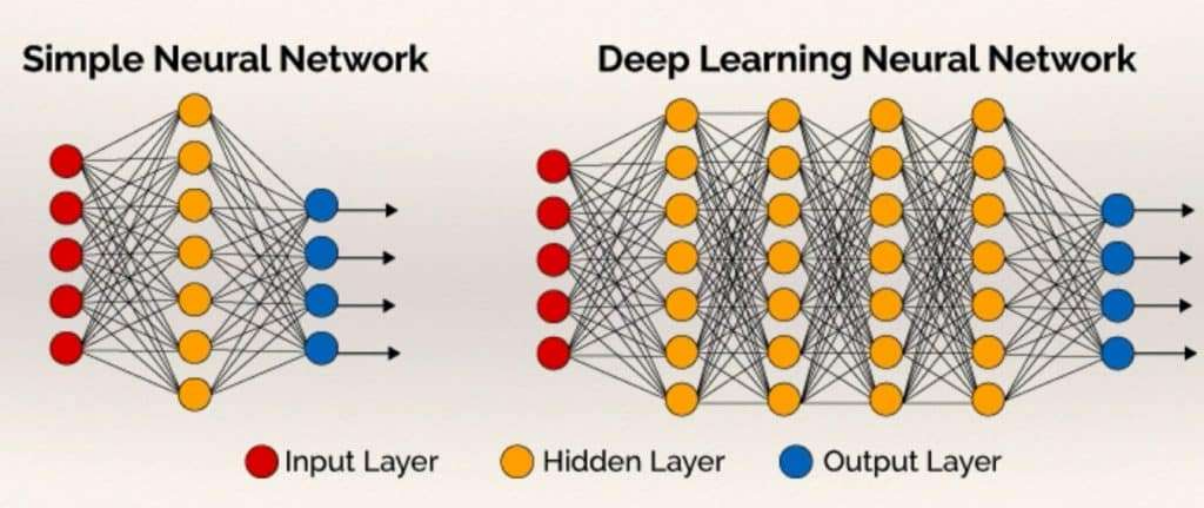

Figure 2. Neural networks' structure

In this study case, we will work with financial data as an input for the input layer (red neurons). Yellow neurons do not generate an output but process the data and created connections between the financial data provided and the dependent variables. Finally, we will obtain the output through the blue neurons (output layer). In our case the output layer will provide us the prediction for the gross domestic product, the inflation rate, and the unemployment rate.

When processing the information in neuron, activation functions are considered. First, the linear combination of the weights and inputs is calculated. Activation functions are the way to transmit the information by output connections (red to yellow, yellow to blue neurons).

In general, activation functions are used to give a "non-linearity" to the model and that the network can solve more complex problems. If all activation functions were linear, the resulting network would be equivalent to a network without hidden layers.

To summarize neural networks, we will work with three main components:

- Layer of input neurons**,** where we will pass information to the network.
- Layer of hidden neurons**,** which will oversee processing the information provided by the input layer.
- Output layer, with which we will obtain the result of the forecast.

As neural networks are not specially made for time series, we will discuss the Recurrent Neural Networks which could take advantage of the dynamic of the series.

2.2.6 Recurrent neural networks

Based on the work "Recurrent Neural Networks and Robust Time Series Prediction" (Connor et al., 1994), we will now discuss the recurrent neural network as they include connections point to the past of the data. As we can see in the figure below, the output Y depend on the last neuron inside the layer.

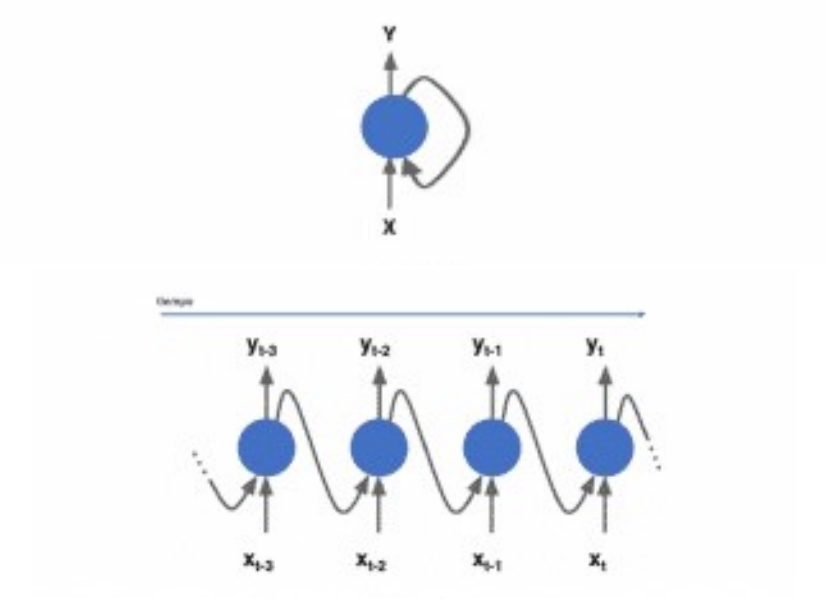

Figure 3. Recurrent Neural Network cell and Timestep

A) TIMESTEP

A recurring neuron receives the entry X of the previous layer, as well as its own output of the previous time to generate its output. This is called timestep and is illustrated below.

Now recurring neuron body has two sets of parameters, one that applies to the entry of data it receives from the previous layer and another set that applies it to the data entry corresponding to the vector output of the previous moment.

B) MEMORY CELL

Since the departure of a recurring neuron at a certain moment of time is a function of the previous time moments, it could be said that a recurring neuron has in a certain way memory. The part of a neuronal network that preserves a state over time is usually called Memory Cell (or simply Cell).

In conclusion, as neural network can take advantage of not linear relations, recurrent neural network can take advantage of the initial information of the inputs. Even though the interpretation of the weights of both models its far from what PCA and PLS can offer, we presume that the forecast in term of MAPE should be concluding.

Continuously, we will explain how to model transfer function to take advantage of the dynamic of the series and apply it to the previous multivariate models we have seen so far.

3. Exploration of the dataset and determination of the relationship between financial markets and macroeconomic indicators.

Once the methodology, the data and the statistical techniques used in this paper have been introduced, the focus will be on the results obtained. Firstly, PCA is going to be used as an exploratory method to determine the relationships between financial markets indicators. The same approach will be performed on the macroeconomic indicators. Then, once understood the correlation structures within each space, PLS is going to be used to study the relationship between both spaces X and Y. Finally, several statistical methods will be used to forecast the main macroeconomic indicators using the financial markets indices as regressors, aiming at determining which are the best fitting methods for this problem.

3.1 ARE FINANCIAL MARKETS CORRELATED?

Historically, financial markets were created to be able to finance, through the purchase of underlying assets, any entity fluidly and reliably. Thus, constituting an agile mechanism that connects financial agents with a surplus and those with a deficit. It should also be remembered that financial markets are usually very liquid and, therefore, they may experience a lot of volatility in relatively short periods.

To explore the financial markets indicators space, a PCA is carried out. In the first place, the constructed model must be validated according to the SPE chart (Figure 4) (detection of rare individuals) and Hotelling's T^2 chart (Figure 5) (detection of extreme individuals) obtained after extracting 7 principal components with a goodness of fit (R^2) and goodness of prediction (Q^2) close to 98% (Figure 6).

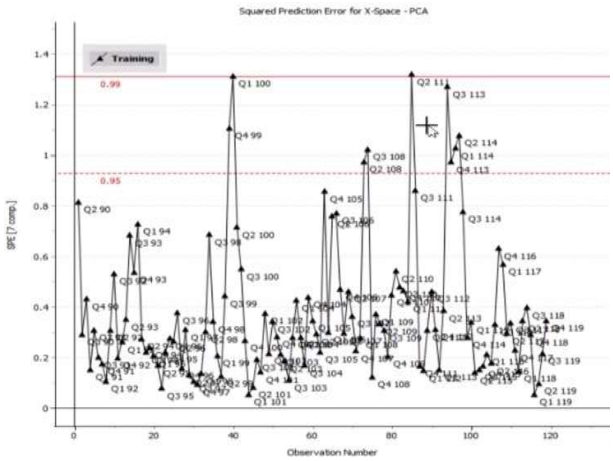

Figure 4. SPE chart for PCA on financial markets indicators

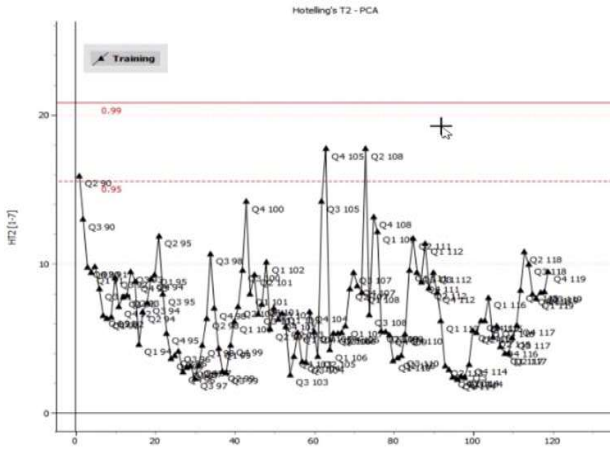

Figure 5. Hotelling´s T^2 chart for PCA financial markets indicators

Considering that the database has 119 quarters, and none exceeds the 99% confidence limit and, given that it is expected that on average 1% of the observations will be outside these limits, the PCA model can be validated.

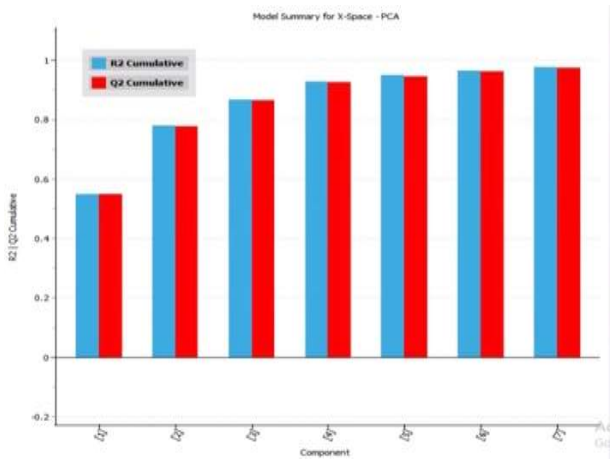

Figure 6. PCA model summary

Figure 6 shows that the two first principal components explain almost 80% of the variability. Given that the objective of this PCA is not predictive but an exploratory tool that will improve the general understanding of the data, we present in the following the results in terms of the two first components.

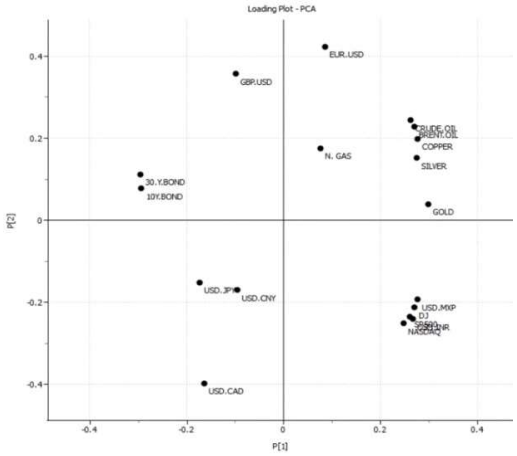

Figure 7. PCA loading plot (p_1/p_2) of the financial markets' indicators

Figure 7 shows the p_1/p_2 loading plot revealing the relationships between the financial indices used in the database, i.e., commodities (Gold, silver, oil), indices (Nasdaq, Dow Jones, S&P 500), bonds, and some FOREX indices. As we can see, certain financial indices are correlated (i.e., those that lie close or in the antipodes, and far away from the centre). For example, there is a clear positive correlation between the following indices, "Crude oil", "Brent oil", "Copper" and "Gold", but also between the following indices, "Dow Jones", "Nasdaq", "GDP" and "SP500". A negative correlation can also be observed, for example, between the price of gold and the American bond at 30 years. To appreciate these relationships, the scatterplots of the most important and clear relationships are shown in Figures 8 to 15.

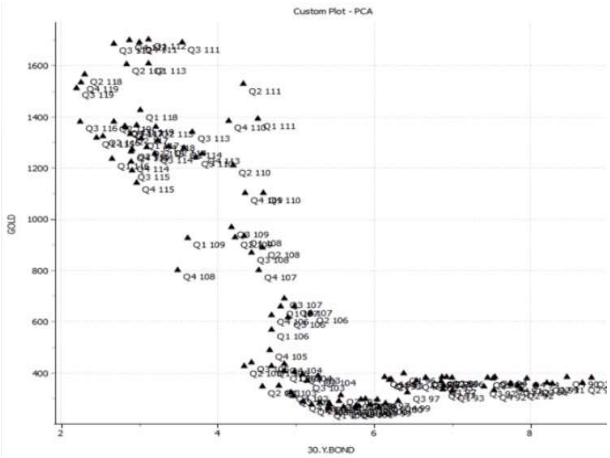

Figure 8. Scatterplot - Gold & 30 years bond

In Figure 8, a scatterplot of the price of gold and the interest rate of the 30-year bond is shown. A clear negative correlation between the gold prices and the 30 years bond can be observed. This is coherence with the antipodal position of both variables in the loading plot of Figure 7. Notice that the current values of the gold are above 700$ since the Subprimes crisis, and this relation is more lineal in this section.

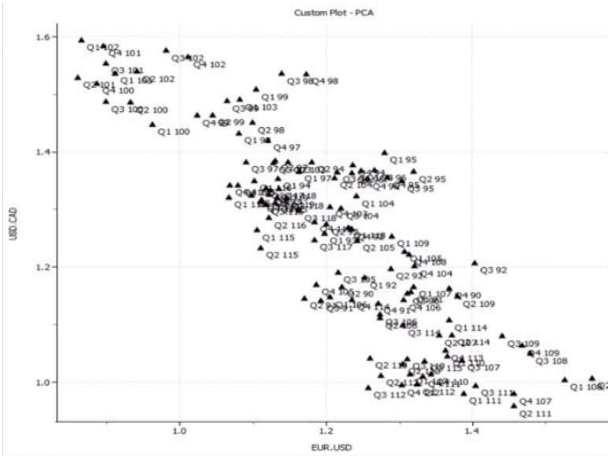

Figure 9. Scatterplot - USD/CAD & EUR/USD

In Figure 9, a scatterplot of the USD/CAD and the EUR/USD is shown. In coherence with the antipodal position of both variables in the loading plot of Figure 7, a clear negative correlation can also be observed when scatterploting the USD/CAD and EUR/USD forex pairs, meaning that it appears that when euro is more expensive than the dollar the relation between the dollar and the Canadian dollar is weakened.

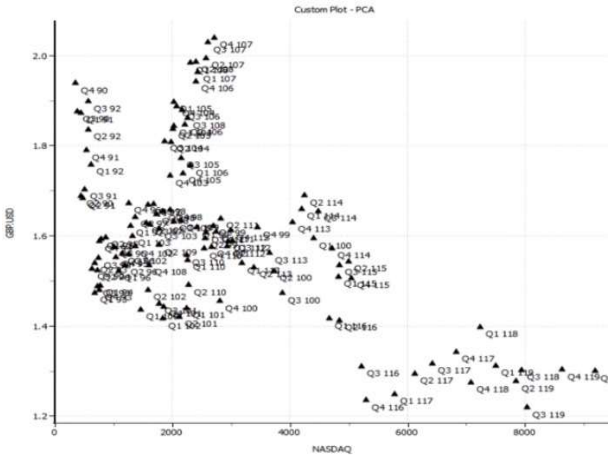

Figure 10. Scatterplot GBP/USD - Nasdaq

Figure 10 shows a scatterplot of the price of GBP/USD and the Nasdaq. A negative relationship is observed between both indices (this is coherence with the antipodal position of both variables in the loading plot of Figure 7).

Figure 11 shows the scatterplot between 30 years bond - 10 years bond. A strong positive relationship is shown (note that both variables are found together and far away from the center of the loading plot of Figure 7) meaning that when one of them rises a raise of the other is very likely to happen. The same behaviour is observed when studying the relationship between the brent oil and the crude oil (both variables are found together and far away from the centre of the loading plot of Figure 7).

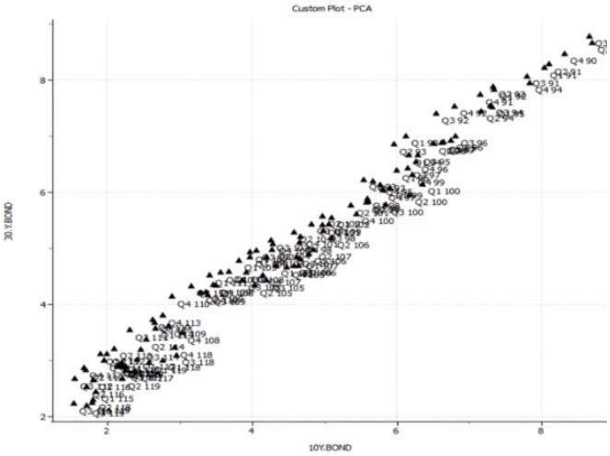

Figure 11. Scatterplot - 30-year bond & 10-year bond

In Figure 12, a scatterplot of the price of copper and the price of silver is shown. A clear positive correlation can be observed between the silver and copper prices (note that both variables are found together and far away from the center of the loading plot of Figure 7). Notice that this relationship appears to be more volatile along the years.

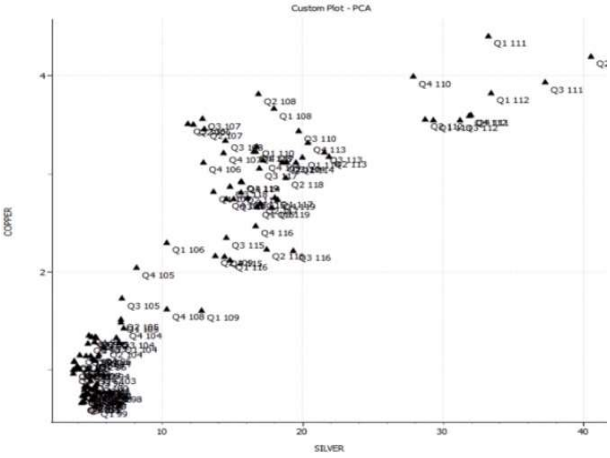

Figure 12. Scatterplot copper- silver

In Figure 13 a strong relation between Nasdaq & Dow Jones, the two main indicators in United States, can be observed. Note that both variables are found together and far away from the center of the loading plot of Figure 7.

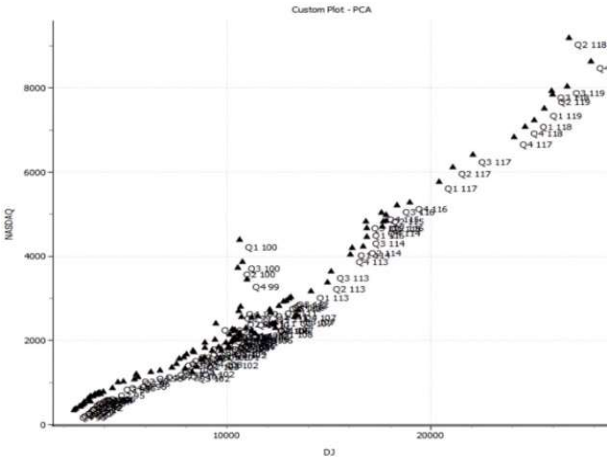

Figure 13. Scatterplot Nasdaq & Dow Jones Industrial

In Figure 14, a positive correlation can be also observed between the Nasdaq technological index and the USD/INR (both variables are found together and far away from the center of the loading plot of Figure 7).

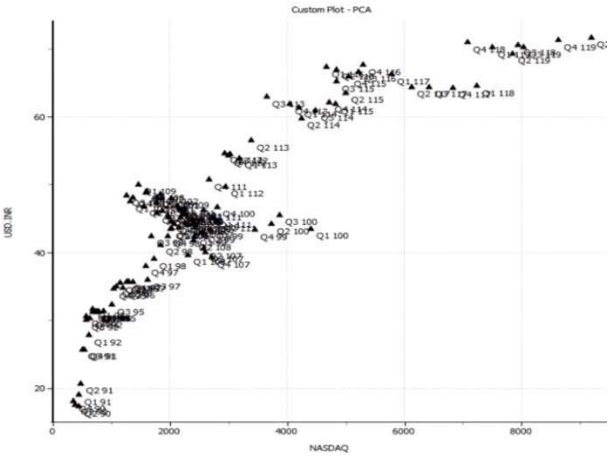

Figure 14. Scatterplot USD/INR & Nasdaq

Figure 15 shows a negative correlation between the 10 year bond and the Nasdaq (this is coherence with the antipodal position of both variables in the loading plot of Figure 7).

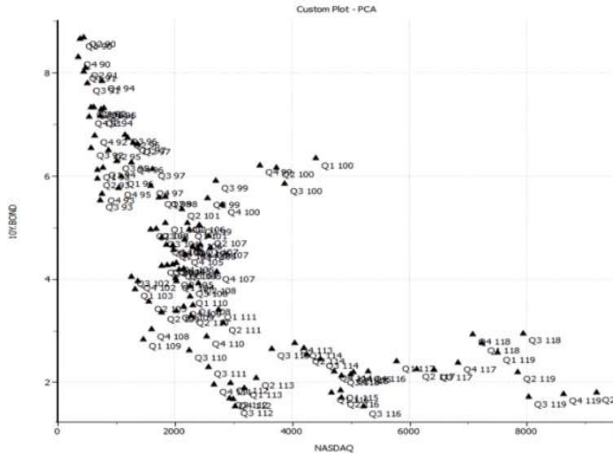

Figure 15. Scatterplot 10 years bond - Nasdaq

Once the main relationships between financial market indicators have been identified, it is important to study how these sources affect the distribution of individuals. For this, the scores associated with the two first principal components are represented.

From the sources of variability identified, we can see how the indices have evolved along time since 1990 (Figure 16 and 17). The scores are distributed cyclically in the same way in which the economy has evolved: economic expansion from the 90s to 2007, recession due to the Subprime Crisis, and economic recovery. By coloring the score plot with the value of one index, it can be observed how this index varies along the distribution of the observations. Figures 16 and 17 show this evolution for Nasdaq and crude oil prices.

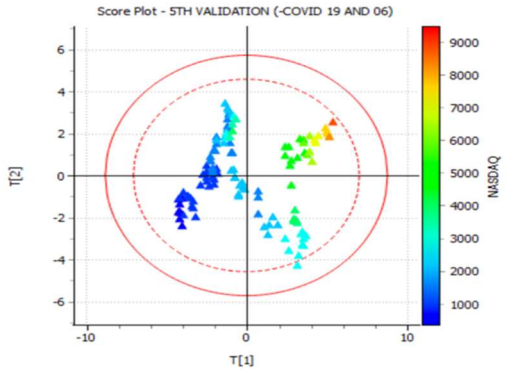

Figure 16. Nasdaq variability reflected on the t_1/t_2 score plot.

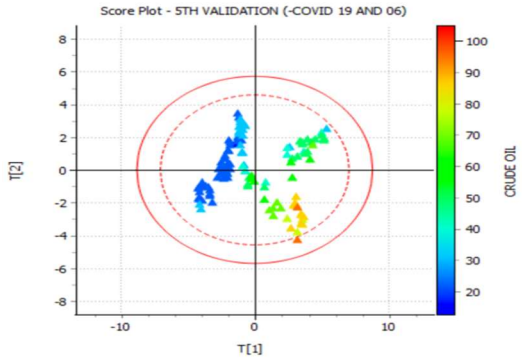

Figure 17. Crude oil variability reflected on the t_1/t_2 score plot.

As we can observed in Figure 7, the Nasdaq and the crude oil are not correlated. We can observe that the Nasdaq has evolved in accordance to the macroeconomic cycles, while the crude oil is naturally more volatile and does not fit perfectly to economic cycles, presenting high values during economic expansions and low values during economic recessions. As crude oil prices are more volatile it is difficult to make an exact assumption about it. Nevertheless, we can observe that during the crisis of

the subprime the oil prices have been at its highest (as we can see in Figure 17, red points mean highest values).

We will now follow our analysis by analysing the relations between the macroeconomic indices: the inflation rate, the Gross domestic Product and the unemployment rate.

3.2 ARE MACROECONOMIC INDICATORS CORRELATED?

Once the relationships between all financial indicators of the X space have been extensively studied, in this section the main ties between the macroeconomic indicators will be evaluated. For this purpose, a PCA is carried out. In the first place, the constructed model must be validated according to the SPE chart (Figure 18) (detection of rare individuals) and Hotelling's T^2 chart (Figure 19) (detection of extreme individuals) obtained after extracting 7 principal components with a goodness of fit (R^2) and goodness of prediction (Q^2) close to 80% (Figure 20).

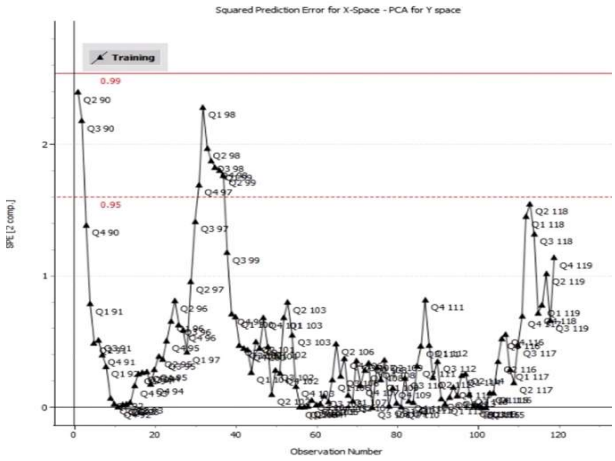

Figure 18. SPE for macroeconomic indicators

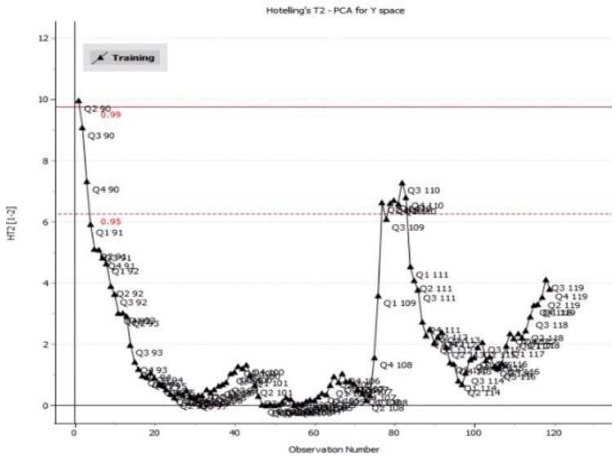

Figure 19. Hotelling´s T^2 for macroeconomic indicators

Considering that the database has 119 quarters, and none exceeds the 99% confidence limit and, given that it is expected that on average 1% of the observations will be outside these limits, the PCA model can be validated.

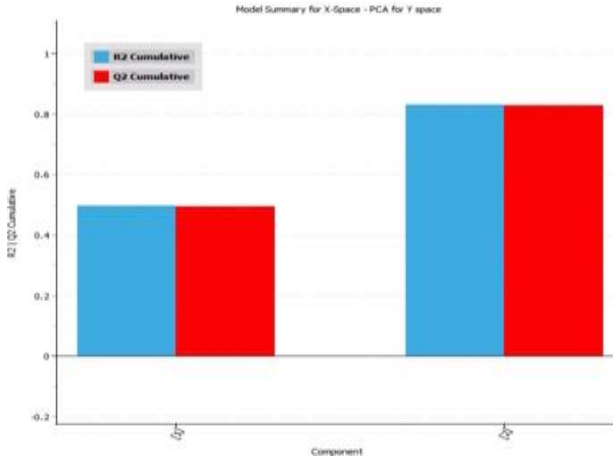

Figure 20, Model summary for macroeconomic indicators

Figure 20 shows that the two first principal components explain almost 80% of the variability. Given that the objective of this PCA is not predictive but exploratory, trying to improve the general understanding of the data, we present in the following the results in terms of the two first components.

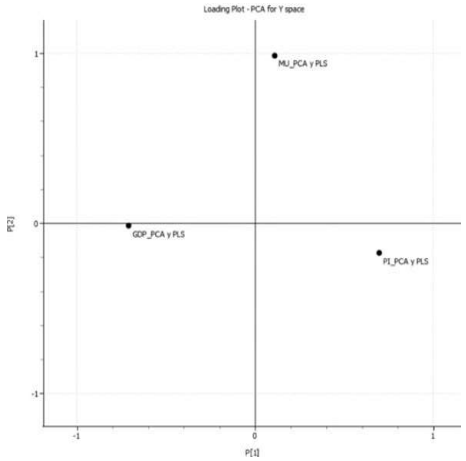

Figure 21. Loading plot for macroeconomic indicators

Figure 21 shows the p_1/p_2 loading plot revealing the relationships between the macroeconomic indicators. It is very interesting to observe that GDP is negatively correlated with the inflation rate (π), but that unemployment seems not to be correlated with the other macroeconomic indices.

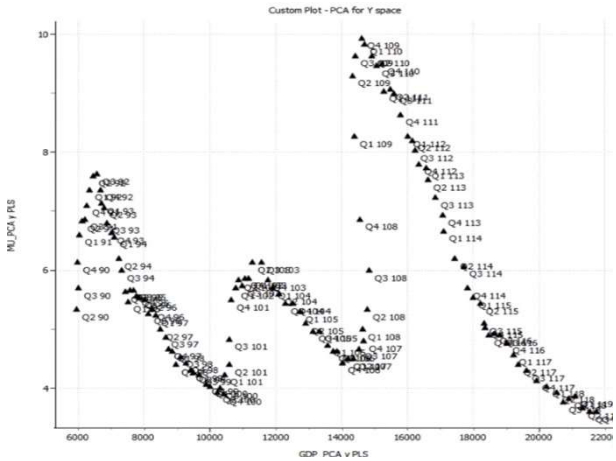

Figure 22. Scatterplot unemployment & GDP

Based on the loading plot (Figure 21), there does not appear to be some correlation between inflation and the unemployment rate. The last relationship deserves its explanation (see Figure 22). When the economy is expanding, unemployment rate shrinks and when the GDP goes into recession, they stagnate and unemployment rates skyrocket. Although the PCA has not been able to detect the relationship between these two variables, a very particular relationship can be drawn between these two indicators.

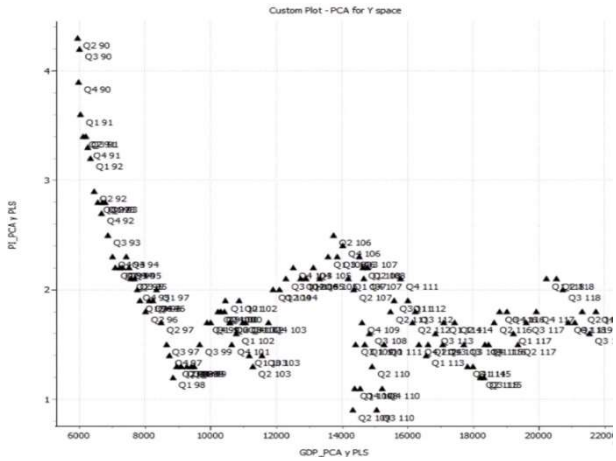

Figure 23. Scatterplot inflation rate & GDP

As it has been detected in the loading plot (Figure 21), there is a negative relationship between GDP and inflation for low GDP values. For higher values this relation seems to be compromised, as seen in Figure 23.

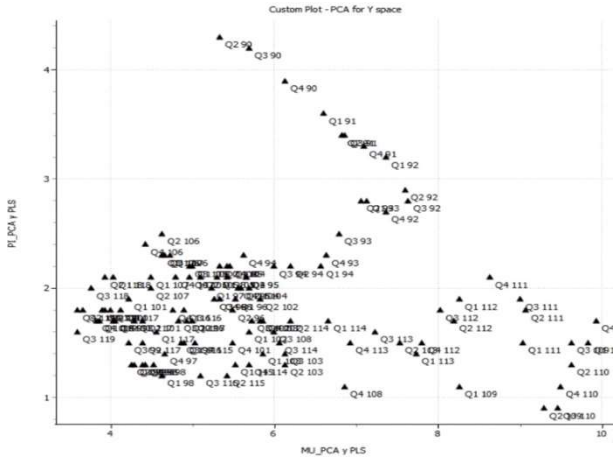

Figure 24, Scatterplot inflation rate & unemployment rate

The Phillips curve states that inflation and unemployment have an inverse relationship. Higher inflation is associated with lower unemployment and vice versa. The Phillips curve was a concept used to guide macroeconomic policy in the 20th century, but was called into question by the stagflation of the 1970's. This is a rather interesting interpretation since it has been known that the Phillips curve has not worked perfectly for a long time, and in the period studied it seems that this inverse structure between unemployment and inflation rates has been broken for most recent years (Figure 24).

To deepen the understanding of the relationships between these macroeconomic indicators, the scores associated with the main components 1 and 2 are represented. This way, it is possible to identify how the indicators have evolved throughout the years since 1990. The scores are distributed cyclically in the same way in which the economy has evolved. Process of economic expansion from the 90s to 2007, the process of recession due to the Subprime Crisis, and economic recovery. In Figure 25, we can corroborate this statement. We can observe that during the first period mentioned above the gross domestic product has raised significantly, and then due to the subprime crisis a significant shrinking of the gross domestic product can be observed. Until then, its value has raised.

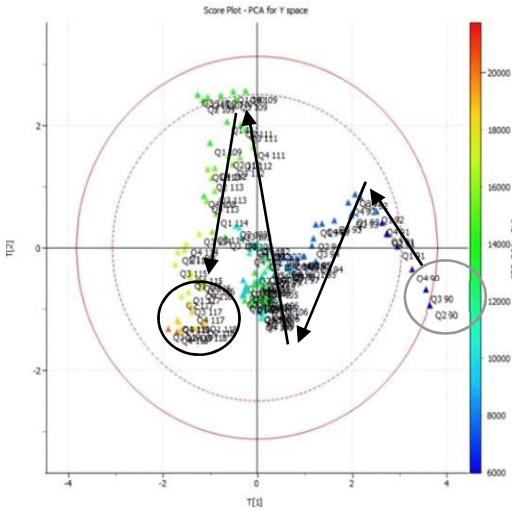

Figure 25. GDP variability on scores

Once both spaces have been studied and explored, the objective of this work is to determine if it is possible to forecast the main macroeconomic indicators from the financial markets' information. For this, as a pre-step of the overall analysis, a PCR is going to be performed and if the markets and the indicators are correlated the regression model should explain a good amount of the variability of each indicator.

3.3 DETECTION OF THE EXISTENCE OF CORRELATION BETWEEN THE FINANCIAL MARKETS AND THE MACROECONOMIC INDICATORS

At first, using the scores obtained with the Principal Component Analysis of the financial markets, we will perform a Principal Component Regression (PCR) analysis to investigate if an existing correlation between macroeconomic indicator and financial indices can be found. Secondly, as complementary analysis, we will perform a Random Forest to determine the most meaningful variables for each indicator. This is done because some of the models we will build later cannot use the information of the 19 variables. For this reason, the Random Forest would permit an identification of the most relevant variables for each indicator. Once identified the relationship between the financial markets and macroeconomic indicators we will study those relationships thanks to the Partial Least Squares method. Finally, as the data we are working with are serially correlated with time, we will use the dynamic version of PLS to take into account the dynamics nature of the time series data.

3.3.1 Identification of a relationship between financial markets and macroeconomic indicators, a principal component regression approach.

As explained in the methodology, from the scores matrix of the Principal Component Analysis fitted on the financial markets indicators, a regression model to predict the macroeconomic indicators will be fitted solving the main problem due to multicollinearity as the scores are independent to one another.

Figure 26 shows the results of the PCR, using the stepwise approach, to predict the Gross Domestic Product (GDP). The regression model has a high goodness of fit (R^2= 98.28%). This prove that an actual correlation exists between financial markets and the Gross Domestic Product.

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	11071,9	50,5481	219,036	0,0000
Comp.1	1225,85	16,6406	73,6663	0,0000
Comp.2	441,142	25,9981	16,9682	0,0000
Comp.3	148,955	39,1603	3,80373	0,0002
Comp.4	188,032	48,2756	3,89498	0,0002
Comp.5	-204,198	84,7254	-2,41011	0,0177
Comp.6	-967,269	96,7691	-9,99564	0,0000
Comp.8	-747,288	138,287	-5,4039	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1,66744E9	7	2,38206E8	839,89	0,0000
Residual	2,92125E7	103	283617,		
Total (Corr.)	1,69665E9	110			

R-squared = 98,2782 percent
R-squared (adjusted for d.f.) = 98,1612 percent
Standard Error of Est. = 532,557
Mean absolute error = 422,961
Durbin-Watson statistic = 0,42226 (P=0,0000)
Lag 1 residual autocorrelation = 0,786227

Figure 26. PCR model for GDP

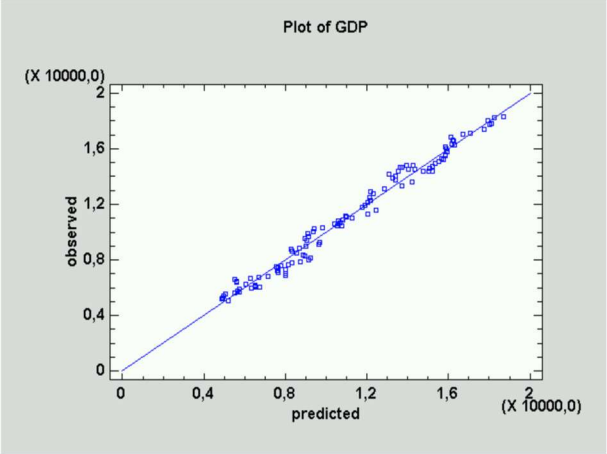

Figure 27. Observed vs predicted plot for GDP

By following a similar procedure for the two other macroeconomic indicators, we obtain PCR models with R^2= 83.5% for the unemployment rate, and R^2=61.2% for the inflation rate. This shows that financial markets contain useful information to predict macroeconomic indicators, especially GDP and unemployment rate.

In conclusion, as it has been seen from the PCR, there is a strong statistical evidence between financial markets and the main economic indicators. This is the reason why this analysis is going to be followed trying to make the most advantage of the relations existing between both spaces. To do this, a PLS methodology will be performed, and therefore a dynamic consideration of transfer functions and exogenous variables are going to be modelled with PLS. Therefore, we will use the Random Forest to identify what variables might fit best for our predicitve analysis.

3.3.2 Random forest – meaningful variables forecasting gross domestic product, inflation rate and unemployment rate

The random forest will be used as an complementary methods to select most meaningful varaibles in our analysis. We will later contrast those selected variables with the variables selected in the Partial Least Squares methods to combine and use most significant financial indices to forecast each macroeconomic indicator.

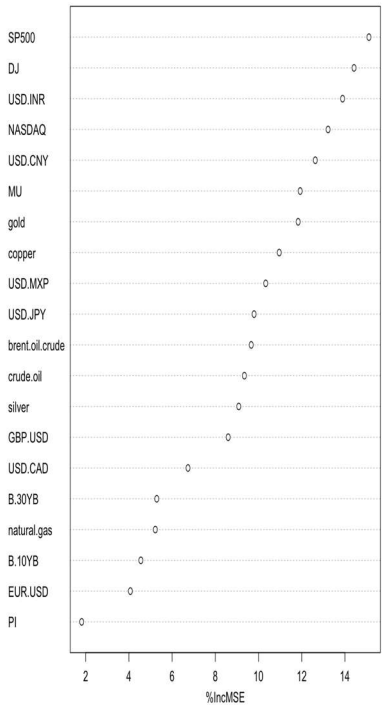

Figure 28. Random Forest for Gross Domestic Product

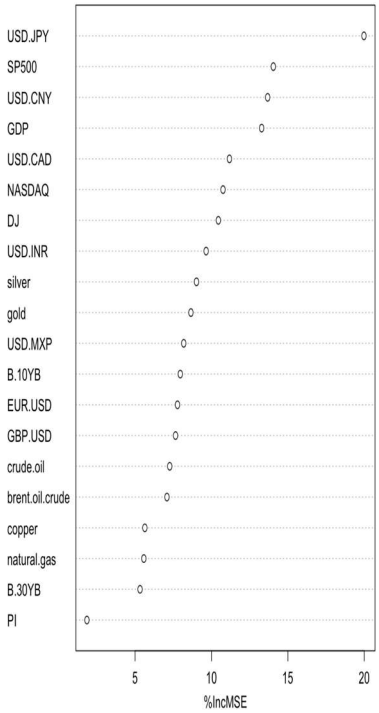

Figure 29. Random Forest for Unemployment Rate

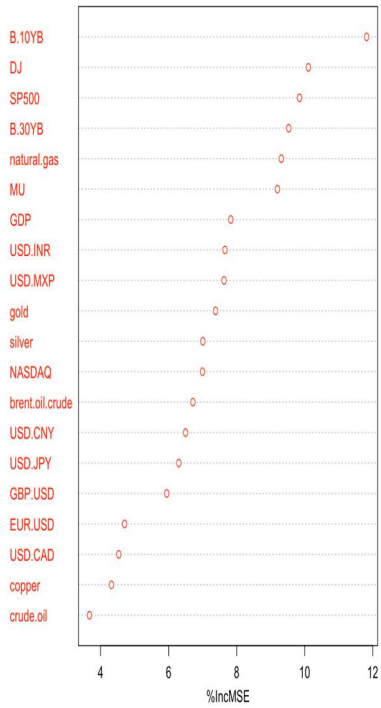

Figure 30. Random Forest for Inflation Rate

In Figure 28, we can observe what are the most important variables to forecast the gross domestic product in RF. In this case, the SP500, the industrial Dow Jones and the USD/INR seems to retrieved information that would be relevant to use for the gross domestic prediction. In this model, a 99.58% of the total variability has been explained. Similarly in Figure 29, we can observe that the best financial markets to use for predicting the unemployment rate in RF are the USD/JPY, the SP500 and the USD/CNY. In this model, a 92% of the variability has been explained. For the inflation rate (Figure 30), the explained variability of the RF only reaches a 35.42%. Even though result are not satisfiying for the indicators, most relevant financial markets seems to be the 10 year bond, the SP500 and the industrial Dow Jones.

3.3.3 Determining the most meaningful relationships between financial markets and macroeconomics indicators.

To finally jointly explore both spaces through latent variables, the PLS will be used. Figure 31 shows the overall goodness of fit and goodness of prediction for the PLS model with 6 components PLS when predicting the three macroeconomic indicators. From the fourth component onwards the increase in the performance indexes is negligible.

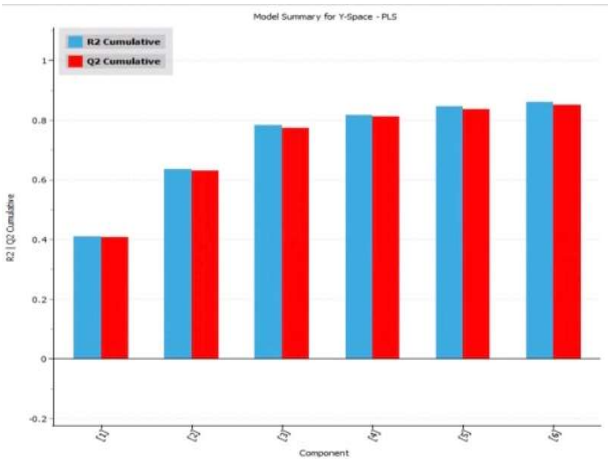

Figure 31. Model summary for PLS model

Figure 32 shows the goodness of fit of each one of the six PLS components when predicting the three macroeconomic indicators. GDP is almost fully explained by the first PLS component, Unemployment rate seems to be largely explained by the second component and a combination of the rest, Inflation, as seen above, is more difficult to predict, and it is a combination of mainly the first 3 PLS components.

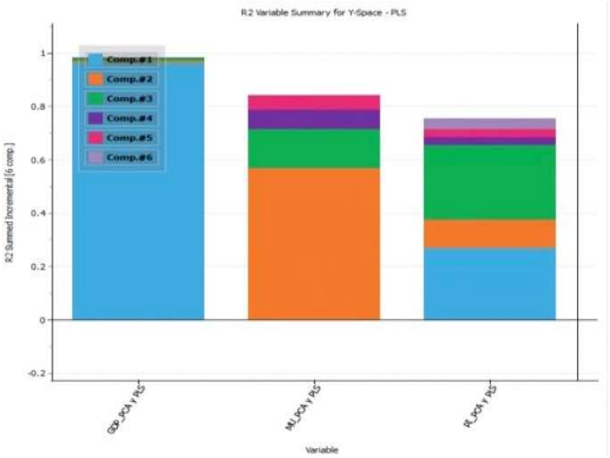

Figure 32, R^2 summary for macroeconomic indicators (Y space)

Figure 33 shows the goodness of fit of each financial market indicator for each one of the six PLS components. As can be seen, stock indices such as the Dow Jones, S & P500, NASDAQ and bonds are mainly explained by the first component. The second component mainly explains some Forex pairs, and the third component already provides less information but mainly explains natural gas and provides information on some Forex pairs.

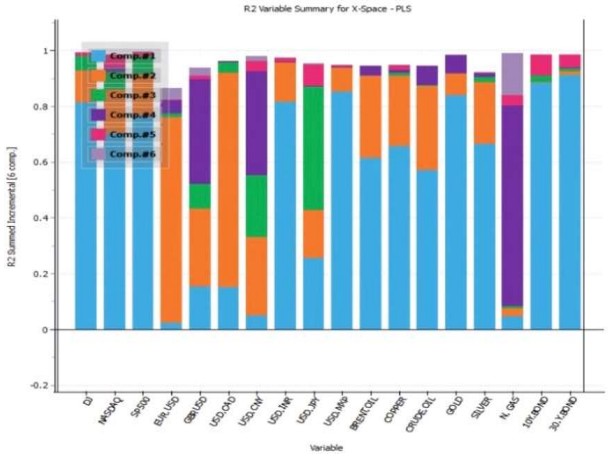

Figure 33. R^2 summary for financial markets indicators (X space)

Based on the previous results only 3 components will be retained. The PLS model is validated following the same procedure pursued in PCA. As can be seen in the SPE plot (Figure 34), three observations exceed the 99% confidence limit, given that the value of these observations does not exceed twice the limit, although slightly anomalous, these observations will be kept in the model.

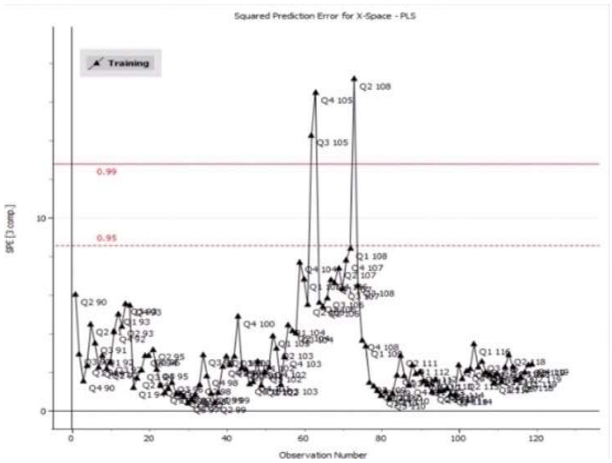

Figure 34. SPE for PLS model

Based on Hotelling's T^2 plot (Figure 35), there is no evidence or appearance of extreme observations.

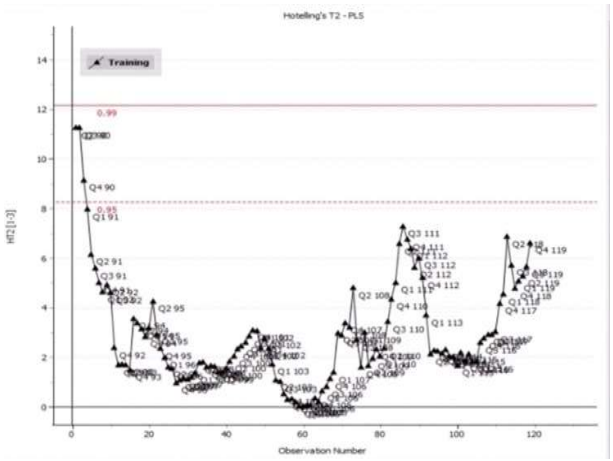

Figure 35. Hotelling´s T^2 for PLS model

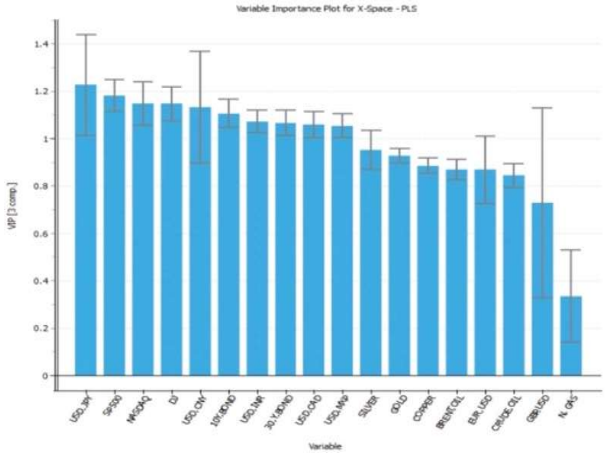

Figure 36. VIP for financial markets indicators

A relative measure of the importance of the predictor variables is VIP (Figure 36). This shows which are the variables that do have a real impact on the prediction and those that do not. To take advantage of the VIP, the analyst must pursue keeping the variables whose VIP value is greater than 1, and delete those that have a VIP lower than 0.5. Thus, the rest of the variables should be subject to study their validity and contribution within the PLS model.

Furthermore, the relationships between the two spaces are observed in the weighting scatterplot of w_1* c_1 / w_2* c_2 (Figure 37).

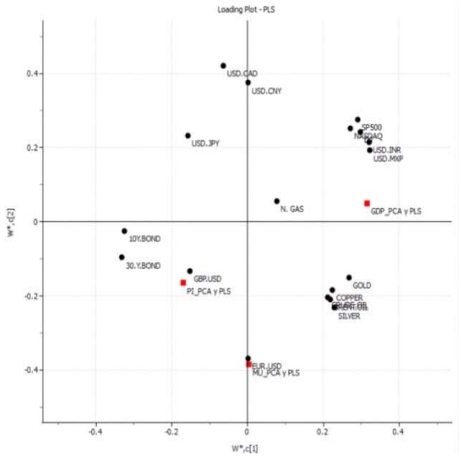

Figure 37. Weighting scatterplot for the two first PLS components

As already revealed in the PCA, the same relationships are observed again within each space in the PLS model. However, when the direct relationships observed are analysed, GDP is positively correlated to major stock indices such as the Dow Jones, the NASDAQ and the S&P500, and negatively correlated with 10- and 30-year bonds. Regarding unemployment, the clearest positive correlation is found with the EUR / USD pair, while it is negatively correlated with the USD / CAD and USD / CNY pair. Regarding the inflation rate, the most relevant correlation seems to be with the GBP / USD, and clear negative correlation can be detected between the inflation rate and the SP500, the Nasdaq and the forex pairs (USD / INR and USD / MXP).

These relationships are illustrated in Figures 38, 39, 40, 41 and 42. This way, it can be confirmed that there is a relationship between the different indicators mentioned above.

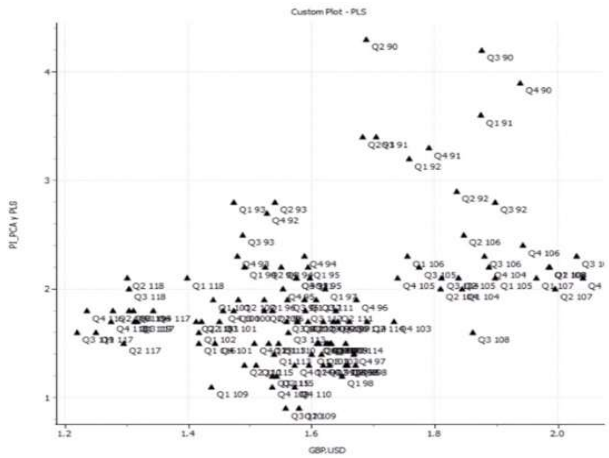

Figure 38. Scatterplot Inflation Rate & GBP/USD

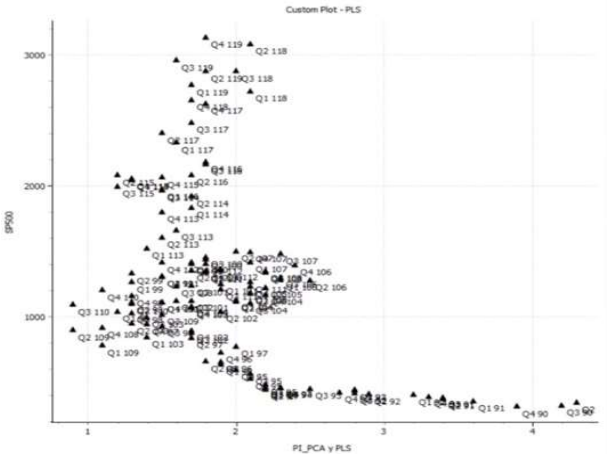

Figure 39. Scatterplot S&P500 & Inflation Rate

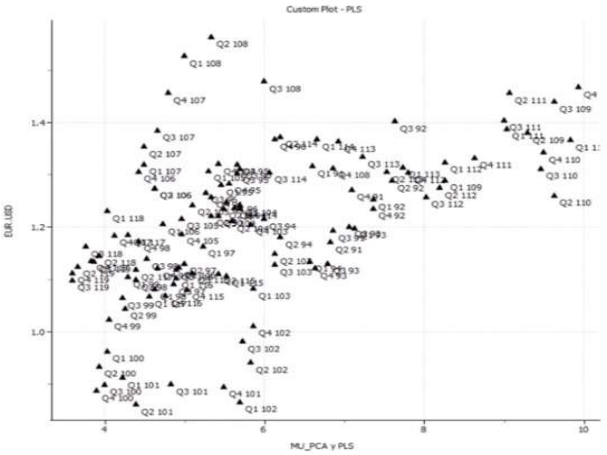

Figure 40. Scatterplot EUR/USD & Unemployment Rate

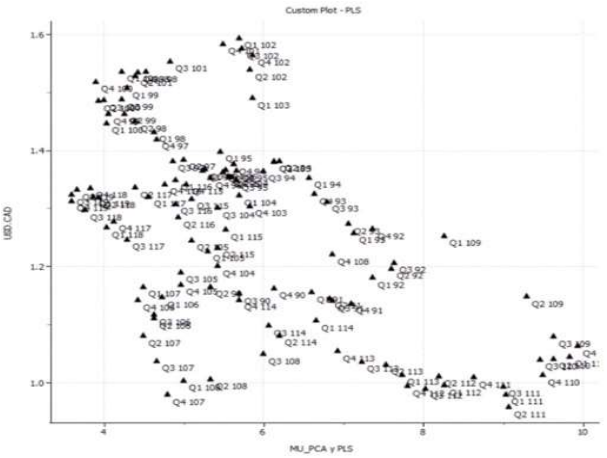

Figure 41. Scatterplot USD/CAD & Unemployment Rate

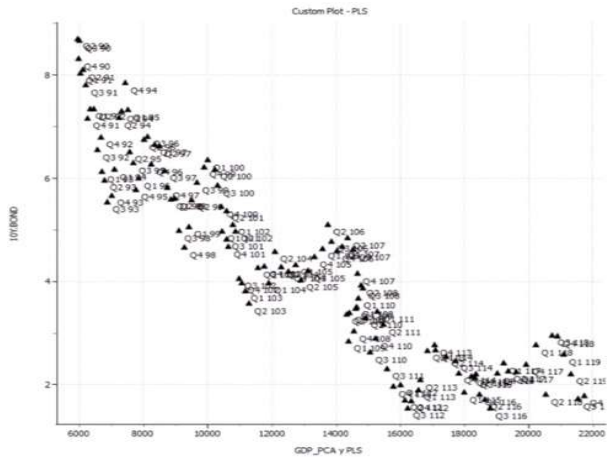

Figure 42. Scatterplot 10 years bond & GDP

Figures 43, 44 and 45 show the regression coefficients corresponding to each of the financial markets indicators in prediction each one of the macroeconomic indicators, GDP, IR and UR, respectively. 95% Jackniffe confidence intervals for regression coefficients are also shown (see the forks). Those regressors with forks including the zero value should be considered non-statistically significant.

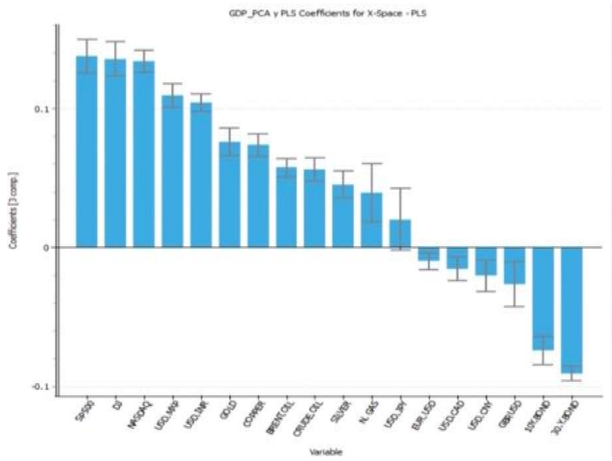

Figure 43. PLS regression coefficients for GDP

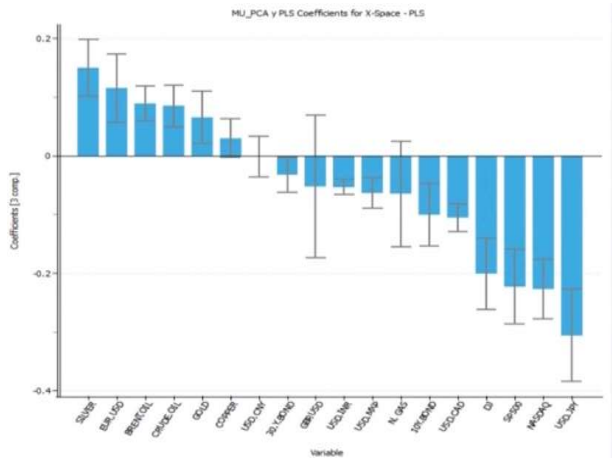

Figure 44. PLS regression coefficients for Inflation Rate

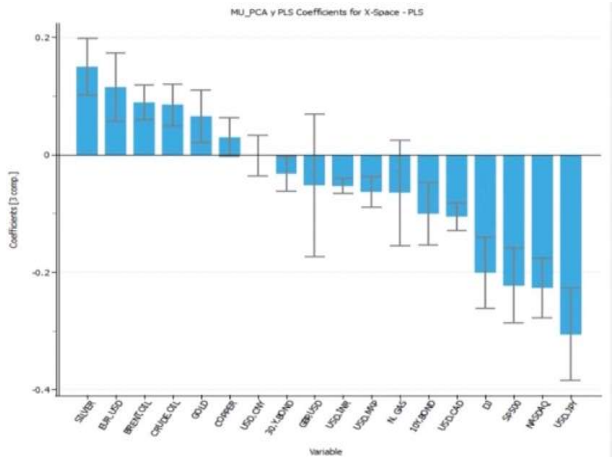

Figure 45. PLS regression coefficients for Unemployment Rate

Figures 46, 47 and 48 show the observed and fitted PLS time series values for each one of the macroeconomic indicators, GDP, IR and UR, respectively. 95% confidence intervals for the predictions are also shown. Although it is possible to observe some deviations in general, the fit of the PLS model is quite good even for the Inflation Rate.

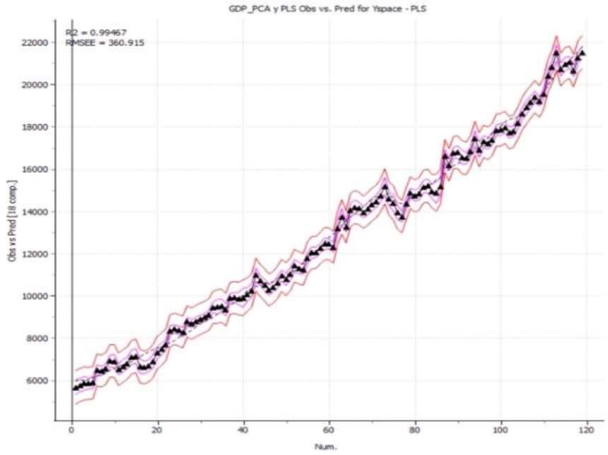

Figure 46. Observed and PLS fitted time series values for GDP

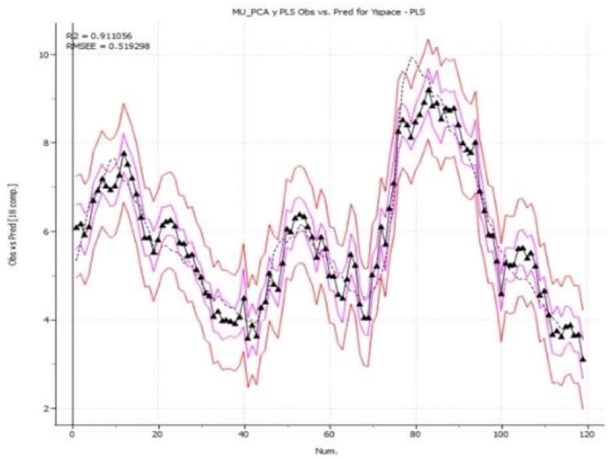

Figure 47. Observed and fitted PLS time series values for Inflation Rate

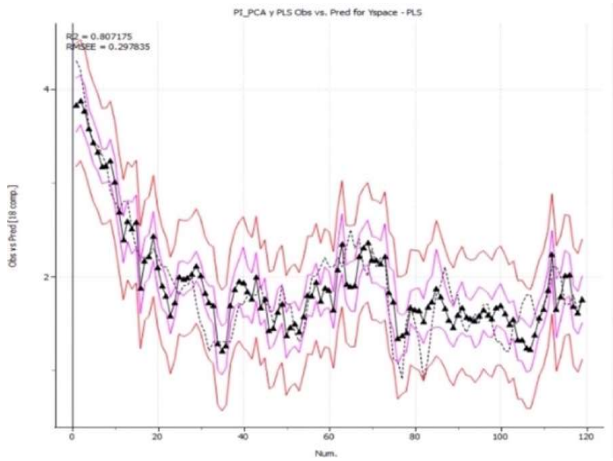

Figure 48. Observed and fitted PLS time series values for Unemployment Rate

Obviously PLS is a strong analytical method, but it can be improved to make even better forecasts and interpretation. This is a reason why, a transfer function with exogenous variables is going to be modelled with PLS allowing the modelling of the dynamics of the time series dataset.

3.3.4 Further study of the relationship between macroeconomic indicators and financial markets, a dynamic consideration

As it has been observed previously, the PLS offers a very good interpretation of the database and can relate both spaces efficiently. However, the problem that is being dealt with is one of time series, therefore, modelling a transfer function with exogenous variables with PLS is an efficient way to model the dynamics of the time series database. For this, it is proposed to include as regressors all the financial markets indicators and their corresponding lagged variables delayed until 12 times, and the lagged variables delayed till 12 times for the three macroeconomic indicators. This way, it is assumed that there might be an influence of some variables on those at present up to a 3-year range.

Next, the PLS model is validated in the same exact way as done before in PCA. As can be seen in the SPE graph (Figure 49), there is no evidence or appearance of anomalous data. In the Hotelling's T^2 plot (Figure 50) two observations exceed the limit, considering the number of observations and that the value of these observations does not exceed twice the limit, these observations will be kept in the analysis.

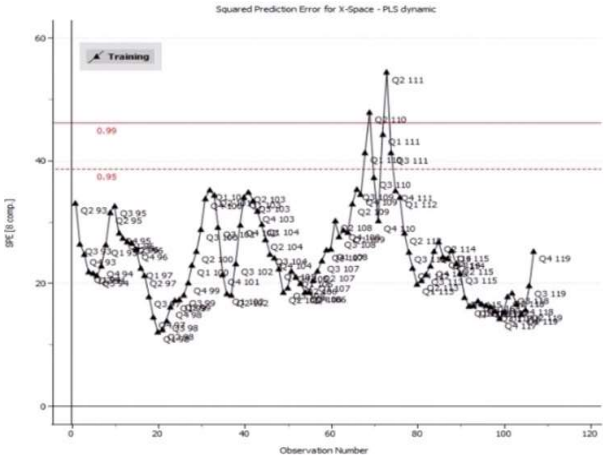

Figure 49. SPE plot for dynamic PLS

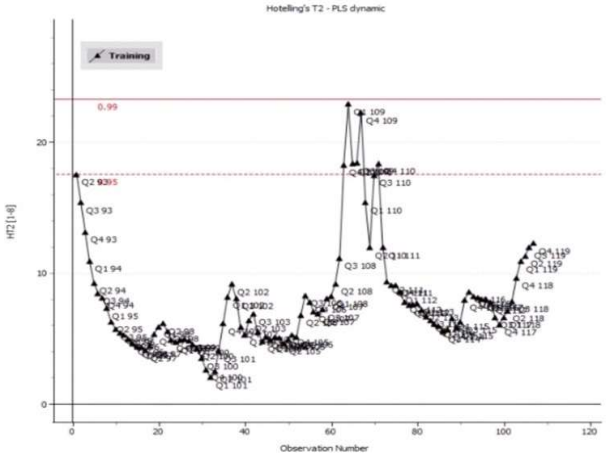

Figure 50. Hotelling´s T^2 plot for dynamic PLS

Considering that the database has 119 quarters and only two exceed the 99% confidence limit and, given that it is expected that on average 1% of the observations will be outside these limits, the PLS model can be validated. Figure 51 shows that eight PLS components explain more than 90% of the variability of the response variables (macroeconomic indicators). Nevertheless, a four-component model has an R^2 greater than 80% and is much more parsimonious.

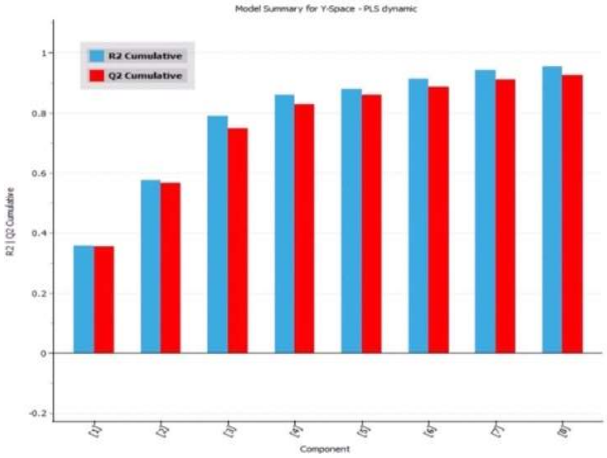

Figure 51. Model summary for dynamic PLS

By having a much larger number of predictor variables, it is difficult to make an exhaustive analysis of which variables are explained by one or more components. However, in general, the conclusions match those obtained previously in the non-dynamic PLS. Regarding the lagged variables, it is very interesting to note that the lagged GDP variables are mainly explained by the first component (the component that explain GDP, see Figure 53). For the lagged Unemployment Rate variables, they are mainly explained by the second component (the component that explain UR, see Figure 53.) Regarding the Inflation Rate, the explained variability varies along the magnitude of the lag (Figure 52).

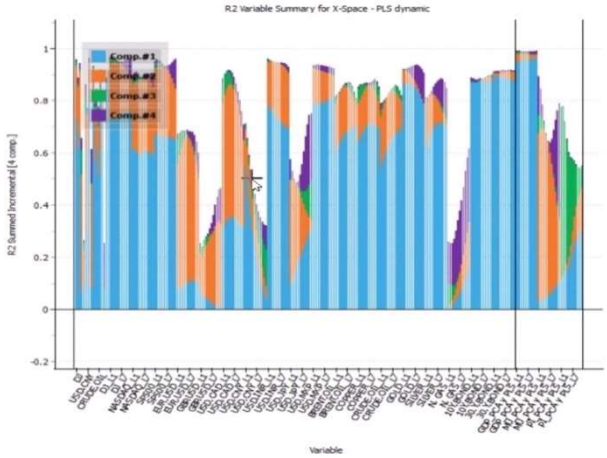

Figure 52. Explained variability for financial markets with dynamic PLS

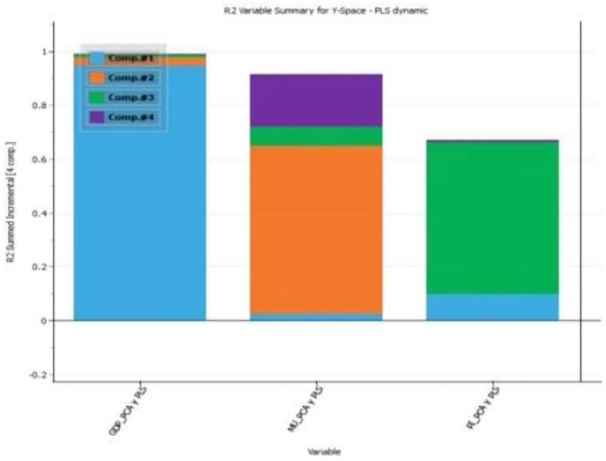

Figure 53. Explained variability for macroeconomic indicators with dynamic PLS

At the level of space Y (see Figure 53), GDP is almost fully explained by the first component. In turn, Unemployment Rate seems to be largely explained by the second component and a combination of the rest. Inflation Rate, as seen above, is less explained than the other two macroeconomic indicators, and is basically explained by the third PLS component.

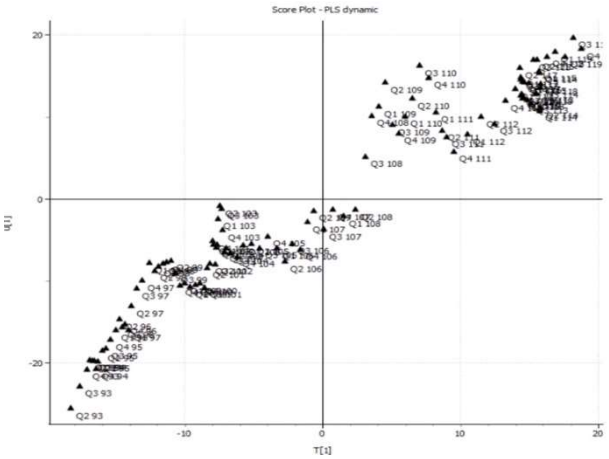

Figure 54. Inner relationship of the dynamic PLS model.

Before being able to analyse the PLS model, it must be ensured that the inner relationship between the u and t scores is linear, since otherwise, transformations and/or the inclusion of interactions in the model should be considered. As can be seen in Figure 54, the relationship is linear enough to be able to continue with the study of the model.

The direct relationships that can be found from the weighting scatterplot $w_1^* c_1$ / $w_2^* c_2$ graph are discussed below.

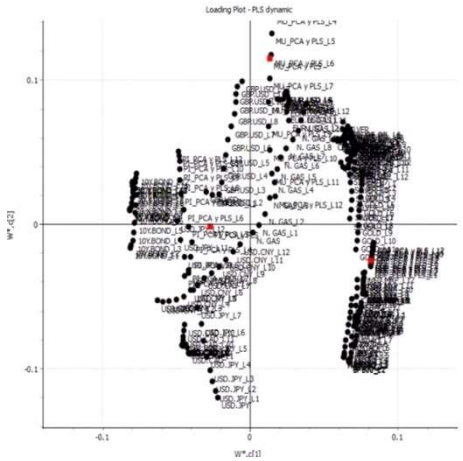

Figure 55. Weighting plot for dynamic PLS model.

As studied in PCA, the same relationships are observed again within each space. However, when the direct relationships observed are analysed, GDP is positively correlated to major stock indices such as the Dow Jones, the NASDAQ and the S & P500. It is also negatively correlated with 10- and 30-year bonds. Regarding Unemployment Rate, the clearest positive correlation is found with the EUR / USD pair, being also negatively correlated with the USD / CAD and USD / CNY pair. Furthermore, it can be seen how the response and lagged variables are correlated with each other and as expected, how this relationship declines as the magnitude of the lag increases.

The relationship between each macroeconomic indicator and its lagged variables is very different depending on the macroeconomic index that is being studied. The following figures show scatterplots between each response variable and its exogenous variables shifted 3 years (12 periods). As can be expected this relationship is strong for GDP (Figure 56), mild for the Unemployment Rate (Figure 57), and weak for the Inflation Rate (figure 59).

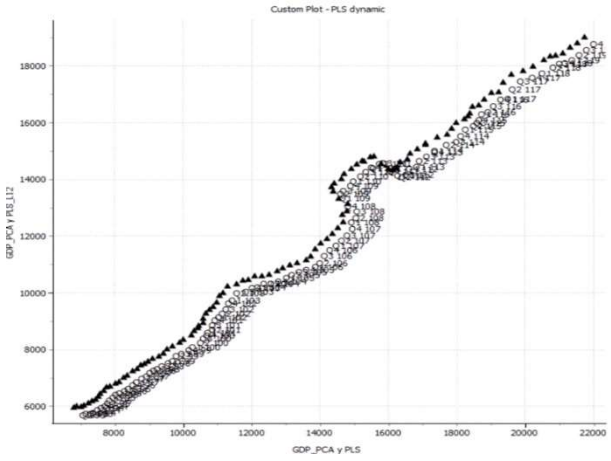

Figure 56. Scatterplot between GDP_t and GDP_{t-12}

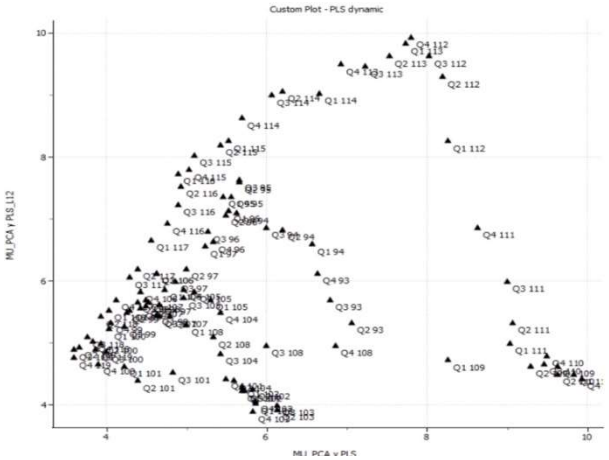

Figure 57. Scatterplot between Unemployment Rate $_{t}$ and Unemployment Rate $_{t-12}$

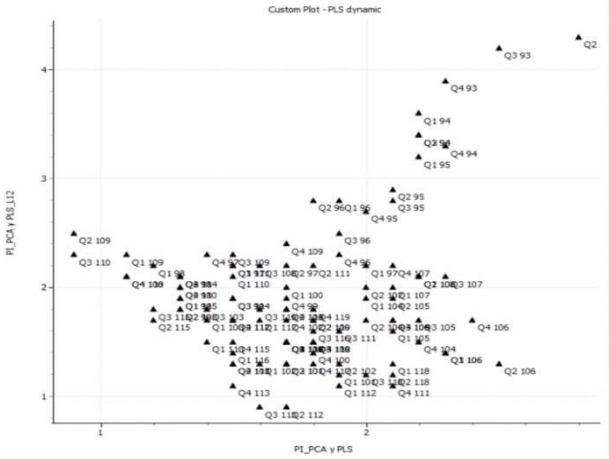

Figure 58. Scatterplot between Inflation Rate $_{t}$ and Inflation Rate $_{t-12}$

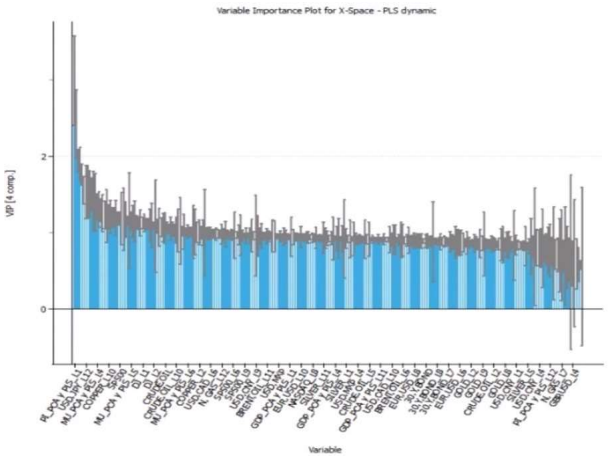

Figure 59. VIP for dynamic PLS model

Figure 59 shows the VIP for the regressors. There are several variables with VIP > 1, indicating a high importance in the dynamic PLS model.

As already mentioned, in order to reduce the prediction uncertainty, the predictor variables that are not statistically significant in the model for each of the macroeconomic indicators should be eliminated.

Figure 60 shows the PLS regression coefficients (with their associated 95% Jackniffe confidence intervals) for predicting GDP. It can be observed that several of these are not statistically significant (its corresponding Jackniffe interval contains the zero value), therefore it is possible to delete those variables a fit a new pruned PLS model (Figure 61).

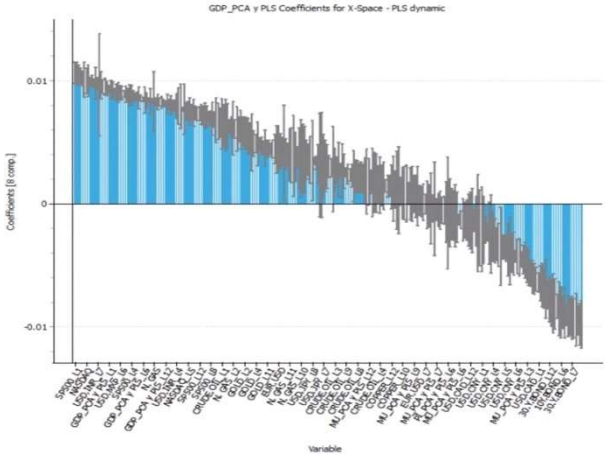

Figure 60. Regression coefficients for GDP with dynamic PLS model

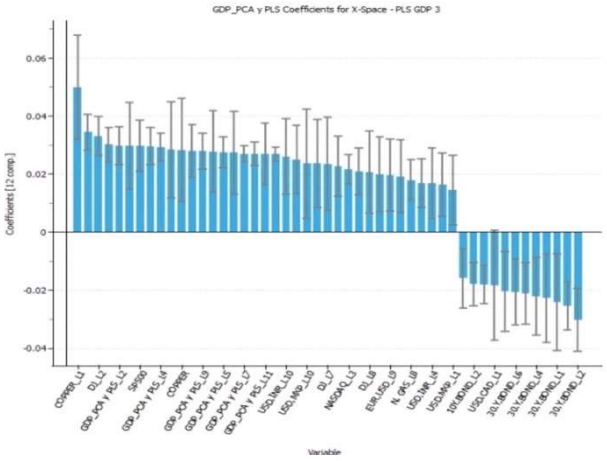

Figure 61 Regression coefficients for GDP with pruned dynamic PLS model

The same process has been repeated for the other two macroeconomic indicators aiming at obtaining a consistent model when forecasting the value of each response separately. Figure 62 and 63, show the regression coefficients for predicting the Unemployment Rate for the initial and pruned PLS models, respectively. Similar information is obtained in Figures 64 and 65 for predicting the Inflation Rate.

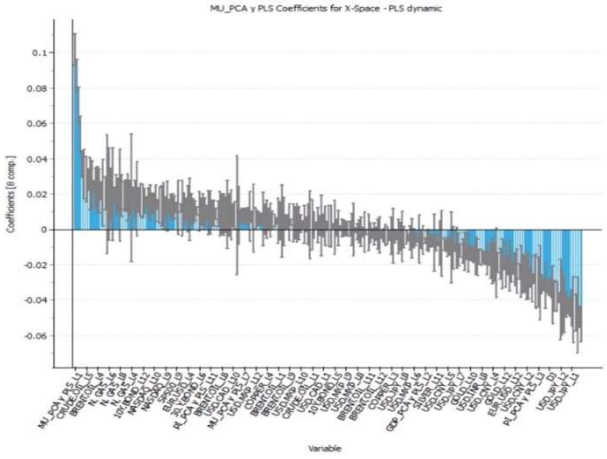

Figure 62. Regression coefficients for Unemployment Rate with dynamic PLS model

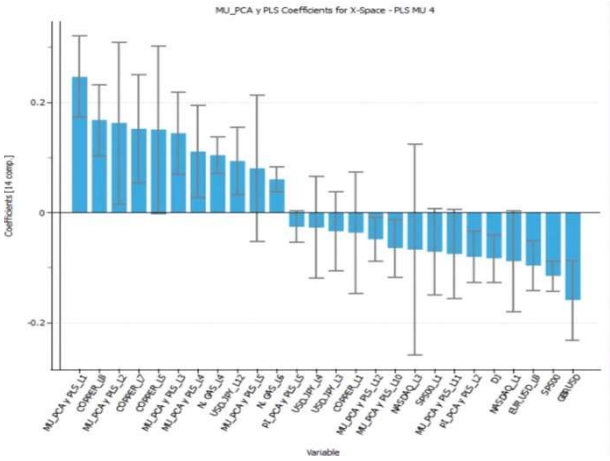

Figure 63. Regression coefficients for Unemployment Rate with pruned dynamic PLS model

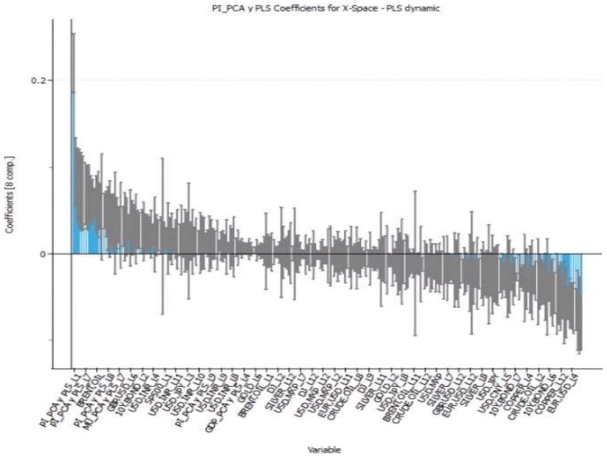

Figure 64. Regression coefficients for Inflation Rate with dynamic PLS model

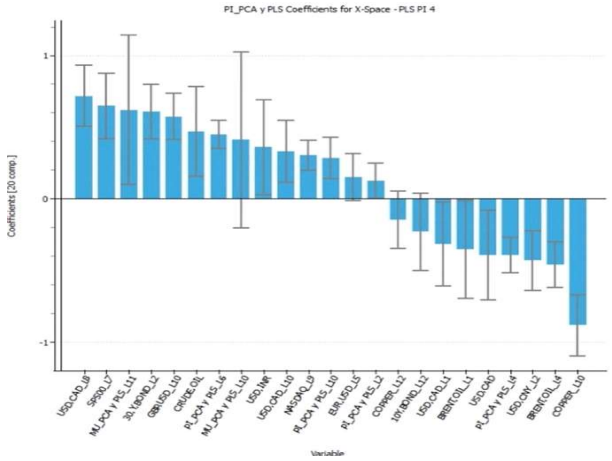

Figure 65. Regression coefficients for Inflation Rate with pruned dynamic PLS model

Figures 66, 67 and 68 show the observed and fitted PLS time series values for each one of the macroeconomic indicators, gross domestic product, the inflation rate and the unemployment rate, respectively, with its 95% confidence intervals for the predictions. By comparing these plots with respect to those shown in Figures 46, 47 and 48, it is clear how the inclusion of dynamics (by means of lagged variables) improves the predictions for the three macroeconomic indicators.

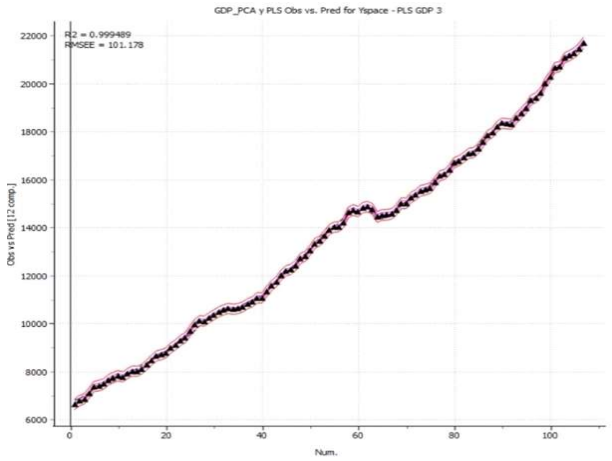

Figure 66. Observed and PLS fitted time series values for GDP with dynamic PLS

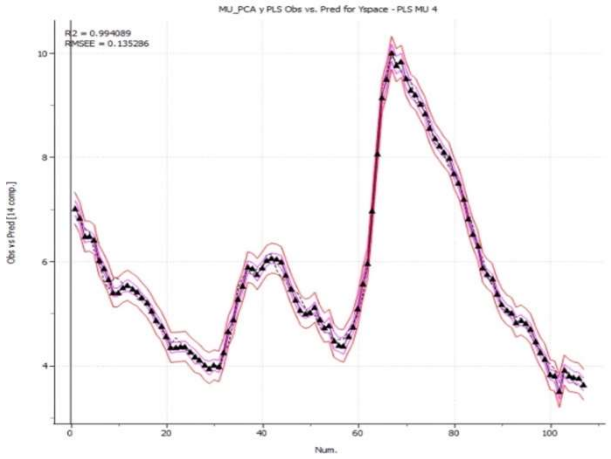

Figure 67. Observed and PLS fitted time series values for unemployment with dynamic PLS

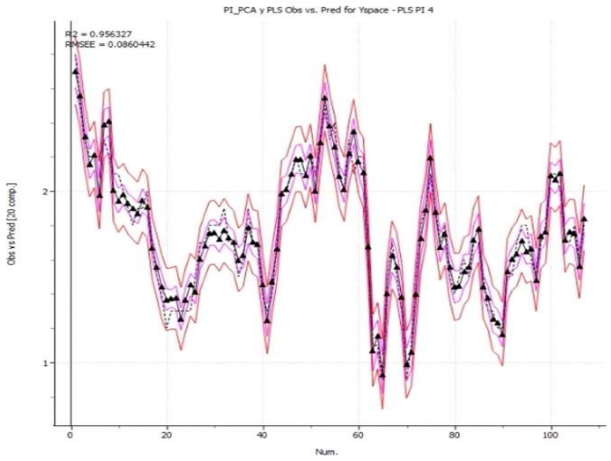

Figure 68. Observed and PLS fitted time series values for inflation rate with dynamic PLS

4. Forecasting the gross domestic product, the inflation, and the unemployment rate with statistical methods.

Once the connection between the two spaces has been studied and established, thanks to the PCA and the PLS models, the next stage of this work will be to take advantage of the existing correlations to predict macroeconomic indicators. For this, using the R software, and the libraries presented below, we will build and validate different statistical models as we did before with the dynamic Partial Least Squares model. To do this, we will first have to carry out a descriptive analysis of the series in question, to know the main limitations of the models that will be built next.

In following sections our analysis will be only shown for the inflation rate as the methodology followed is similar for the gross domestic product and the unemployment rate. Results of all models for each indicator will be shown in the last section of Chapter 4.

To compare the models performance, we will split the data in a **training set** (data used to build the model, 1990 to 2018 years) and a **validation set** (last year of our data, 2019). Once the prediction has been made for year 2019, we will calculate the MAPE between the validation set and the prediction of each model. This way we obtain a systematical and reliable comparative methodology.

4.1 DESCRIPTIVE ANALYSIS OF THE SERIES.

As shown in Figure 69 the Inflation Rate is quite volatile and has varied along the period studied. We will remember that the series of the Inflation Rate begins in the second quarter of the year 1990 until the last quarter of 2018, being the period analysed throughout this work.

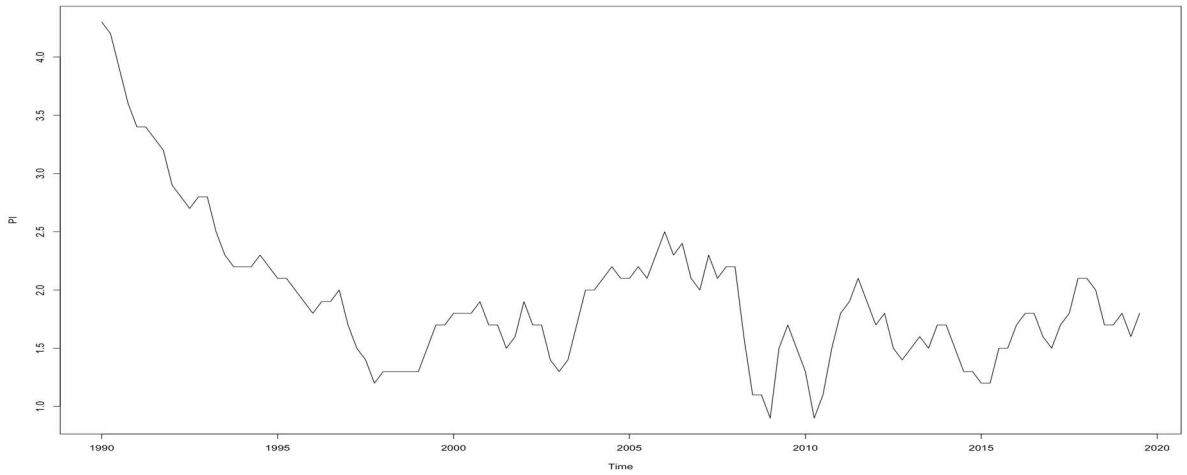

Figure 69. Inflation rate time series

The Inflation Rate time series has a negative trend from the 1990s to the 2000s. However, after this period the trend does not seem to be completely clear, and it presents a more pronounced random component. However, after the Subprime crisis (2008) we can observe that the inflation rate has augmented until 2018. About the variance, this is quite constant throughout the period analysed. At first glance, a clear seasonality in the series is not detected either. From this analysis, the different models described in the previous sections will be presented to predict the inflation rate for the coming year. For this, a cross-validation method will be used, in which a training set and a validation set will be used. This method will be useful to be able to compare how the models behave and thus compare their performance in terms of mean absolute percentage error (MAPE) on the validation set.

Note that the same procedure will be repeated for the Unemployment Rate and the Gross Domestic Product. Results we will be summarized at the end of the document.

4.2 ARIMA

To start our predictive analysis, we will focus on the ARIMA models. To build such model, we should first prevent that the series is stationary in trend and variance, and then, using the ACF and PACF, we would select the MA and AR order of the model. Once the ARIMA model has been chosen, we will care about its validation (white noise of residuals) before we can make a forecast. As said before, all the predictive analysis will be performed thanks to R project statistical software, and particularly for ARIMA models we will use the function auto.arima from the package “forecast”. This function straightforwardly selects the best order depending on the series analysed. Once obtained the tentative model (auto.arima function), the objective will be to ensure that the model can be validated and later, we will proceed to the prediction.

At first, we will check if the Inflation Rate series is stationary. To do it we perform the Augmented Dickey-Fuller Test using the function adf.test from the package "tseries".

Augmented Dickey-Fuller Test

Ho: Time series is not stationary
H1: Time series is stationary

Dickey-Fuller = -3.1316, Lag order = 4, p-value = 0.1069 > 0.05
alternative hypothesis: stationary

As the p-value is superior to the significance level (5%) we cannot reject the null hypothesis and we accept that the series is not stationary. To prevent it we will differentiate the series once to eliminate the trend effect. We obtain the following series for the Inflation Rate (Figure 70).

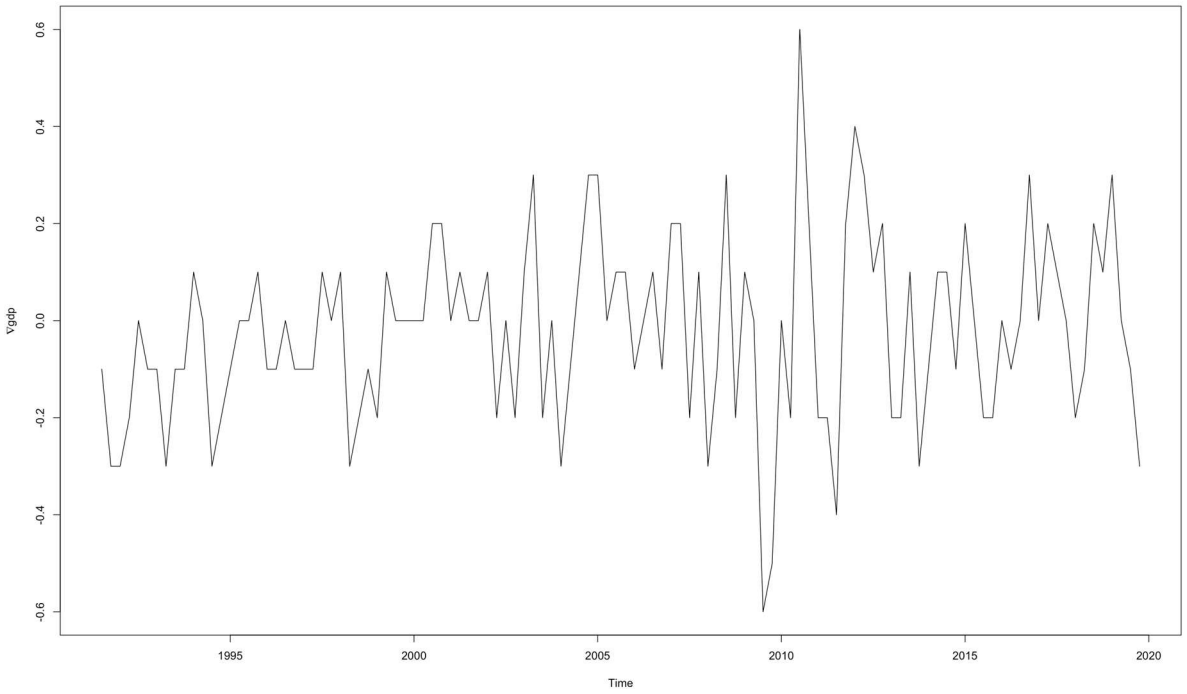

Figure 70. Inflation rate with one regular difference

Augmented Dickey-Fuller Test

Ho: Time series is not stationary
H1: Time series is stationary

Dickey-Fuller = -5.6206, Lag order = 4, p-value = 0.01 < 0.05
alternative hypothesis: stationary

Now we can reject the null hypothesis (i.e., accept that the series is stationary) and start modelling the ARIMA model. Using the function auto.arima, we get to this tentative ARIMA(1,1,1)(0,0,1)$_{[4]}$, one-

order AR part and a one-order MA part as well. But also a certain seasonality has been detected and modelled by the SMA(1).

Although the auto.arima function works with the stationary function and the difference by itself, we must take care of this assumption. Now we will check if the series has a constant trend and a constant variance. Looking at the figure below we can confirm that the series is stationary and so on start using and analysing the model.

We need to ensure that the model parameters are meaningful and can be used in the model. As we can see the AR, MA and SMA parts are statistically significant with a p-value lower than 0.05. Based on this result, we can guarantee that these parameters are conclusive and can be used in the model to predict macroeconomic indicators.

Table of ARIMA model coefficients			
ARIMA	AR1	MA1	SMA1
Coefficient	0.97	-0.74	-0.78
P-value	0.00	0.00	0.00

Before we can proceed, we will need to analyse whether the residuals are indeed a white noise. For this, we will detect if the residuals have a mean close to 0, that the variance is constant and that we cannot detect any autocorrelation.

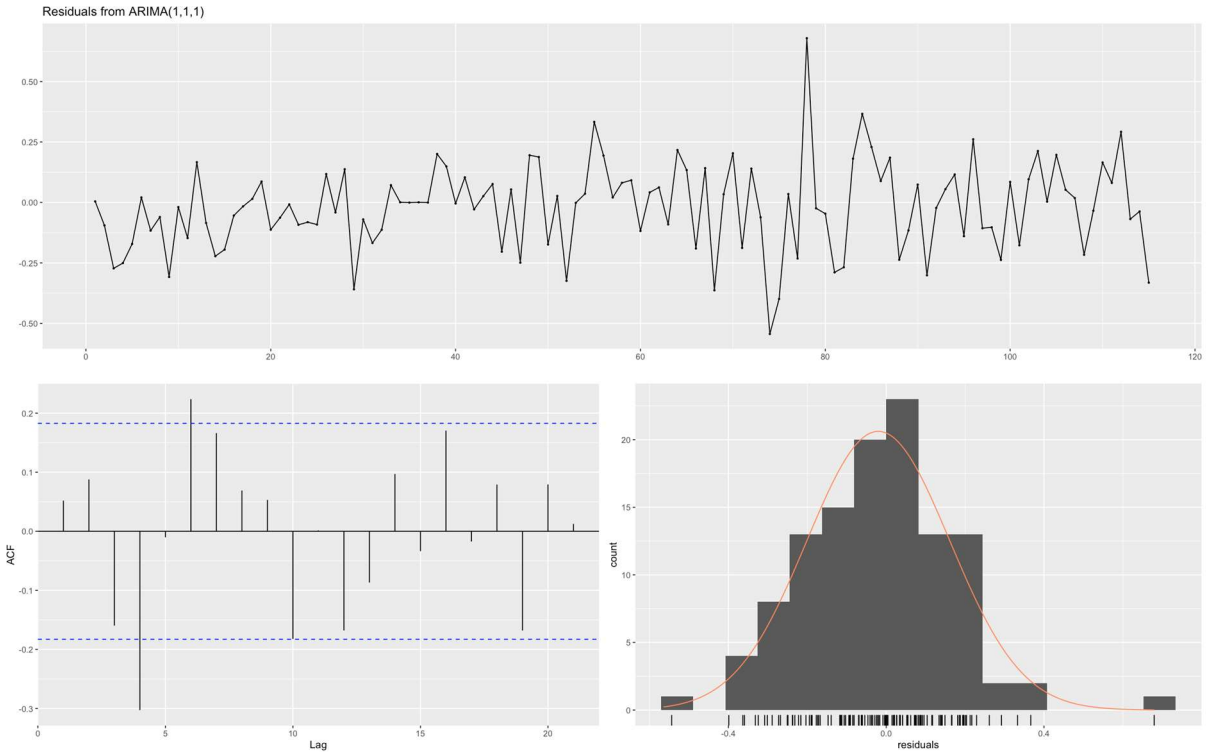

Figure 71. Residual analysis for ARIMA model

Based on Figure 71, we cannot identify an anomaly in terms of variance, and the mean of the residuals seems to be close to 0, moreover, certain normality of the residuals can be assumed. As for the

autocorrelation, it is possible that a certain vulnerability of this hypothesis can be detected. However, it is only the 4th coefficient that results statistically significant, possibly due to the randomness of the sampling and not of the model itself. To complete this analysis, we will perform the Ljung-Box Test that permits to detect if the residuals are autocorrelated:

Ljung-Box test
Data: Residuals from ARIMA(1,1,1)(0,0,1)[4]

Ho: residuals are not autocorrelated
H1: residuals are autocorrelated

Q* = 9.7804, df = 5, p-value = 0.0817 > 0.05
Model df: 3. Total lags used: 8

As the p-value is superior to 0.05, we can assume that no autocorrelation has been detected in the residuals with a 95% of statistical evidence.

To corroborate the normality of the residuals and if they are a white noise, we will construct a normal probability plot (Figure 72).

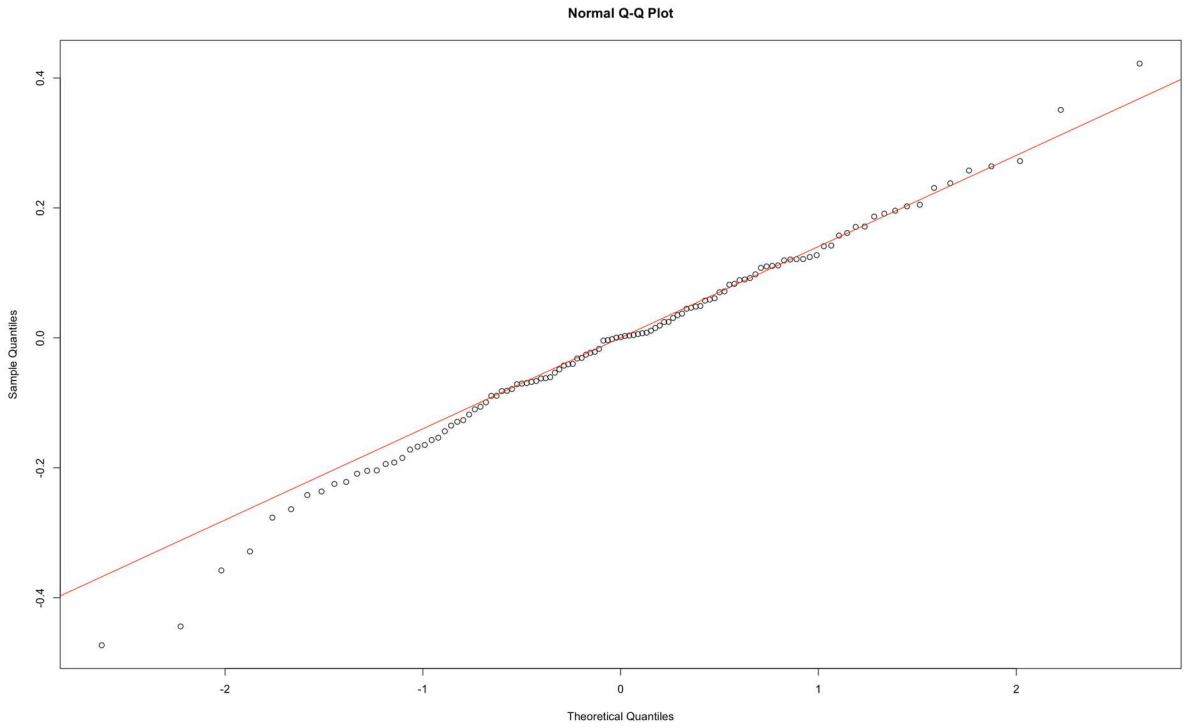

Figure 72, Normal Probability Plot of residuals for ARIMA model

From all the previous analyses, it can be affirmed that the residuals follow a white noise process and that consequently the fitted ARIMA model can be exploited to predict the Inflation Rate. When predicting the Inflation Rate with the training set, we obtain the following values for the next year predicted. We next compare those results to the validation set to obtain the estimation of the MAPE:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set	1.55	1.56	1.49	1.66
Validation Set	1.7	1.8	1.6	1.8
MAPE	15.72%			

Notice that the results obtained in term of MAPE are not outstanding (15,72%), even though we can surely notice that the results are close to the reality. As a conclusion of the ARIMA models, we can ensure that for this time series and this time axis the use of ARIMA models makes it possible to obtain a reliable estimate of inflation rates.
We will now deepen our analysis by including other exogenous variables correlated with Inflation Rate to see if it is possible to improve the prediction made previously. We will thus construct the ARIMAX models.

4.3 ARIMAX

For this model, we will select one or more financial market to predict the inflation rate. We will first try with the financial indices most correlated with the Inflation Rate, detected thanks to the PLS analysis carried out previously: GBP.USD (positively correlated) and S&P500 and the Nasdaq (negatively correlated). As no satisfying ARIMAX model has been obtained with those variables, we have explored more combinations and shown as followed the combination which best result has obtained. We finally construct the ARIMAX models with the exogenous variable EUR / USD.

To build an ARIMAX model, we will use the same methodology as before for the construction of the ARIMA model, but we will add in the auto.arima function an argument "xreg" which will include the matrix of the values of the stock indices that we will finally have to select to create the best predictive model.

Before using ARIMAX model we must transform and differentiate the series that we will use as regressors in the same way we transformed the Inflation Rate time series. Once the regressors (i.e. exogenous) time series have been differentiated will try different regressors and combination of them to show later on the best ARIMAX model we found so far. The auto.arima function form the package “forecast” we obtain the following tentative model ARIMAX(2,1,2)(1,0,1)$_{[4]}$, The function auto.arima selected a two-order AR part and a two-order MA part as well. But also, a certain seasonality has been detected and modelled by the SMA(1) and SAR(1). Before proceeding to the forecasting, we will check the statistical significance of the parameters. Notice that all of them has passed the test with a p-value near to 0, but the MA(1) whom p-value is equal to 0.19.

Once this information has been considered, we will decide to proceed with the analysis of the residuals of this model to validate and to be able to predict.

Table of ARIMAX model coefficients						
ARIMAX part	AR1	AR2	MA1	MA2	SMA1	EUR/USD
Coefficient	-0.03	0.94	0.14	-0.70	-0.84	0.39
P-value	0.06	0.00	0.19	0.00	0.00	0.015

Before we can proceed with the prediction, we will need to analyse whether the residuals are indeed white noise. For this, we will detect if the residuals have a mean close to 0, that the variance is constant and that we cannot detect any autocorrelation.

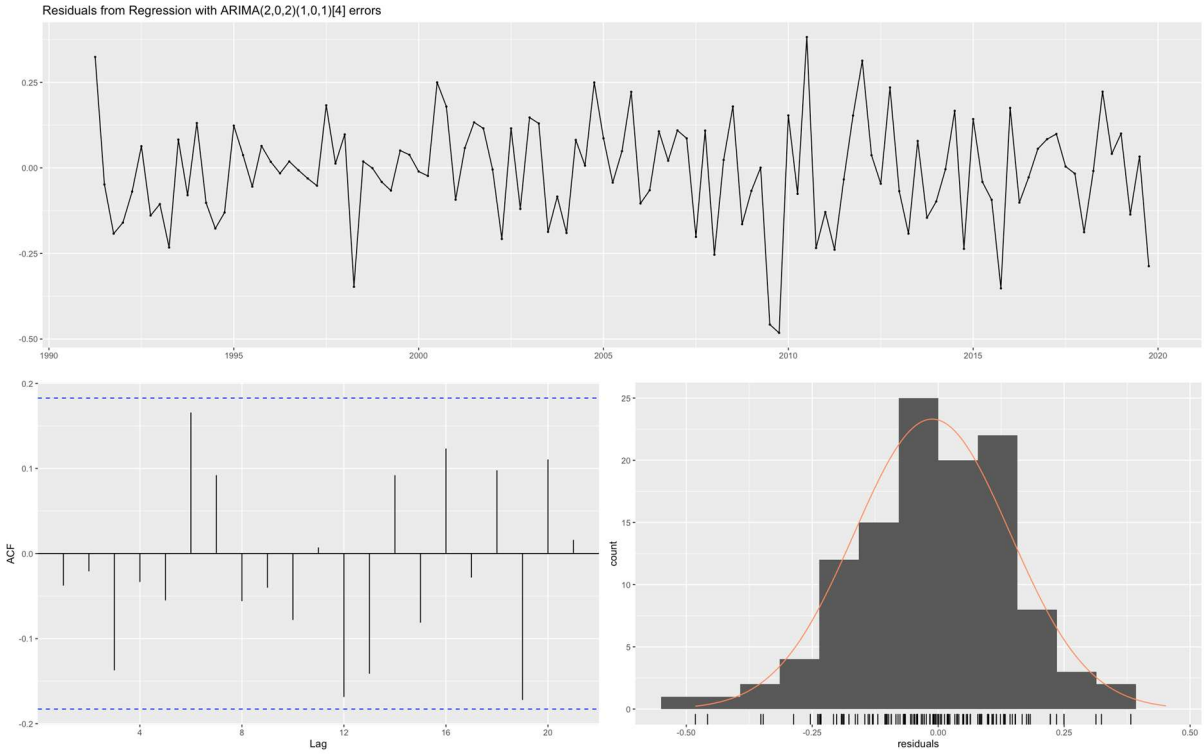

Figure 73. Residual analysis for ARIMAX model

Based on Figure 73, we cannot identify an anomaly in terms of variance, and the mean of the residuals seems to be close to 0, moreover, certain normality of the residuals can be assumed. From the ACF, we cannot observe any coefficient exceeding the threshold, in this way it can be assumed that the residuals are not autocorrelated and, therefore, by meeting all the requirements of a white noise, the constructed ARIMAX model can be validated. To complete this analysis, we will perform the Ljung-Box Test that permits to detect if the residuals are autocorrelated:

Ljung-Box test

data: Residuals from Regression with ARIMA(2,0,2)(0,0,1)[4] errors

Ho: residuals are not autocorrelated
H1: residuals are autocorrelated

Q* = 6.3842, df = 3, p-value = 0.09434 > 0.05
Model df: 6. Total lags used: 9

As the p-value is superior to 0.05, we can assume that no autocorrelation has been detected in the residuals with a 95% of statistical evidence.

To corroborate the normality of the residuals and if they are a white noise, we will construct a normal probability plot (Figure 74).

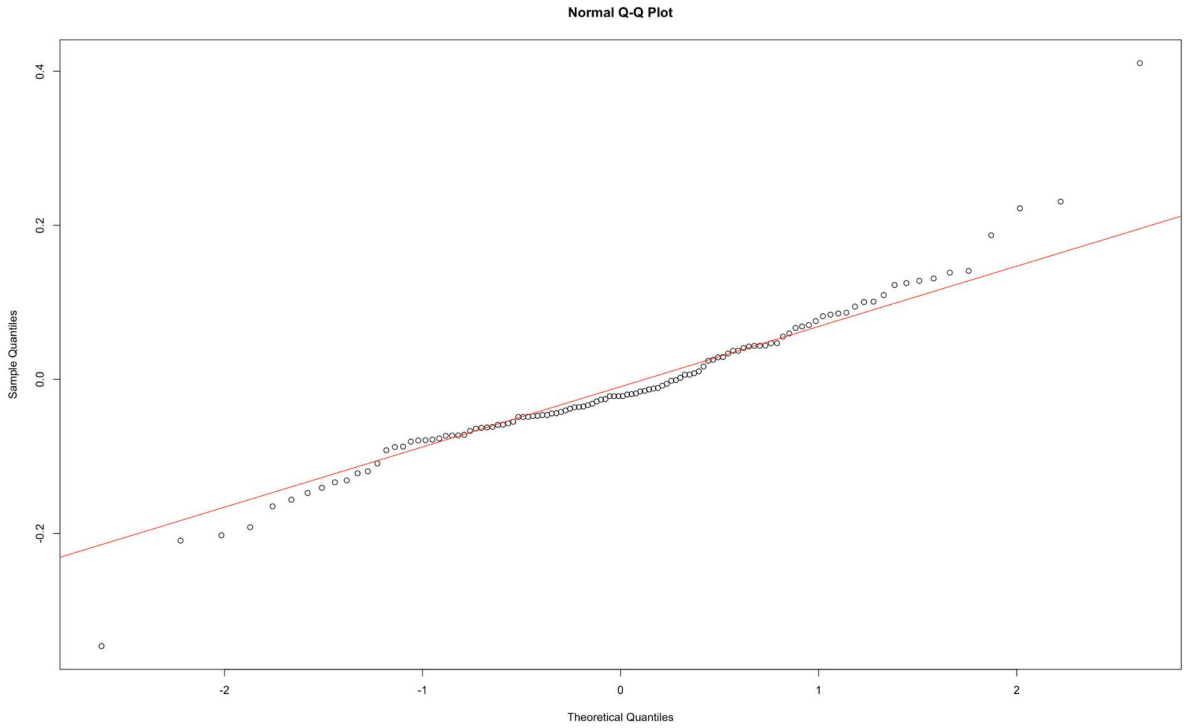

Figure 74. Normal Probability Plot of residuals from ARIMAX model

When predicting the inflation rate with the training set, we obtain the following values for the next year predicted. We next compared those results to the validation set to obtain the estimation of the MAPE:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set	1.55	1.56	1.49	1.66
Validation Set	1.7	1.8	1.6	1.8
MAPE	9.04%			

First, we took advantage of the correlation structures detected in the PLS model. To determine which financial indicators are most likely to improve the prediction made further. Afterwards, we used a more sophisticated model allowing to include the new regressors in the model to improve the prediction made before. As we can see, the predictions are more precise and come closer to reality. We obtain a MAPE of 9.04% improving the predictions from the ARIMA model.

In conclusion, we can with this first approach ensure that for this time axis and for the Inflation Rate the use of stock market indices as a new source of data in the field of prediction is a viable source that allows improving predictions of macroeconomic time series.

4.4 VAR MODELLING

As an alternative to ARIMAX models, we are going to build a VAR model. Notice that we will not take advantage of the full potential of VARs model as we only want to predict the macroeconomic indicator, so only the first predictive equation will be used. As discussed before in the Partial Least Squares methodology, financial indices apparently related to the Inflation Rate are the SP500, the Nasdaq, the GBP/USD and the Dow Jones Industrial. Nevertheless, the best combinations will be tried and shown in the following.

To build VARs models we will use the function VAR from the package "vars". Following the same procedure as ARIMAX models we must make sure that the series we work with are stationary and the same differential order have been applied. As discussed before, the inflation rate time series is made stationary after applying one regular difference. Therefore, one regular difference will be used on financial markets used.

Lately we identify that the GBP / USD is strongly and positively correlated to the inflation rate. This is the reason why we decided to work with this variable. We develop the VAR model with different indices and multiple indices. Best results predictive results have been obtained using the GBP / USD alone. To illustrate how series evolve together we plot them in Figure 75.

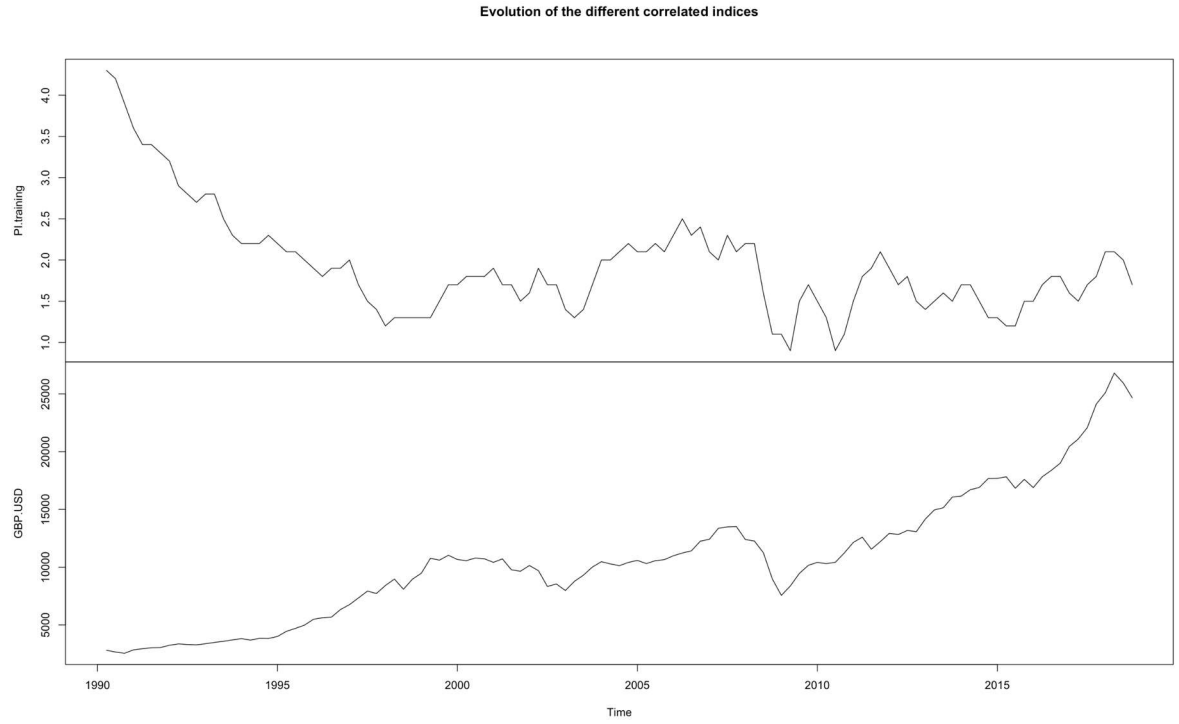

Figure 75. Inflation Rate & GBP / USD time series

To select the VARs order we will use the function VARselect from the package "vars". Once the function has been running, four criteria are proposed: the AIC, HQ, SC and FPE. To remain conservative in our analysis when selecting the order of the model, we will base ourselves mainly on the results offered by HQ and SC criteria.

AIC(n)	HQ(n)	SC(n)	FPE(n)
4	**1**	**1**	4

As the SC criterion proposes a 1 order VAR modelling, we will follow on with this VAR model order. Once fitted, we obtain the next two equations, one for the Inflation Rate depending on the pair GBP / USD, and an equation for the GBP / USD depending on the Inflation Rate.

$$\nabla Inflation\ Rate_{\mathrm{t}} = \alpha_1 + (1-\phi_{11})\nabla\, Inflation\, rate_{t-1} + \left(1-\phi_{12}\right)\nabla(GBP\ /\ USD)_{t-1} + w_{t,1}$$
$$\nabla(GBP\ /\ USD)_{\mathrm{t}} = \alpha_2 + (1-\phi_{21})\nabla(GBP\ /\ USD)_{t-1} + \left(1-\phi_{22}\right)\nabla Inflation\, rate_{t-1} + w_{t,2}$$

As we can observe, since the Inflation Rate and the GBP / USD were used to fit the VAR model, we obtain two predictive equations depending on each other (one for each index).

To validate the fitted model, we need to check that the residuals are white noise, i.e., zero mean and uncorrelated with the previous periods. To do this, we run the serial.test function. As we can see no autocorrelation is detected by the Portmanteau test as the p-value is higher than alpha (5%). So, we can assume that the residuals of the model are white noise.

Portmanteau Test (asymptotic)

Ho: residuals are not autocorrelated
H1: residuals are autocorrelated

data: Residuals of VAR object Model1
Chi-squared = 26.116, df = 16, p-value = 0.05241 > 0.05

Another aspect to consider is the presence of **heteroscedasticity**, to detect it, we will use the ARCH effect, as a multivariate estimation of the heteroscedasticity of the model. Commonly, in series such as stock prices there could be excessive volatility thereby changing the variance of the residuals, far from our assumption of constant variance.

ARCH (multivariate)

Ho: variance is constant
H1: variance is not constant

data: Residuals of VAR object Model1
Chi-squared = 153.54, df = 135, p-value = 0.1312 > 0.05

Again, the results of the ARCH test show no degree of heteroscedasticity as we accept the null hypothesis. Therefore, we conclude that there are no ARCH effects in this model.

A soft but desirable assumption is the normality of the distribution of the residuals. To test for the normality of the residuals, we use the normality.test function in R which brings in the Jarque-Bera test, the Kurtosis Test, and the Skewness test.

JB-Test (multivariate)

Ho: residuals are normally distributed

H1: residuals are not normally distributed

Chi-squared = 6.024, df = 4, p-value = 0.1974 > 0.05

Skewness only (multivariate)

Chi-squared = 2.843, df = 2, p-value = 0.2414 > 0.05

Kurtosis only (multivariate)

Chi-squared = 3.181, df = 2, p-value = 0.2038 > 0.05

As no test has been resulting to reject the null hypothesis, based on all the three results, we can accept that the residuals of this model are normally distributed. So on, we will perform a stability test to detect the presence of structural breaks of the residuals. We use the function stability from the package "vars". If at any point in the graph the residuals cross the red critical bounds of 95%, then a structural break might be detected. As we can observe as followed in Figure 76, no structural break has been observed.

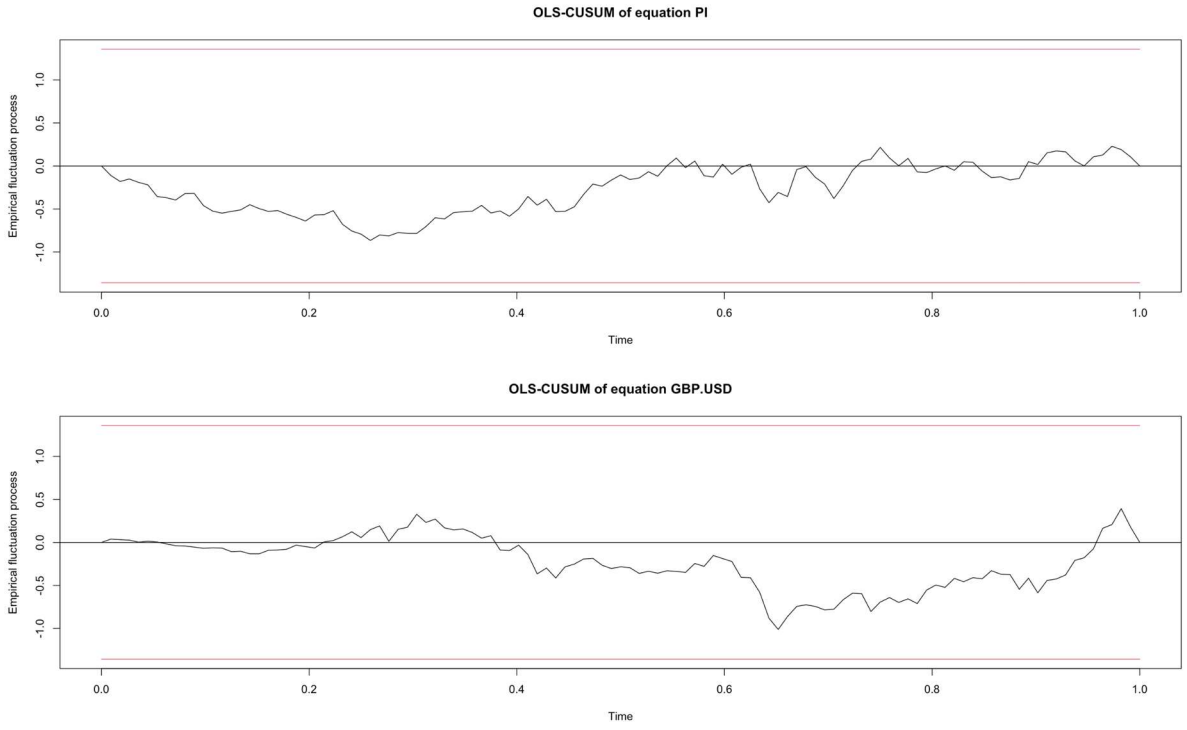

Figure 76. OLS-CUSUM detection of structural breaks of residuals.

Following VAR analysis, we will test for an overall Granger causality testing each variable in the system against all the others. As we said, there could be a unidirectional, bidirectional, or no causality relationships between the Inflation Rate and the GBP / USD.

Granger causality

H0: Inflation Rate do not Granger-cause GBP.USD
H1: Inflation Rate do Granger-cause GBP.USD

data: VAR object Model1
F-Test = 0.70576, df1 = 1, df2 = 218, p-value = 0.4018 > 0.05

H0: No instantaneous causality between: Inflation Rate and GBP.USD
H1: Instantaneous causality between: Inflation Rate and GBP.USD

data: VAR object Model1
Chi-squared = 6.9342, df = 1, p-value = 0.008456 < 0.05

As we can observe thanks to the Granger causality test, The p-value of the first test is above the significance level and therefore we cannot assume a Granger causality-effect relationship. However, we can reject with high statistical significance the null hypothesis that there is no instantaneous causality between the Inflation Rate and the GBP / USD. As an instantaneous causality has been detected we will in Figure 77, represent the Impulse Response Function (IRF). As we can see on the first periods, it seems that the GBP / USD has a high impact on the Inflation Rate, and that over time this effect is diluted.

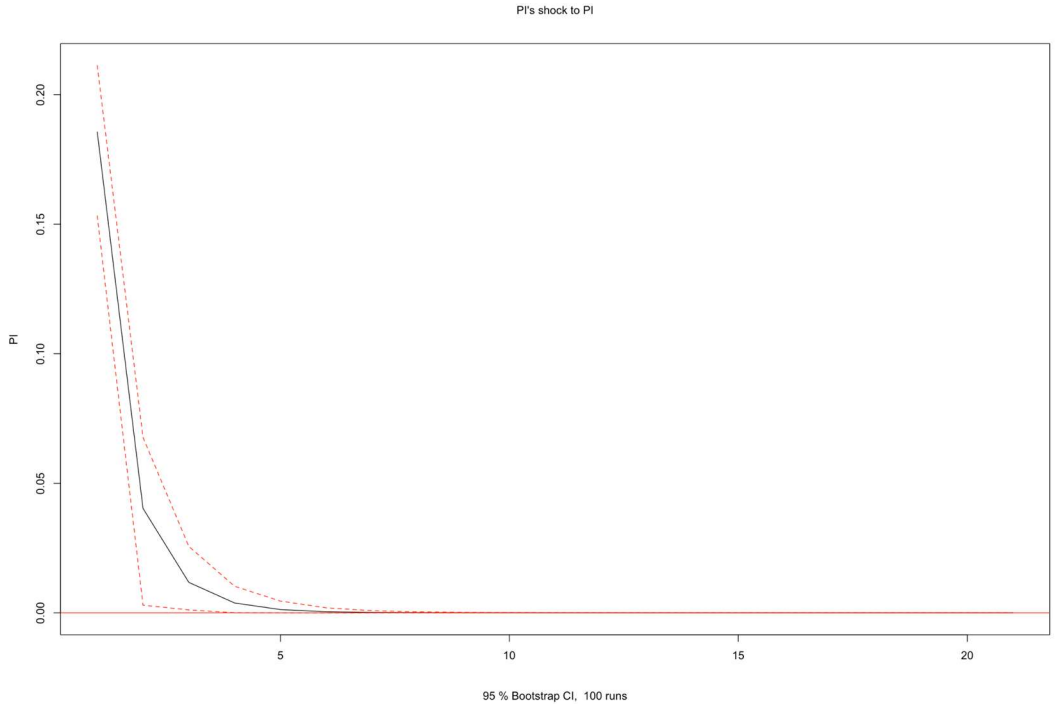

Figure 77. VAR model Impulse Response Function for inflation rate (PI).

Once we have studied and validated the VAR model, we will perform the prediction of the Inflation Rate for the next year. We use the predict function from the package "tseries". We will set the forecast horizon to 4 periods ahead or a full-year forecast (Figure 78).

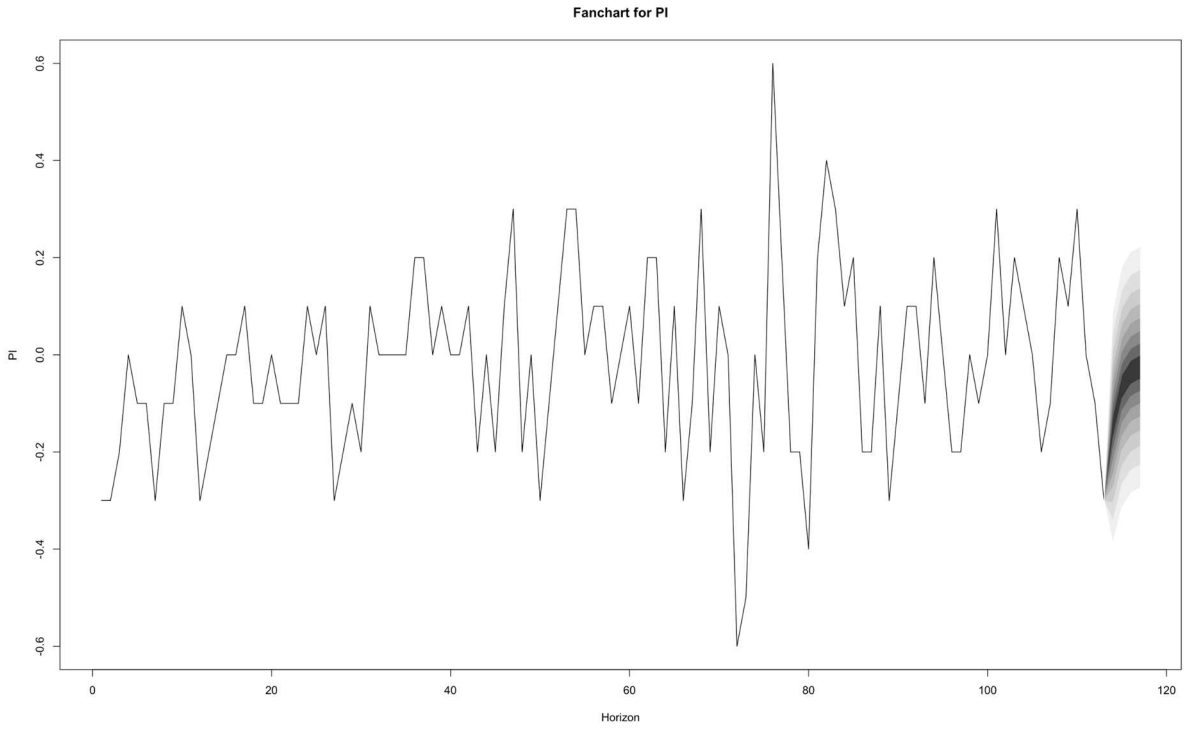

Figure 78. Forecasting with VAR model

When predicting the Inflation Rate with the training set, we obtain the following values for the next year predicted. We next compared those results to the validation set to obtain the estimation of the MAPE:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set	-0.99	-0.94	-1.09	-1.10
Validation Set	-1.00	-0.90	-1.20	-0.8
MAPE	12.95%			

As we can observe the resulting MAPE equals to 12.95%, offering a better result than the ARIMA model, reinforcing the idea that some information in the financial indices can be used to improve the predictions made by the traditional models. Nevertheless, this model has not improved the results obtained in the ARIMAX.

Now that we have been able to study and predict the Inflation Rate using the main econometric models, the next stage of this work will focus on the use of alternative models, to contrast the results obtained previously. Therefore, we will develop next, a transfer function model estimated using three techniques: Partial Least Squares, Neural Networks, and finally Recurrent Neural Networks. Note that for some of these models we will use the lagged variables to model the dynamics of the series used.

4.5 PARTIAL LEAST SQUARES TO ESTIMATE A TRANSFER FUNCTION WITH EXOGENOUS VARIABLES

In this section we will use the dynamic PLS model fitted in Chapter 3. Predictions will be obtained using the R package. When predicting the inflation rate with the training set, we obtain the following values for the next year predicted. we next compared those results to the validation set to obtain the estimation of the MAPE:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set	1.71	1.65	1.65	1.57
Validation Set	1.7	1.8	1.6	1.8
MAPE	6.20%			

The MAPE obtained thanks to the PLS model with five components is 6.20%. Note that the number of components affects the predictive performance. If we select 2 components, we obtain a MAPE of 15,67%, if we select 30 components, we obtain a MAPE of 48,67%, and it is with 5 components that we obtain the best value of MAPE equal to 6,2%.

4.6 NEURAL NETWORKS TO ESTIMATE TRANSFER FUNCTION WITH EXOGENOUS VARIABLES

Once the PLS models have been used to model a transfer function to predict macroeconomic indicators, we will turn our attention to Neural Networks, and for this we will develop a Neural Network with the lagged variables, in this way we can ensure that the model considers the dynamics of the series.

For this model we will use all financial markets data and the lagged inflation rate five times. As followed, we can observe the network being created (Figure 79).

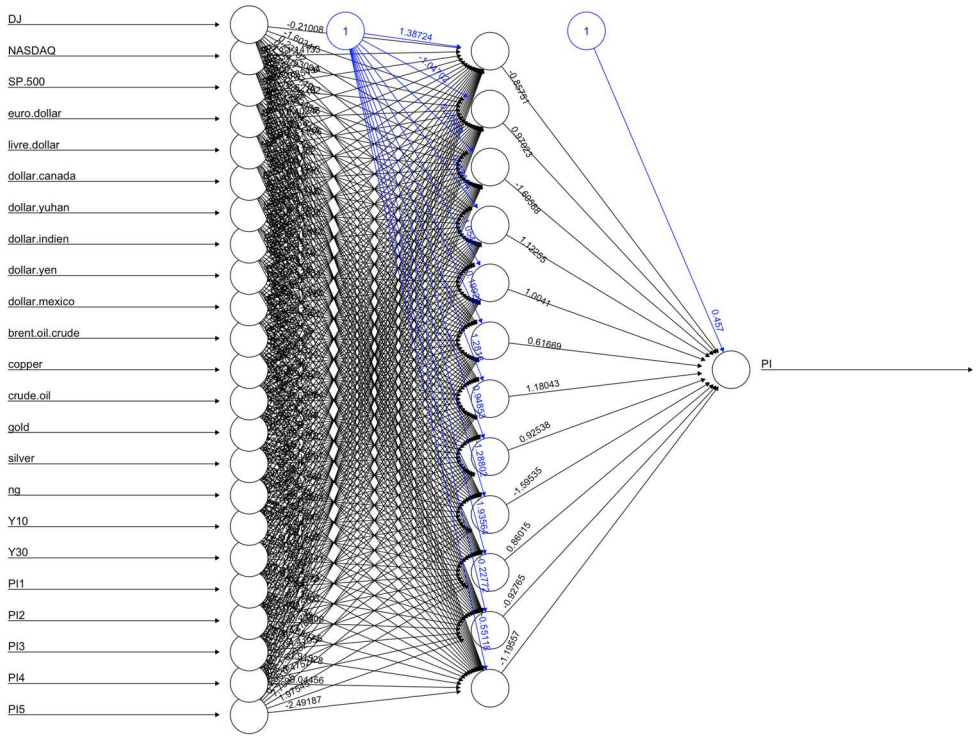

Figure 79. Modelling Transfer Function Model with neural networks for Inflation Rate

When predicting the Inflation Rate with the training set, we obtain the following values for the next year predicted. We next compared those results to the validation set to obtain the estimation of the MAPE:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set	1.58	1.65	1.54	2.05
Validation Set	1.7	1.8	1.6	1.8
MAPE	8.26%			

For the construction of this model, 12 hidden layers have been selected, and the resulting MAPE is 8.26%, slightly higher than dynamic PLS and quite similar to ARIMAX model. It should be noted that neural networks can be made much more complex by increasing the number of nodes and layers, and modifying the activation functions within the nodes. Finding the best neural network is beyond the scope of this work, but the objective is to observe that with a sophisticated network it has been possible to find prediction results that are like other more classic models and that the use of the neural networks in the predictive work of macroeconomic indicators and financial indices is not ruled out in terms of the quality of the results obtained.

4.7 RECURRENT NEURAL NETWORKS

Neural networks are well known to offer good predictive result but are not especially designed for time series. For this reason, we will use the recurrent neural networks develop by Elman and Jordan.

When predicting the Inflation Rate with the training set, we obtain the following values for the next year predicted. We next compared those results to the validation set to obtain the estimation of the MAP:

Forecast	2019-Q1	2019-Q2	2019-Q3	2019-Q4
Training set (Elman)	1.38	1.61	1.74	1.58
Training set (Jordan)	1.49	1.63	1.76	1.64
Validation Set	1.7	1.8	1.6	1.8
MAPE (Elman)	12.26%			
MAPE (Jordan)	9.98%			

Even though recurrent neural networks work better for time series, it seems that the traditional neural networks performed better for this specific prediction.

4.8 SUMMARY OF RESULTS FOR ALL MACROECONOMIC INDICATORS AND ALL OTHER RELEVANT MODELS.

Following a similar procedure as explained in the previous section for predicting the Inflation Rate, we have fitted different models for predicting the other two macroeconomic indicators: GDP and Unemployment Rate. Figure 80 shows the MAPE obtained for all responses and fitted models.

MAPE	Inflation Rate	Gross Domestic Product	Unemployment Rate	Average results
Dynamic Partial Least Squares	6,20%	0,36%	1,83%	2,80%
Arimax modeling	9,04%	2,26%	2,21%	4,50%
Partial Least Squares	8,16%	2,71%	5,50%	5,46%
Transfer Functions Modelling - Neural Networks	8,26%	3,50%	4,97%	5,58%
Vars	12,95%	3,24%	1,99%	6,06%
Arima	15,72%	0,20%	3,66%	6,53%
Recurrent Neural Networks (Jordan)	9,98%	0,19%	12,40%	7,52%
Neural Networks	11,05%	3,75%	8,80%	7,87%
Recurrent Neural Networks (Elman)	12,26%	1,19%	14,12%	9,19%
Average	8,92%	2,41%	3,30%	4,88%

Figure 80. Summary of models

Based on the overall results, we can see that dynamic Partial Least Squares methods has obtained the best results for this dataset. ARIMAX models has also obtained overall good results. As a limitation of these results, we must be aware that the validation set correspond to year 2019 and working with another validation set might change the result of this analysis.

5. Conclusions

At first, the main goal of this work was to determine if the data obtained from the financial indices could be used to improve the predictions made in macroeconomics, in particular to obtain better estimates of the behaviour of macroeconomic indicators such as the GDP, the Inflation Rate, and the Unemployment Rate.

To achieve such a task, we began our analysis with an exploratory study of financial indices (X space). Thanks to the PCA we were able to identify the main correlation structures present in the dataset. A similar analysis was undertaken with the macroeconomic indicators (Y space). In a second step, the objective was to relate both the X space and the Y space, for this we used the Partial Least Square method. Thanks to the PLS, we have been able to detect which indices have a clear relationship, and thanks to this we have used this information to improve the prediction models. But before, it must be remembered that we work with time series and therefore that it is expected a strong dynamic in the database. Therefore, we used the PLS model to estimate a transfer function with exogenous variables. Once the analyses with the PLS and PCA models were completed, we were able to gain a clear understanding of our database. Moreover, as each macroeconomic indicator showed a clear relationship with one or more financial indices, we decided to continue the analysis by moving on to the construction of prediction models.

For this, firstly, to obtain a standard of prediction, we fitted an ARIMA, ARIMAX and VAR models. The results obtained were very conclusive and positive because the predictions were sufficiently close to reality to be able to obtain a valid estimate when obtaining information on the future values of macroeconomic indices. Once fitted the main econometric models, we contrasted these with less common models in econometrics such as neural networks. The results obtained were satisfactory but generally less useful in term of interpreting results as those obtained by the main econometric models. We must note the outstanding results obtained by the dynamic PLS models when estimating the transfer function.

Thanks to this analysis, we have been able to see how the statistical models are able to take advantage of the information available in the financial markets and thus improve the prediction of the Gross Domestic Product, the Inflation Rate, and the Unemployment Rate. The results obtained are very satisfactory since predictions are very close to reality. Thus, offering a possibility of predicting with consistency and precision the most relevant indicators for the government.

In general, the possibility of being able to predict macroeconomic indicators would allow the government to have more precise information on macroeconomic components beforehand. If so, these types of statistical analysis could improve the applications of fiscal and monetary policies in qualitative terms. This is possible since, for example, when predicting the inflation rate, if a drastic increase in it is expected, the government, when applying the corresponding monetary policies, may take this prediction into account to try to reduce this forecasted increase.

In conclusion, we have been able to determine that it is possible to take advantage of the abundant information of the financial markets to improve the predictions of the main macroeconomic indicators. It is for this reason that we propose the use of techniques such as Principal Component Analysis and Partial Least Squares to obtain a satisfactory exploratory analysis of the data used. In a second stage, it is proposed to use a dynamic modelling of the PLS models as well as the ARIMAX models to predict

the indices. In this way, when applying fiscal and monetary measures, economic agents could consider both qualitative and quantitative information on how the different indicators interact and what variables have a statistical effect on them.

As a possible continuation of this study, the variables on which the government can act, such as public spending, could be modelled within the models, and sensitivity analyses used to determine the effect of increasing or reducing this spending, and thus try to control inflationary periods. Similarly, both the growth of the Gross Domestic Product and the Unemployment Rate could be controlled and modelled.

6. References & R packages

6.1 REFERENCES

Boumans, M. (2012). Observations in a Hostile Environment: Morgenstern on the Accuracy of Economic Observations. *History of Political Economy*, *44*(Supplement 1), 114–136. https://doi.org/10.1215/00182702-1631806

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/c3ay41907j

Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. San Francisco: Holden-Day.

Connor, J. T., Martin, R. D., & Atlas, L. E. (1994). Recurrent Neural Networks and Robust Time Series Prediction. In *IEEE Transactions on Neural Networks* (Vol. 5, Issue 2, pp. 240–254). https://doi.org/10.1109/72.279188

Einav, L., Leibtag, E., & Nevo, A. (2008). On the accuracy of Nielsen Homescan data. *Economic Research Report*, *69*, 34. http://www.ers.usda.gov/Publications/ERR69/ERR69.pdf

García Diaz, J. C. (2016). *Predicción en el dominio del tiempo.* Valencia: Univeristat Politècnica de València.

Gunderman, R. B., & Chen, M. (2015). The limits of economics. *Journal of the American College of Radiology*, *12*(9), 981–982. https://doi.org/10.1016/j.jacr.2015.05.024

Hill, T., O'Connor, M., & Remus, W. (1996). Neural network models for time series forecasts. *Management Science*, *42*(7), 1082–1092. https://doi.org/10.1287/mnsc.42.7.1082

Hoskuldsson, A. (2003). Regression Methods. *Data Handling in Science and Technology*, *2*(C), 165–189. https://doi.org/10.1016/S0922-3487(08)70226-0

Kitov, I. (2011). GDP Growth Rate and Population. *SSRN Electronic Journal*, 1–60. https://doi.org/10.2139/ssrn.886660

Mankiw, G. (2016). *Macroeconomics* (Seventh Ed).

Montgomery, D. C., Peck, E. A., & Vining, G. G. (n.d.). *Introduction to linear analysis* (Vol. 148).

Morgenstern, O. (1974). On the accuracy of economic observations: Foreign trade statistics **Editor's note: This reprint of ch. IX of Morgenstern's work On the Accuracy of Economic Observations omits the first three paragraphs. Figure and table numbers have been re-numbered to co. In *Illegal Transactions in International Trade*. NORTH-HOLLAND PUBLISHING COMPANY. https://doi.org/10.1016/b978-0-444-10581-3.50014-8

Stern, D. I. (1993). Energy and economic growth in the USA. A multivariate approach. In *Energy Economics* (Vol. 15, Issue 2, pp. 137–150). https://doi.org/10.1016/0140-9883(93)90033-N

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

6.2 SOFTWARE AND PACKAGES

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Achim Zeileis and Gabor Grothendieck (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. doi:10.18637/jss.v014.i06 - Package(zoo)

Achim Zeileis, Christian Kleiber, Walter Kraemer and Kurt Hornik (2003). Testing and Dating of Structural Changes in Practice. Computational Statistics & Data Analysis, 44, 109-123. - Package(strucchange)

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/ - Package(lmtest)

Adam Petrie (2020). regclass: Tools for an Introductory Class in Regression and Modeling. R package version 1.6. https://CRAN.R-project.org/package=regclass - Package(regclass)

Adrian Trapletti and Kurt Hornik (2020). tseries: Time Series Analysis and Computational Finance. R package version 0.10-48. - Package(tseries)

Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra - Package(factoextra)

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.1. https://CRAN.R-project.org/package=Ggally - Package(GGally)

Bernhard Pfaff (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars. Journal of Statistical Software 27(4). URL http://www.jstatsoft.org/v27/i04/. - Package(vars)

Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2020). pls: Partial Least Squares and Principal Component Regression. R package version 2.7-3. https://CRAN.R-project.org/package=pls - Package(pls)

Christoph Bergmeir, Jose M. Benitez (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7), 1-26. URL http://www.jstatsoft.org/v46/i07/. - Package(RSNNS)

Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot - Package(cowplot)

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2020). Hmisc: Harrell Miscellaneous. R package version 4.4-1. https://CRAN.R-project.org/package=Hmisc - Package(Hmisc)

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. - Package(ggplot2)

Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr - Package(tidyr)

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl - Package(readxl)

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr - Package(dbplyr)

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, _forecast: Forecasting functions for time series and linear https://pkg.robjhyndman.com/forecast/>. - Package(forecast)

Jeffrey A. Ryan and Joshua M. Ulrich (2020). quantmod: Quantitative Financial Modelling Framework. R package version 0.4.18. https://CRAN.R-project.org/package=quantmod - Package(quantmod)

Joshua Ulrich (2020). TTR: Technical Trading Rules. R package version 0.24.2. https://CRAN.R-project.org/package=TTR - Package(TTR)

Matt Dowle and Arun Srinivasan (2020). data.table: Extension of `data.frame`. R package version 1.13.2. https://CRAN.R-project.org/package=data.table - Package(data.table)

Max Kuhn (2020). caret: Classification and Regression Training. R, package version 6.0-86. https://CRAN.R-project.org/package=caret - Package(caret)

Mehmet Balcilar (2019). mFilter: Miscellaneous Time Series Filters. R package version 0.1-5. https://CRAN.R-project.org/package=mFilter - Package(mFilter)

Pfaff, B. (2008) Analysis of Integrated and Cointegrated Time Series with R. Second Edition. Springer, New York. ISBN 0-387-27960-1 - Package(vars)

R Core Team (2020). foreign: Read Data Stored by 'Minitab', 'S', version 0.8-80. https://CRAN.R-project.org/package=foreign 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... R package 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... R package version 0.8-80. https://CRAN.R-project.org/package=foreign - Package(foreign)

Rami Krispin (2020). TSstudio: Functions for Time Series Analysis and Forecasting. R package version 0.1.6. https://CRAN.R-project.org/package=Tsstudio - Package(TSstudio)
Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5 - Package(lattice)

Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical - Package(FactoMineR)

Stefan Fritsch, Frauke Guenther and Marvin N. Wright (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet - Package(neuralnet)

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot - Package(corrplot)

Terry M. Therneau, Patricia M. Grambsch (2000). _Modeling SurvivalData: Extending the Cox Model_. Springer, New York. ISBN0-387-98784-3. - Package(survival)

Therneau T (2020). _A Package for Survival Analysis in R_. Rpackage version 3.2-7, <URL:https://CRAN.R-project.org/package=survival>. - Package(survival)

Tim Bergsma (2018). datetime: Nominal Dates, Times, and Durations. R package version 0.1.4. https://CRAN.R-project.org/package=datetime - Package(datetime)

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0 - Package(MASS)

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 - Package(tidyverse)

Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489. - Package(ggfortify)

Zeileis A, Croissant Y (2010). “Extended Model Formulas in R: Multiple Parts and Multiple Responses.” _Journal of Statistical Software_, *34*(1), 1-13. doi: 10.18637/jss.v034.i01 (URL:https://doi.org/10.18637/jss.v034.i01). - Package(Formula)