ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

*MSc in Business Economics with Analytics*

**BRUNILDA ZEQIRAJ**

# *Applications of Analytical Methods in Business Economics & Strategy*

## *Part 1*

## TABLE OF CONTENTS Pages

## Abstract

How have some European countries been affected by global crises during the previous few decades? Are there any economic metrics that demonstrate how these events have affected these countries? What methodologies will we employ to better comprehend and analyze these statistics, as well as to develop financial forecasts? These are some of the questions that our study will attempt to address.

## Introduction

In this study it will be presented how machine learning methodologies could be applied in economics and finance and it will also be demonstrated how economic forecasts can be generated using various methodologies. More specifically Austria and Belgium are the countries examined in this analysis and in which these methods will be applied.

The data was obtained from Eurostat, and some 'data cleaning' was required in order to do this research. Outliers were corrected, NA values were eliminated or rectified, and lastly Seasonalities, if they existed, were corrected to make the variables stationary in order to be able to apply some methodologies and make the results more reliable before the study began.

This study's goals include a better understanding of the variables through statistical indicators and visualization, variable selection, and the creation of cross validation forecasts for some of the indicators. To achieve these objectives, a number of models will be utilized, including ARMA, PCA, PLS, Ridge, Lasso, and k – means clustering models. All methods are applied in real data using the R Project for Statistical Analysis as the main scientific software.

So, let's get started!

In order to carry out this study, we will need certain variables to apply machine learning methodologies to. These 45 Eurostat variables that will be used are separated into 10 Hard and 35 Soft variables, with Hard variables referring to real economic data and Soft variables referring to predictions. The meaning of each variable will be explained below for a better understanding of the analysis.

## Hard Variables

- *Production in industry*-monthly data (HD1) refers to the output of industrial establishments and covers sectors such as mining, manufacturing, electricity, gas and steam and air-conditioning. This indicator is measured in an index based on a reference period that expresses change in the volume of production output.
- *Euro area 19 international trade [Exports]* -monthly data (HD2) refers to goods and services that are produced in one country and sold to buyers in another in Europe.
- *Euro area 19 international trade [Imports]* -monthly data (HD3) refers to goods or services bought in one country that was produced in another.
- *Production in construction* -monthly data (HD4) index is a business-cycle indicator which measures the monthly changes in production of buildings (residential and non-residential) and of civil engineering (roads, railways, bridges, tunnels, utility projects).
- *Unemployment rate* (HD5) is the percentage of the labor force without a job. It is a lagging indicator, meaning that it generally rises or falls in the wake of changing economic conditions, rather than anticipating them. When the economy is in poor shape and jobs are scarce, the unemployment rate can be expected to rise.
- *Money market interest rates [3months yield]* -monthly data (HD6) means the fixed interest rate per annum which the Bank determines could be obtained by reinvesting a specified Prepaid Installment in the Money Market from the date of prepayment through the Original Payment Date.
- *Euro yield curves [10 Year Yield]* -monthly data (HD7) is a representation of the relationship between market remuneration rates and the remaining time to maturity of debt securities in Europe.
- *Turnover and volume of sales in wholesale and retail trade* -monthly data (HD8) .The volume of sales index, more commonly called the index of the volume of (retail) sales, is the value of retail sales in terms of its volume. The volume of sales index is a volume measure of the retail trade turnover index.
- *Euro / ECU [USD] exchange rates* -monthly data (HD9). The European Currency Unit (ECU) was the official monetary unit of the European Monetary System (EMS) before it was replaced by the euro. The value of the ECU was used to determine the exchange rates and reserves among the members of the EMS, but it was always an accounting unit rather than a real currency.
- *Spread (10Y-3M)* (HD10) refers to the difference between two prices, rates, or yields.

## Soft Variables

- *Production development observed over the past 3 months [Industry]* (SD1) is the complete process of taking an idea from concept to delivery and beyond. Whether you are delivering a brand new offering or enhancing an existing product, the product development cycle begins long before anything gets built.
- *Production expectations over the next 3 months[Industry]* (SD2) is an prediction of what is expected the production will be in the next 3 months

- *Employment expectations over the next 3 months[Industry]* (SD3) summarizes managers' employment plans in four surveyed business sectors (industry, services, retail trade and construction) and thus provides a timely indication of expected changes in dependent employment.
- *Assessment of order-book levels [Industry]* (SD4) refers to the assessment of the electronic list of buy and sells orders for a specific security or financial instrument organized by price level. An order book lists the number of shares being bid on or offered at each price point, or market depth.
- *Assessment of export order-book levels [Industry]* (SD5) refers to the assessment of the electronic list conveying the choice of foreign purchaser to buy goods from the exporter.
- *Assessment of the current level of stocks of finished products [Industry]* (SD6) refers to the assessment of the level of the value of goods that have completed the manufacturing process and are available for distribution to customers.
- *Building activity development over the past 3 months [Construction]* (SD7) means the development of erection, re-erection, making material alteration, or demolition of any building.
- *Evolution of the current overall order books [Construction]* (SD8) is the evolution of the total order quantity accumulated from the best bid (ask) price to this price level.
- *Employment expectations over the next 3months [Constructions]/* (SD9] helps getting a timely indication of expected changes in dependent employment. The indicator is constructed as a weighted average of the employment expectations of managers in the four surveyed business sectors (industry, services, retail trade and construction).
- *Price expectations over the next 3 months [Constructions]* (SD10) are used as reference points to help make final purchase decisions.
- *Business activity (sales) development over the past3 months [Retail]* (SD11) can be summarized as the ideas, initiatives, and activities that help make a business better.
- *Volume of stocks currently hold[Retail]* (SD12) measures the number of shares traded in a stock or contracts traded in futures or options. Volume can indicate market strength, as rising markets on increasing volume are typically viewed as strong and healthy. When prices fall on increasing volume, the trend is gathering strength to the downside.
- *Expectations of the number of orders over the next 3 months [Retail]* (SD13) are how many orders are expected to have over the next 3 months.
- *Business activity expectations over the next 3 months [Retail]* (SD14) are the expectations of the ideas, initiatives, and activities that help make a business better.
- *Employment expectations over the next 3months [Retail] (*SD15) helps getting a timely indication of expected changes in dependent employment. The indicator is constructed as a weighted average of the employment expectations of managers in the four surveyed business sectors (industry, services, retail trade and construction).
- *Business Situation over the past 3 months [Services]* (SD16) is the situation of services used by the business enterprises in conducting the activities of the business.
- *Evolution of Demand over the past 3 months[Services]* (SD17) measures the evolution of demand which is is an economic principle referring to a consumer's desire to purchase goods and services and willingness to pay a price for a specific good or service
- *Expectation of Demand over the next 3 months [Services]* (SD18) is the expectations that buyers have concerning the future price of a good, which is assumed constant when a demand curve is constructed. Buyers' expectations are one of five demand determinants that shift the demand curve when they change.
- *Evolution of employment over the past 3 months[Services]* (SD19) is how the employment was expected to evolve over the past 3 months

- *Expectation of Employment over the next 3 months [Services] (*SD20) helps getting a timely indication of expected changes in dependent employment. The indicator is constructed as a weighted average of the employment expectations of managers in the four surveyed business sectors (industry, services, retail trade and construction).
- *Euro-zone Business Climate Indicator –*monthly data (SD21) is calculated in order to receive a timely composite indicator for the manufacturing sector in the euro area.
- *Construction confidence indicator* (SD22) s the arithmetic average of the balances (in percentage points) of the answers to the questions on order book and employment expectations.
- *Economic sentiment indicator* (SD23) s calculated based on a selection of questions from industry, services, retail trade, construction and consumers at country level and at aggregate level (EU and euro area) in order to track overall economic activity. ESI has been calculated since 1985.
- *Industrial Confidence Indicator* (SD24) is the arithmetic average of the balances (in percentage points) of the answers to the questions on production expectations, order books and stocks of finished products (the last with inverted sign).
- *Retail Confidence Indicator* (SD25) is the arithmetic average of the balances (in percentage points) of the answers to the questions on the present and future business situation, and on stocks (the last with inverted sign).
- *Consumer Confidence Indicator* (SD26) is the arithmetic average of the balances (in percentage points) of the answers to the questions on the past and expected financial situation of households, the expected general economic situation and the intentions to make major purchases over the next 12 months.
- *Financial situation over the last 12 months /over next 12 months* (SD27)/ (SD28) refers to what the conditions in finance were and going to be. Finance is the process of raising funds or capital for any kind of expenditure. It is the process of channeling various funds in the form of credit, loans, or invested capital to those economic entities that most need them or can put them to the most productive use.
- *General economic situation over the last 12 months /over next 12 months* (SD29)/ (SD30) is he complex of elements which, in a given period, characterize the condition or state of a country or region's ability to produce goods, services and other resources with exchange value.
- *Price Trends over the last 12 months /over next 12 months* (SD31)/ (SD32) is used to determine the balance between a security's demand and supply. The percentage change in the share price trend shows the relative supply or demand of a particular security, while volume indicates the force behind the trend.
- *Unemployment Expectations over the next 12 months[Consumer]* (SD33) is defined as the projected value for the number of unemployed people over the next 12 months.
- *Major purchases over the next 12 months [Consumer]* (SD34) is a list of things of the biggest purchases that cost the most such as buying and selling a house, college education or graduate school, taxes ,a new car etc.
- *Savings over the next 12 months [Consumer]* (SD35) is the portion of income not spent on current expenditures over the next 12 months. In other words, it is the money set aside for future use and not spent immediately.

And of course one of the most important variables in economics is Gross domestic product (GDP) which is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. We will not use the GDP as much as we would want because the data is quarterly rather than monthly, making it harder to use in our methods.

## Descriptive statistics

The first stage in the study is to produce a table of summary statistics to get a better view of the variables and how they fluctuate. *Summary statistics* is a part of descriptive statistics that summarizes and provides the gist of information about the sample data. In this table will be included mean, median, minimum value, maximum value, standard deviation, skewness, kurtosis, Jarque-Bera test.

*Mean* is a type of average. It is the sum (total) of all the values in a set of data, such as numbers or measurements, divided by the number of values on the list. To find the mean, add up all the values in the set. Then divide the sum by how many values there are.

*Median* is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average. The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values.

*Minimum Value* is data value that is less than or equal to all other values in our set of data.

*Maximum Value* is data value that is most than or equal to all other values in our set of data.

*Standard Deviation* is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

*Skewness* refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

*Kurtosis* is a measure of the combined weight of a distribution's tails relative to the center of the distribution. When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.

*The Jarque-Bera's test* null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero. With a p-value >0.05, one would usually say that the data are consistent with having skewness and excess kurtosis zero.

The summary statistics for Belgium and Austria are presented in the four tables below. The statistical indicators for the hard variables are found in Tables 1 and 2, whereas the statistical indicators for the soft variables are found in Tables 3 and 4.

| ***Variables*** | ***Mean*** | ***SD*** | ***Minimum*** | ***Median*** | ***Maximum*** | ***Skewness*** | ***Kurtosis*** | ***JB Pval*** |
|---|---|---|---|---|---|---|---|---|
| **HD1** | 0.160 | 3.054 | -11.896 | 0.100 | 11.669 | -0.135 | 4.642 | 0.000 |
| **HD2** | 0.494 | 4.610 | -12.533 | 0.274 | 19.363 | 0.484 | 4.542 | 0.000 |
| **HD3** | 0.328 | 4.492 | -14.641 | 0.695 | 17.812 | -0.153 | 3.849 | 0.029 |
| **HD4** | 0.024 | 2.384 | -12.538 | 0.000 | 11.561 | 0.177 | 8.778 | 0.000 |
| **HD5** | -0.02 | 0.185 | -0.500 | 0.000 | 0.600 | 0.257 | 3.841 | 0.000 |
| **HD6** | -0.003 | 0.092 | -0.513 | -0.001 | 0.327 | -1.318 | 10.263 | 0.000 |
| **HD7** | -0.015 | 0.151 | -0.407 | -0.037 | 0.484 | 0.668 | 3.668 | 0.000 |
| **HD8** | 0.040 | 1.900 | -7.476 | 0.000 | 8.887 | -0.047 | 8.215 | 0.000 |
| **HD9** | -0.027 | 2.119 | -7.287 | -0.022 | 6.36 | -0.196 | 3.813 | 0.028 |
| **HD10** | -0.007 | 0.167 | -0.477 | -0.028 | 0.646 | 1.061 | 5.282 | 0.000 |

**Table 1.Summary statistics of Belgium -Hard Variables**

| ***Variables*** | ***Mean*** | ***SD*** | ***Minimum*** | ***Median*** | ***Maximum*** | ***Skewness*** | ***Kurtosis*** | ***JB Pval*** |
|---|---|---|---|---|---|---|---|---|
| **HD1** | 0.372 | 1.773 | -5.031 | 0.334 | 8.997 | 0.657 | 6.614 | 0.000 |
| **HD2** | 0.333 | 3.170 | -10.926 | 0.204 | 13.868 | 0.205 | 4.676 | 0.000 |
| **HD3** | 0.313 | 5.261 | -15.763 | 0.462 | 15.315 | 0.117 | 3.395 | 0.405 |
| **HD4** | 0.264 | 2.834 | -8.829 | 0.367 | 9.913 | -0.175 | 4.101 | 0.003 |
| **HD5** | -0.007 | 0.409 | -1.100 | 0.000 | 1.100 | 0.162 | 2.876 | 0.594 |
| **HD6** | -0.003 | 0.092 | -0.513 | -0.002 | 0.327 | -1.304 | 10.175 | 0.000 |
| **HD7** | -0.018 | 0.148 | -0.407 | -0.038 | 0.484 | 0.649 | 3.693 | 0.000 |
| **HD8** | -0.080 | 2.005 | -13.277 | 0.000 | 7.116 | -1.629 | 14.442 | 0.000 |
| **HD9** | -0.013 | 2.115 | -7.287 | -0.018 | 6.360 | -0.206 | 3.850 | 0.021 |
| **HD10** | -0.010 | 0.165 | -0.477 | -0.032 | 0.646 | 1.084 | 5.462 | 0.000 |

**Table 2.Summary statistics of Austria -Hard Variables**

| *Variables* | *Mean* | *SD* | *Minimum* | *Median* | *Maximum* | *Skewness* | *Kurtosis* | *JB Pval* |
|---|---|---|---|---|---|---|---|---|
| **SD1** | 0.007 | 5.388 | -20.400 | 0.150 | 19.100 | -0.117 | 4.525 | 3E-05 |
| **SD2** | 0.089 | 3.944 | -10.800 | 0.000 | 10.700 | 0.111 | 3.128 | 7E-01 |
| **SD3** | 0.167 | 3.057 | -10.200 | 0.000 | 9.700 | 0.065 | 3.869 | 3E-02 |
| **SD4** | 0.048 | 4.279 | -12.100 | 0.300 | 12.200 | -0.133 | 3.118 | 7E-01 |
| **SD5** | -0.025 | 5.185 | -17.100 | 0.350 | 13.200 | -0.316 | 3.443 | 7E-02 |
| **SD6** | 0.039 | 3.789 | -13.500 | 0.150 | 11.500 | -0.197 | 3.968 | 8E-03 |
| **SD7** | 0.188 | 4.673 | -13.800 | 0.100 | 18.100 | 0.139 | 4.387 | 2E-04 |
| **SD8** | 0.073 | 2.493 | -9.7000 | 0.200 | 6.100 | -0.288 | 3.547 | 6E-02 |
| **SD9** | 0.150 | 2.508 | -7.5000 | 0.300 | 7.800 | 0.095 | 3.786 | 6E-02 |
| **SD10** | 0.231 | 2.717 | -9.0000 | 0.100 | 10.000 | 0.261 | 4.706 | 9E-07 |
| **SD11** | 0.053 | 9.981 | -30.600 | 0.000 | 28.600 | -0.085 | 3.726 | 9E-02 |
| **SD12** | -0.060 | 5.028 | -14.900 | -0.050 | 14.700 | 0.005 | 3.583 | 2E-01 |
| **SD13** | 0.039 | 5.359 | -11.900 | 0.300 | 15.900 | 0.162 | 3.038 | 6E-01 |
| **SD14** | 0.053 | 6.989 | -24.400 | 0.000 | 24.500 | -0.221 | 4.105 | 2E-03 |
| **SD15** | 0.025 | 4.259 | -13.400 | 0.000 | 13.600 | -0.038 | 3.163 | 9E-01 |
| **SD16** | 0.193 | 5.973 | -21.400 | -0.300 | 18.300 | 0.227 | 3.830 | 2E-02 |
| **SD17** | 0.225 | 7.764 | -19.300 | -0.200 | 22.700 | 0.028 | 3.060 | 1E+00 |
| **SD18** | -0.220 | 6.038 | -18.800 | -0.150 | 14.700 | -0.208 | 3.311 | 3E-01 |
| **SD19** | 0.068 | 8.477 | -25.200 | 1.000 | 24.900 | -0.251 | 2.990 | 3E-01 |
| **SD20** | 0.201 | 6.072 | -18.600 | 0.400 | 14.900 | -0.281 | 3.584 | 6E-02 |
| **SD21** | 0.013 | 0.193 | -0.8000 | 0.000 | 0.800 | -0.411 | 5.828 | 0E+00 |
| **SD22** | 0.133 | 2.046 | -7.2000 | 0.075 | 6.200 | 0.005 | 3.364 | 6E-01 |
| **SD23** | 0.174 | 3.121 | -11.400 | 0.190 | 11.900 | 0.162 | 4.886 | 1E-07 |
| **SD24** | 0.052 | 2.774 | -7.4000 | 0.100 | 10.400 | 0.195 | 3.584 | 1E-01 |
| **SD25** | 0.088 | 4.868 | -13.500 | 0.200 | 13.700 | 0.021 | 3.253 | 8E-01 |
| **SD26** | -0.032 | 2.021 | -7.3000 | -0.100 | 5.600 | -0.092 | 3.401 | 4E-01 |
| **SD27** | -0.031 | 1.672 | -4.7000 | 0.000 | 4.000 | -0.057 | 2.961 | 9E-01 |
| **SD28** | -0.051 | 1.773 | -5.9000 | -0.100 | 4.800 | -0.268 | 3.486 | 1E-01 |
| **SD29** | -0.102 | 5.462 | -18.300 | 0.100 | 15.000 | -0.380 | 3.837 | 4E-03 |
| **SD30** | -0.057 | 5.488 | -15.500 | -0.200 | 14.700 | 0.120 | 3.135 | 7E-01 |
| **SD31** | 0.149 | 4.126 | -14.800 | 0.300 | 15.600 | 0.043 | 4.177 | 2E-03 |
| **SD32** | 0.083 | 4.636 | -17.500 | -0.200 | 12.000 | -0.137 | 3.417 | 3E-01 |
| **SD33** | -0.263 | 6.482 | -22.540 | -0.600 | 23.100 | 0.230 | 4.440 | 5E-05 |
| **SD34** | -0.003 | 2.488 | -7.1000 | -0.250 | 6.400 | -0.002 | 3.022 | 1E+00 |
| **SD35** | -0.036 | 3.296 | -10.600 | 0.050 | 9.100 | 0.152 | 3.222 | 5E-01 |

**Table 3.Summary statistics of Belgium-Soft Variables**

| *Variables* | *Mean* | *SD* | *Minimum* | *Median* | *Maximum* | *Skewness* | *Kurtosis* | *JB Pval* |
|---|---|---|---|---|---|---|---|---|
| **SD1** | -0.014 | 6.511 | -23.400 | 0.100 | 22.000 | -0.090 | 4.098 | 0.005 |
| **SD2** | 0.144 | 5.436 | -15.100 | 0.000 | 18.100 | 0.359 | 3.865 | 0.004 |
| **SD3** | 0.106 | 4.334 | -14.500 | 0.400 | 12.700 | -0.312 | 3.358 | 0.105 |
| **SD4** | 0.178 | 4.486 | -13.000 | 0.200 | 14.100 | 0.133 | 2.823 | 0.642 |
| **SD5** | 0.177 | 4.285 | -12.900 | 0.200 | 12.500 | -0.170 | 3.253 | 0.458 |
| **SD6** | -0.032 | 3.155 | -11.000 | -0.100 | 9.000 | -0.216 | 3.947 | 0.009 |
| **SD7** | 0.391 | 9.359 | -32.200 | 0.000 | 33.200 | -0.012 | 4.110 | 0.005 |
| **SD8** | 0.417 | 6.261 | -18.200 | 0.700 | 17.600 | 0.167 | 3.152 | 0.555 |
| **SD9** | 0.132 | 7.748 | -27.700 | -0.300 | 24.500 | 0.202 | 3.841 | 0.023 |
| **SD10** | 0.469 | 6.957 | -20.900 | 1.200 | 24.600 | 0.138 | 3.611 | 0.142 |
| **SD11** | 0.296 | 11.224 | -34.200 | 0.900 | 41.400 | 0.012 | 3.679 | 0.134 |
| **SD12** | -0.081 | 5.869 | -23.100 | 0.600 | 15.100 | -0.332 | 3.599 | 0.031 |
| **SD13** | 0.075 | 7.010 | -26.200 | 0.100 | 24.000 | -0.081 | 3.815 | 0.049 |
| **SD14** | -0.371 | 8.992 | -27.100 | 0.100 | 26.100 | -0.176 | 3.885 | 0.019 |
| **SD15** | 0.032 | 5.591 | -14.700 | -0.300 | 13.900 | 0.074 | 2.715 | 0.637 |
| **SD16** | 0.262 | 6.128 | -17.600 | 0.800 | 20.700 | 0.166 | 4.269 | 0.001 |
| **SD17** | 0.189 | 6.405 | -17.900 | 0.600 | 21.600 | 0.041 | 3.777 | 0.070 |
| **SD18** | 0.007 | 5.724 | -20.600 | -0.300 | 18.500 | -0.051 | 3.781 | 0.067 |
| **SD19** | -0.021 | 7.107 | -27.600 | -0.400 | 24.700 | -0.169 | 6.127 | 0.000 |
| **SD20** | 0.067 | 6.639 | -25.800 | 0.000 | 25.700 | 0.007 | 5.310 | 0.000 |
| **SD21** | 0.013 | 0.197 | -0.800 | 0.000 | 0.900 | -0.314 | 6.244 | 0.000 |
| **SD22** | 0.358 | 5.251 | -14.800 | 0.100 | 16.100 | 0.115 | 3.393 | 0.405 |
| **SD23** | 0.265 | 3.016 | -10.058 | 0.179 | 13.290 | 0.564 | 5.623 | 0.000 |
| **SD24** | 0.128 | 2.988 | -10.500 | 0.000 | 10.400 | 0.151 | 3.980 | 0.010 |
| **SD25** | -0.083 | 6.179 | -21.600 | -0.100 | 16.900 | -0.201 | 3.870 | 0.018 |
| **SD26** | 0.074 | 2.084 | -5.100 | 0.100 | 5.900 | 0.012 | 3.300 | 0.675 |
| **SD27** | 0.064 | 2.029 | -6.900 | 0.000 | 6.300 | -0.154 | 3.953 | 0.013 |
| **SD28** | 0.076 | 2.296 | -6.300 | 0.000 | 6.100 | 0.048 | 2.851 | 0.872 |
| **SD29** | 0.114 | 5.019 | -18.300 | 0.400 | 14.100 | -0.097 | 3.652 | 0.134 |
| **SD30** | 0.156 | 5.029 | -18.200 | 0.400 | 19.600 | 0.114 | 4.758 | 0.000 |
| **SD31** | 0.181 | 4.364 | -11.000 | 0.400 | 15.700 | 0.268 | 3.723 | 0.029 |
| **SD32** | 0.102 | 3.981 | -14.700 | 0.300 | 15.000 | -0.133 | 4.394 | 0.000 |
| **SD33** | -0.301 | 5.086 | -22.000 | 0.164 | 16.375 | -0.579 | 5.135 | 0.000 |
| **SD34** | -0.008 | 2.903 | -7.800 | 0.200 | 7.700 | -0.203 | 2.959 | 0.483 |
| **SD35** | -0.140 | 3.091 | -10.200 | -0.100 | 7.500 | -0.190 | 3.103 | 0.509 |

**Table 4.Summary of statistics of Austria-Soft Variables**

# Data Visualization

*Data visualization* is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. Visualization lets you comprehend vast amounts of data at a glance and in a better way. A range of graphs will be shown in this analysis in order to analyze the behavior of some variables as well as the economic standing of the two countries.

The *time series plot* is the most basic and widely used type of data visualization. A time plot is basically a line plot showing the evolution of the time series over time. We can use it as the starting point of the analysis to get some basic understanding of the data, for example, in terms of trend/seasonality/outliers, etc.

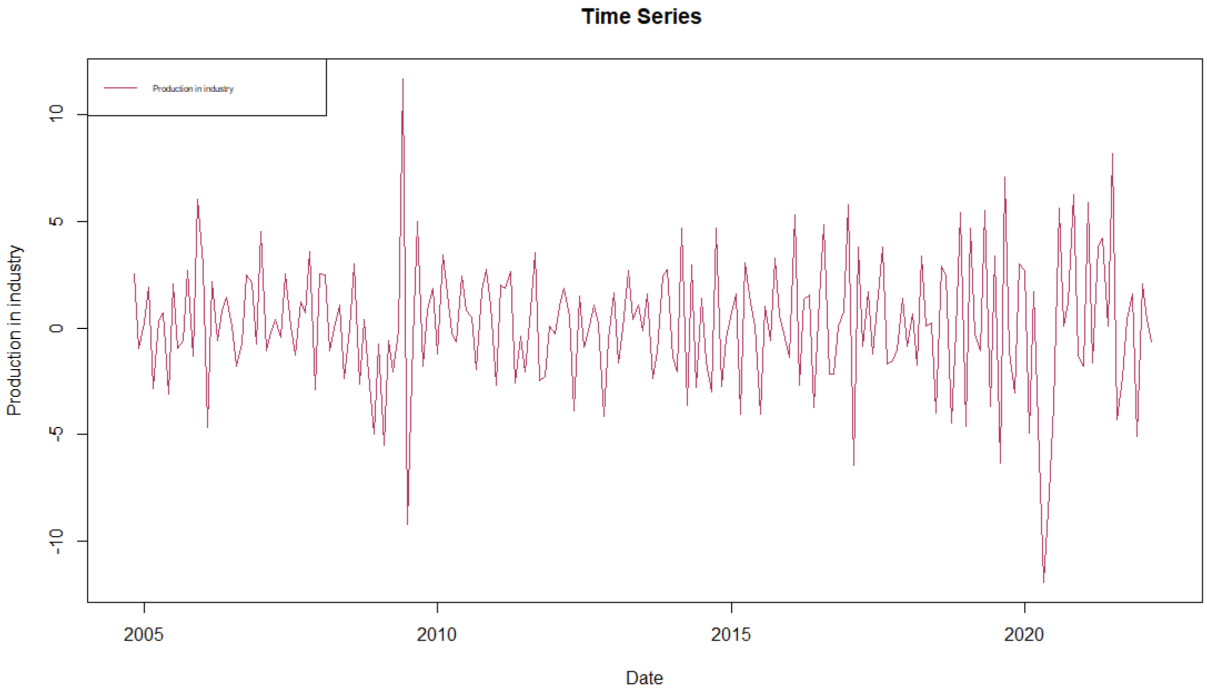


Figure1. Production in Industry as a Time Series Plot for Belgium.

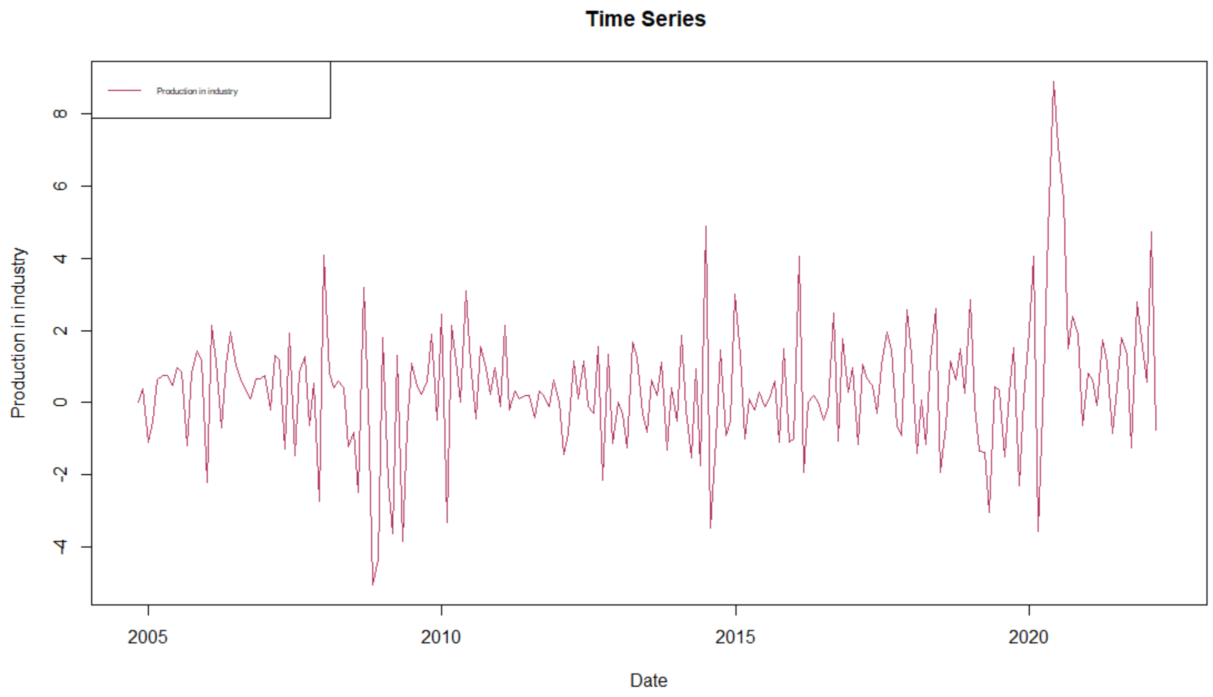


Figure 2. Production in Industry as a Time Series Plot for Austria.

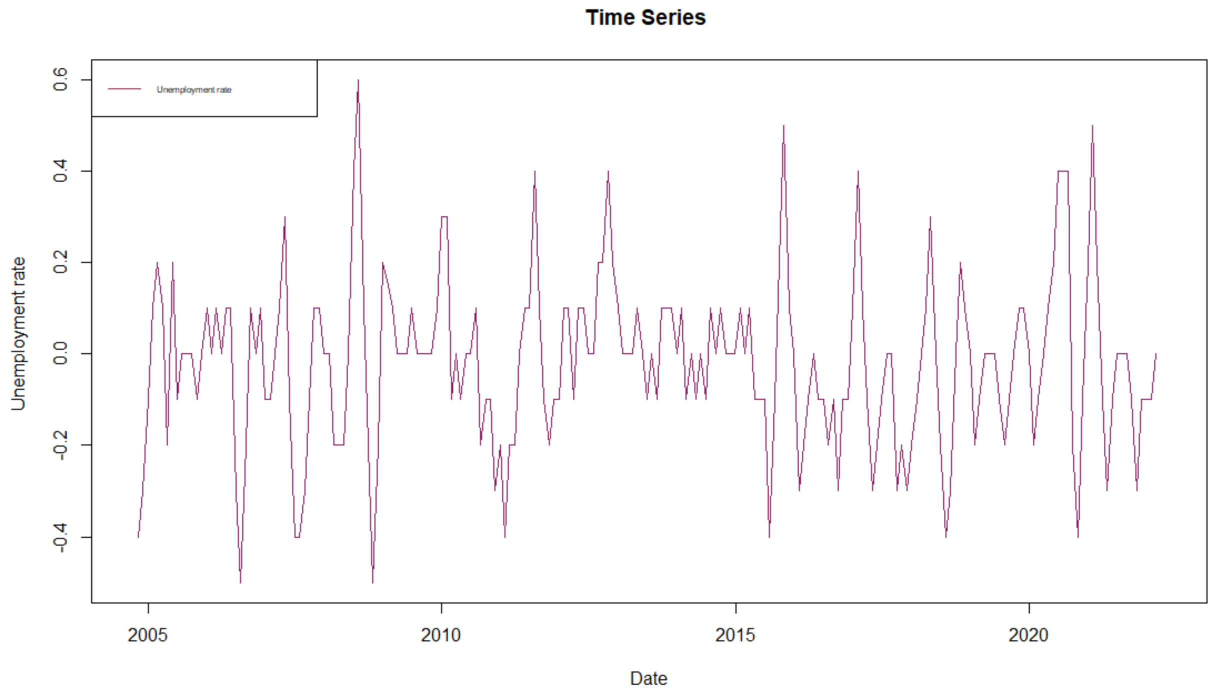


Figure 3. Unemployment rate as a Time Series Plot for Belgium.

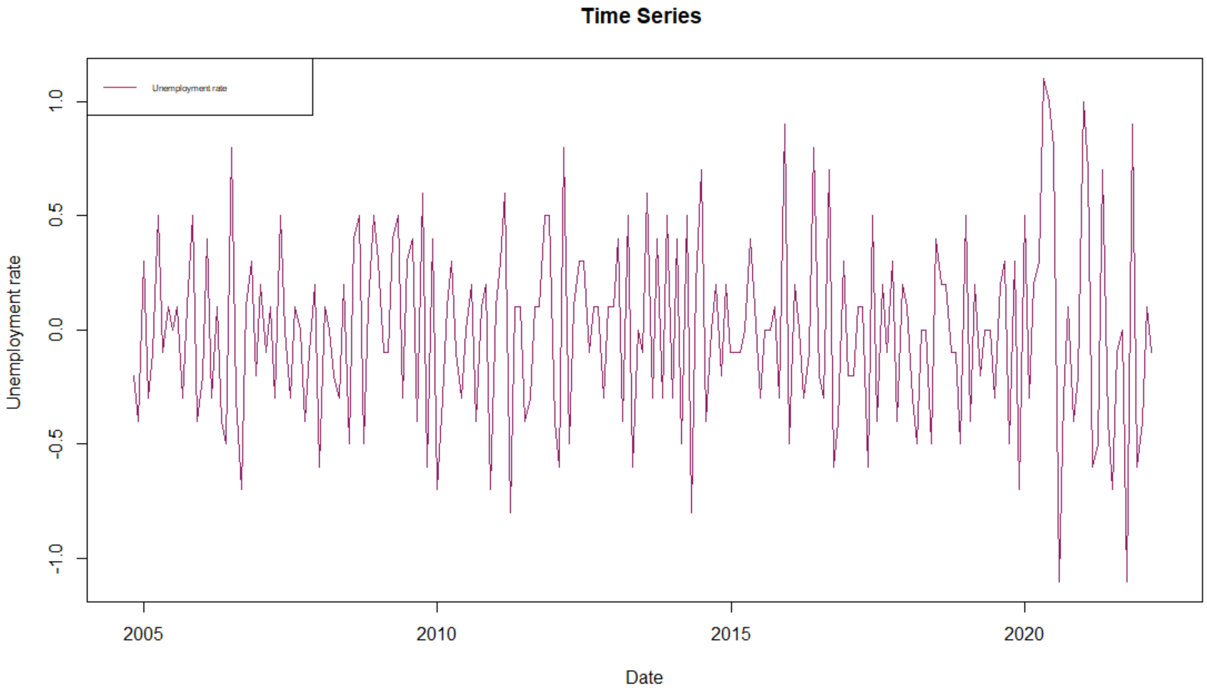


Figure 4. Unemployment rate as a Time Series Plot for Austria.

In Figure 1 and 2, we can observe how the Production in Industry in Belgium and Austria has changed over time. Except for a few occasions, we see a relatively steady variation in both countries. The first date that appears to be a significant drop in both countries is the date of the global financial crisis of 2008. In Belgium more specifically in 2008-2009 two of the country's largest banks - Fortis and Dexia - started to face severe problems, exacerbated by the financial problems hitting other banks around the world. In Austria, on the other hand, something similar may not have occurred, but the global economic crisis has had an impact on the country's GDP. More specifically, until then, the country's GDP was rapidly increasing, with annual growth rates reaching 3.5 percent in 2006 and 2007. However, in the fourth quarter of 2008, Growth declined. As one would expect, industrial production was impacted by the crisis and declined in both countries, which is why we are seeing this drop. The next significant difference between the two plots occurs shortly after the start of the COVID-19 pandemic in 2020. Following the significant increase in cases , Belgium and Austria, like the rest of Europe, went into lockdown after March 2020, when the first cases occurred. As was to be expected, this had a significant and negative impact on both countries' economy, along with industrial production.

Figures 3 and 4 show how the unemployment rate changes over time. We can see that the conduct of the two countries differs in these two graphs. There is a considerable price fluctuation in Belgium, and it is also worth noting that it is inversely linked to the status of the economy, as demonstrated by the fact that the rate increased after the 2008 financial crisis, as well as during the pandemic after 2020. Austria, on the other hand, appears to have a substantially lower unemployment fluctuation rate. There is no noticeable differential when the pandemic of 2020 is excluded, which is still inversely related to the economic position.

The *histogram* is another technique to visualize data. A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins. It is used to summarize discrete or continuous data that are measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form.

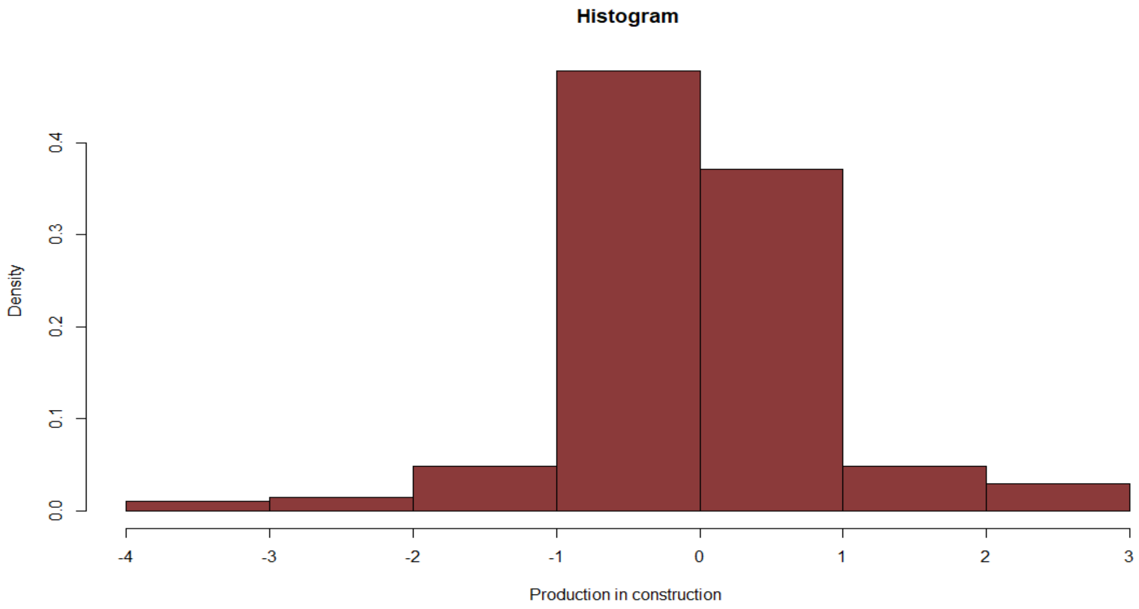


Figure 5.Production in construction of Belgium.

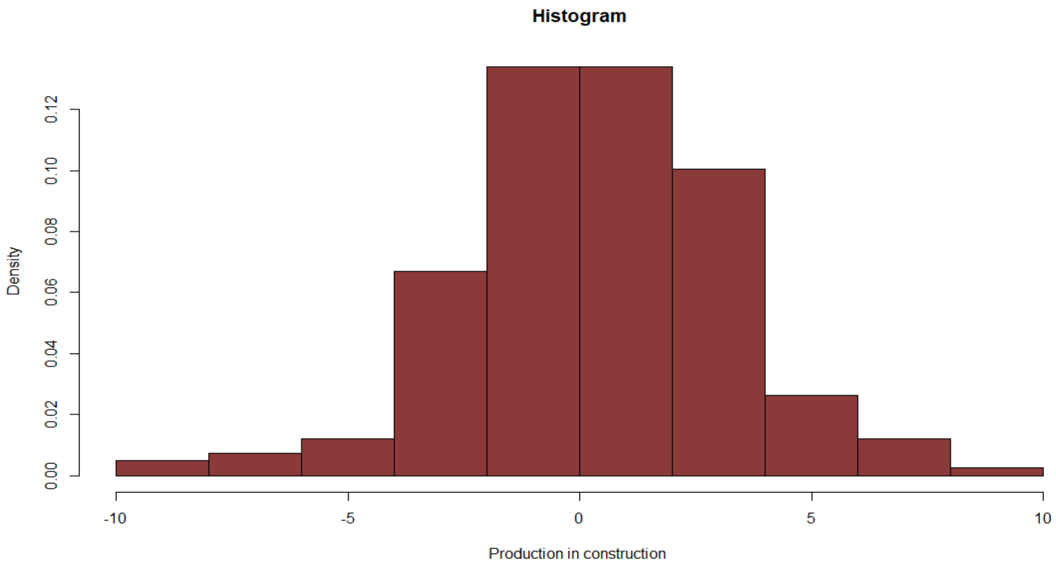


Figure 6.Production in construction of Austria.

The prices of construction production are grouped and dispersed in both countries, as shown in figures 5 and 6. In Belgium, prices for construction production range primarily between -1 and 1, with the majority of prices close to 0, but there are some values at -4 and 3. In comparison, in Austria, prices follow a more normal distribution and there are fewer extremes, with prices ranging from -10 to 10 with most remaining near to 0.

Boxplots are another technique to view data and obtain information about it. In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. In addition to the box on a box plot, there can be lines (which are called whiskers) extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the box-and-whisker plot and the box-and-whisker diagram. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. Box plots can be drawn either horizontally or vertically.

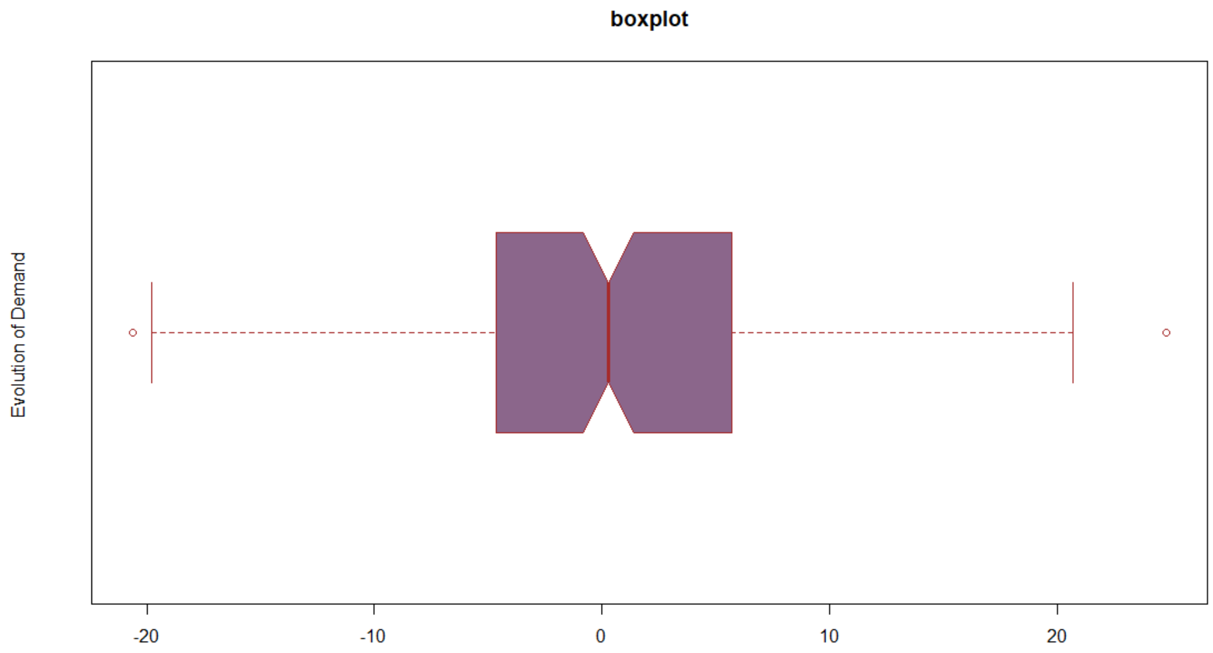


Figure 7.Boxplot of Evolution of Demand for Belgium

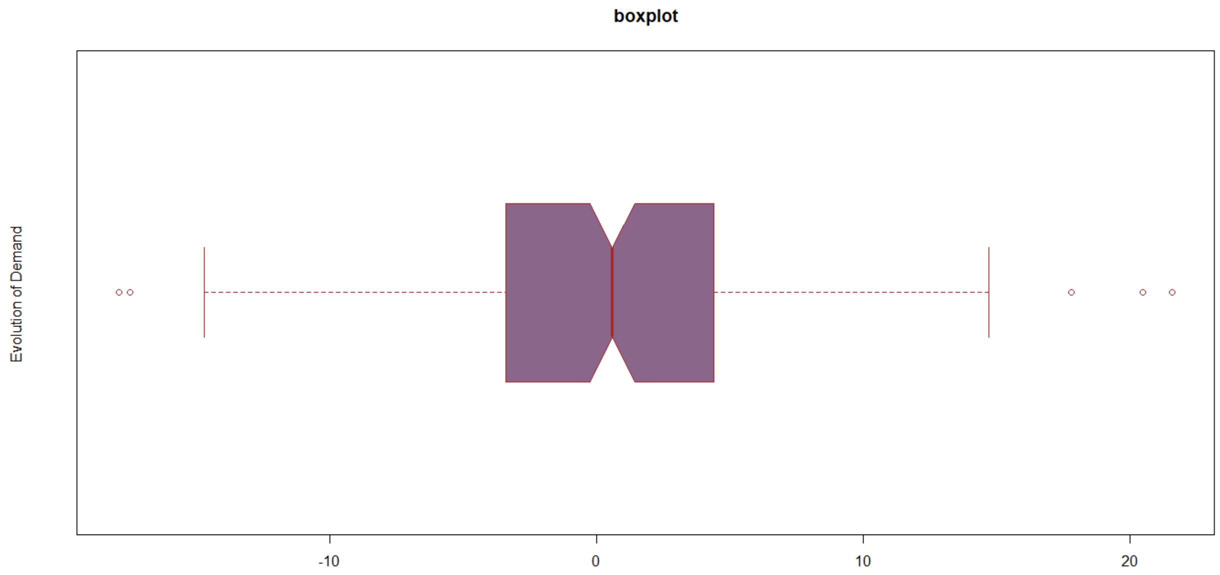


Figure 8.Boxplot of Evolution of Demand for Austria

We can see how the values of Evolution of Demand behave in the boxplots for the two countries in figures 7 and 8. In Belgium, we see that 50% of the Evolution of Demand values are between -4 and +5, with a median slightly less than 0. Within the boundaries, the smallest value is close to -17, while the greatest value is close to +17. We see 2 outliers when we go outside the bounds. In Austria, we see that prices fluctuate similarly to those in Belgium. The median price is slightly higher than 0, with 50% of the values ranging between -4 and +5. Finally we have five outliers again, but this time the majority of these prices are positive rather than negative, as in Belgium.

The study of how a variable might be related to itself in the past or to another variable is of particular importance, and this will be covered in this field of analysis. To begin, the *AR (1) model* will be examined to see how its visualization benefits in better understanding the behavior of a variable. In a multiple regression model we forecast the variable of interest using a linear combination of predictors. In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself.

Thus, an autoregressive model of order p can be written as

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t \ ,$$

Where $\varepsilon_t$ is white noise. This is like a multiple regression but with lagged values of $y_t$ as predictors. We refer to this as an AR(p) model, an autoregressive model of order p. Autoregressive models are remarkably flexible at handling a wide range of different time series patterns. Changing the parameters $\varphi_1 ... \phi_p$ results in different time series patterns. The variance of the error term $\varepsilon_t$ will only change the scale of the series, not the patterns.

For an AR(1) model:

- When $\phi_1=0$ and $c=0$, $y_t$ is equivalent to white noise;
- When $\varphi_1=1$ and $c=0$,$y_t$ is equivalent to a random walk;
- When $\varphi_1=1$ and $c\neq 0$ , $y_t$ is equivalent to a random walk with drift;
- When $\varphi_1<0$ ,$y_t$ tends to oscillate around the mean.

We normally restrict autoregressive models to stationary data, in which case some constraints on the values of the parameters are required.

- For an AR(1) model : $-1< \varphi_1< 1$.
- For an AR(2) model : $-1< \varphi1< 1,\varphi_1+\varphi_2 <1,\varphi_2-\varphi_1 <1$.

When $p\geq 3$, the restrictions are much more complicated.

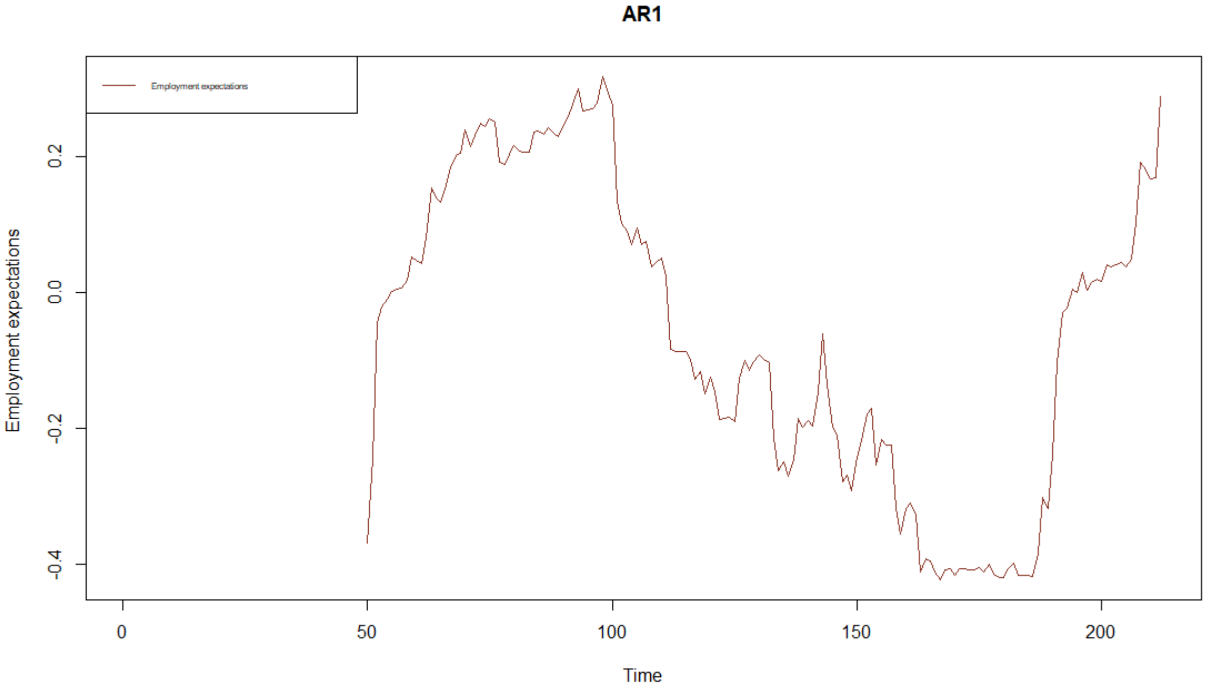


Figure 9. AR(1) model for Employment expectation (SD3) for Belgium

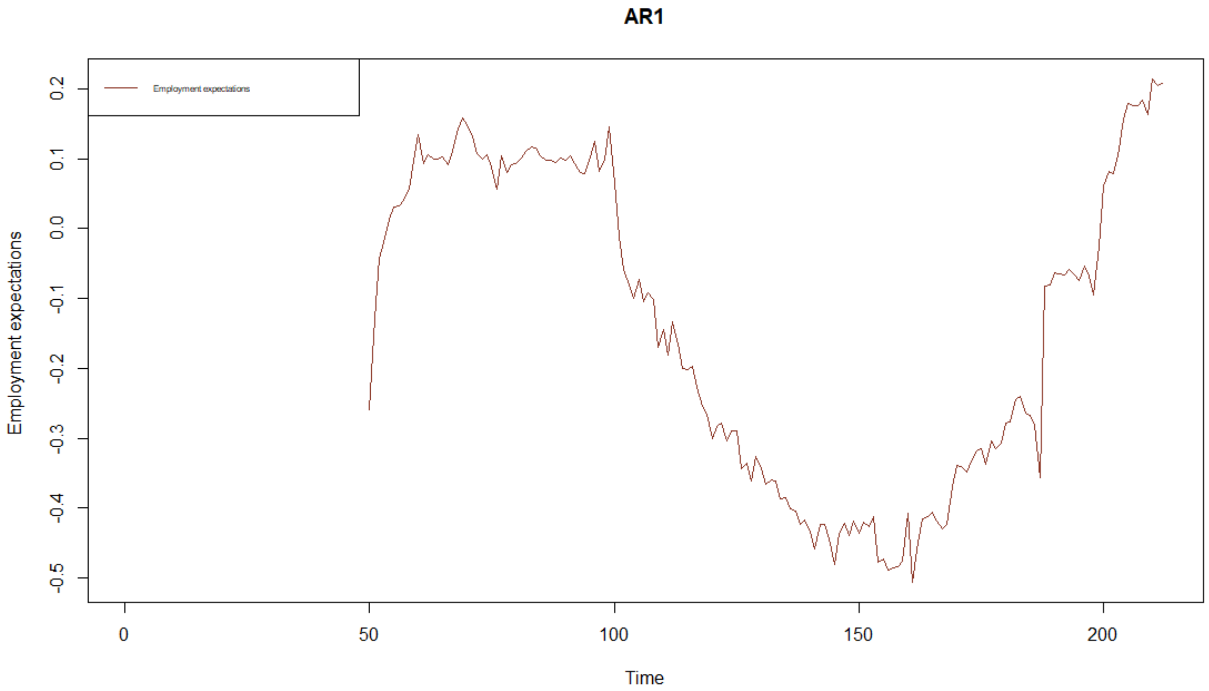


Figure 10. AR(1) model for Employment expectation (SD3) for Belgium

The first thing we observe in Figures 9 and 10 is that the variable Employment Expectations moves remarkably similarly in both countries. The first 50 values of the variables are utilized as the "past," and we can see how these first variables influence the subsequent variables. In both countries, we observe that the first prices are heavily influenced by the recent past, but that as we add variables, the influence drop sharply, and while there is a lot of persistence at first, then it fades. But what's interesting about the variable's behavior is that following a severe drop, the effect of the prior factors rises sharply again, and persistence rises again.

Apart from looking at how prior values of a variable affect subsequent ones, it's also intriguing to look at how variables can be correlated with one another. In statistics, *correlation* or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Although in the broadest sense, "correlation" may indicate any type of association, in statistics it normally refers to the degree to which a pair of variables are linearly related. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. Formally, random variables are dependent if they do not satisfy a mathematical property of probabilistic independence. In informal parlance, correlation is synonymous with dependence. However, when used in a technical sense, correlation refers to any of several specific types of mathematical operations between the tested variables and their respective expected values. Essentially, correlation is the measure of how two or more variables are related to one another. There are several correlation coefficients, often denoted $\rho$ or $r$, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables.

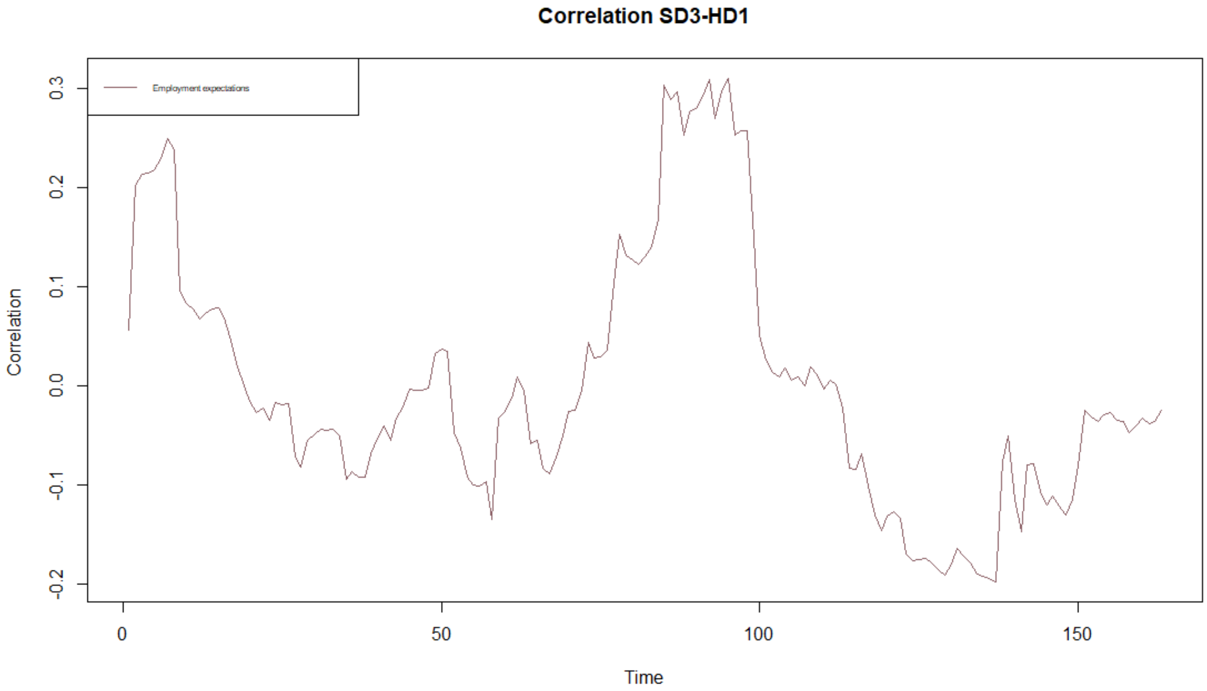


Figure 11.Correlation Expectation of Employment with Production in industry for Belgium.

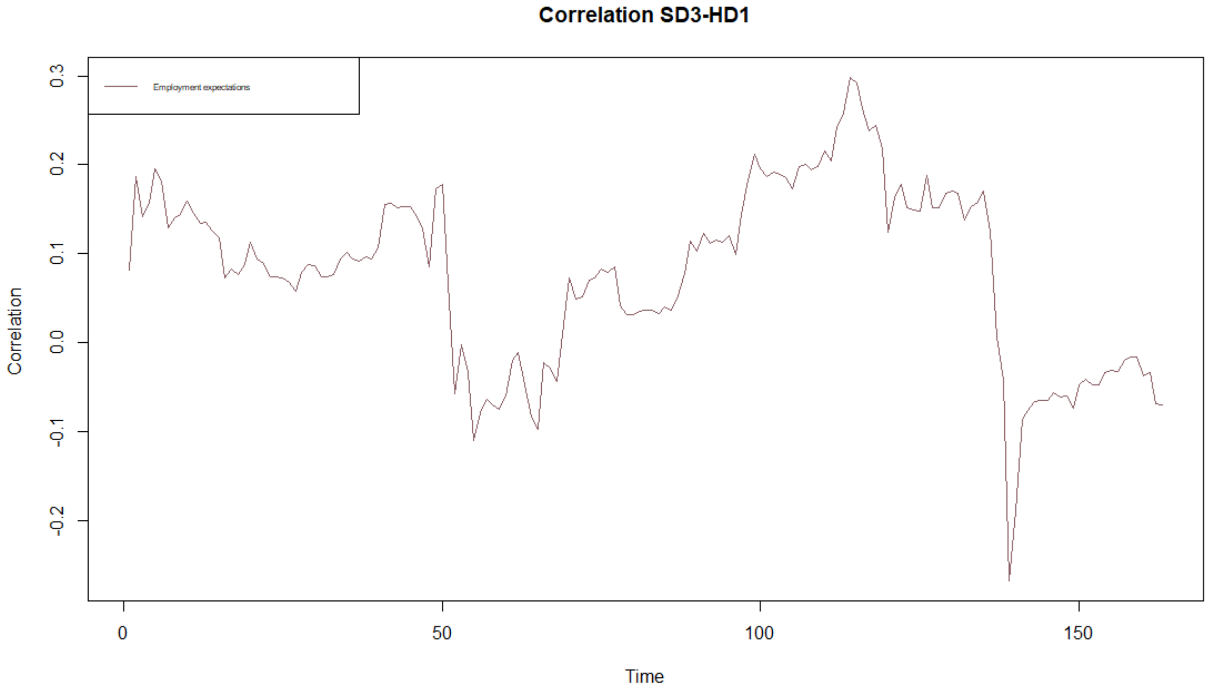


Figure 12.Correlation Expectation of Employment with Production in industry for Austria.

Figures 11 and 12 demonstrate how the variables Expectation of Employment and Production in Industry are correlated in the two countries. In the case of Belgium, we can observe that the correlation of the variables varies wildly. This correlation is mostly positive or close to zero, but it appears that the correlation develops relatively strongly in the first and middle values in a row, implying that Industrial Production had a positive effect on Employment Expectation. Also, for Austria, there has been a considerable fluctuation. And we can see that the prices are mostly positive if we eliminate the most recent prices, where it can be seen a severe drop followed by a sharp rebound. This is most likely because of the Covid-19 pandemic, and this drop appears to be the result of the first lockdown.

*Structural breakpoints* are the next approach to illustrate data that we will discuss in this analysis. In econometrics and statistics, a structural break is an unexpected change over time in the parameters of regression models, which can lead to huge forecasting errors and unreliability of the model in general.

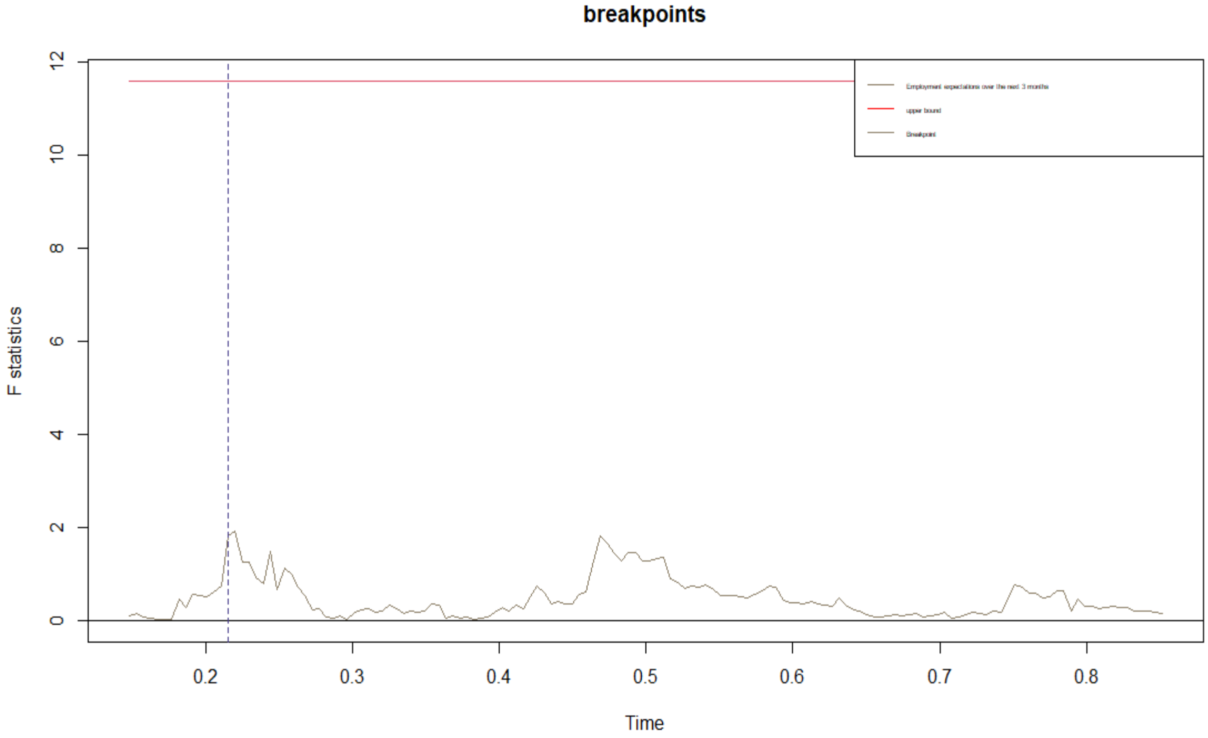


Figure 13.Breakpoint of Consumer Confidence Indicator for Belgium.

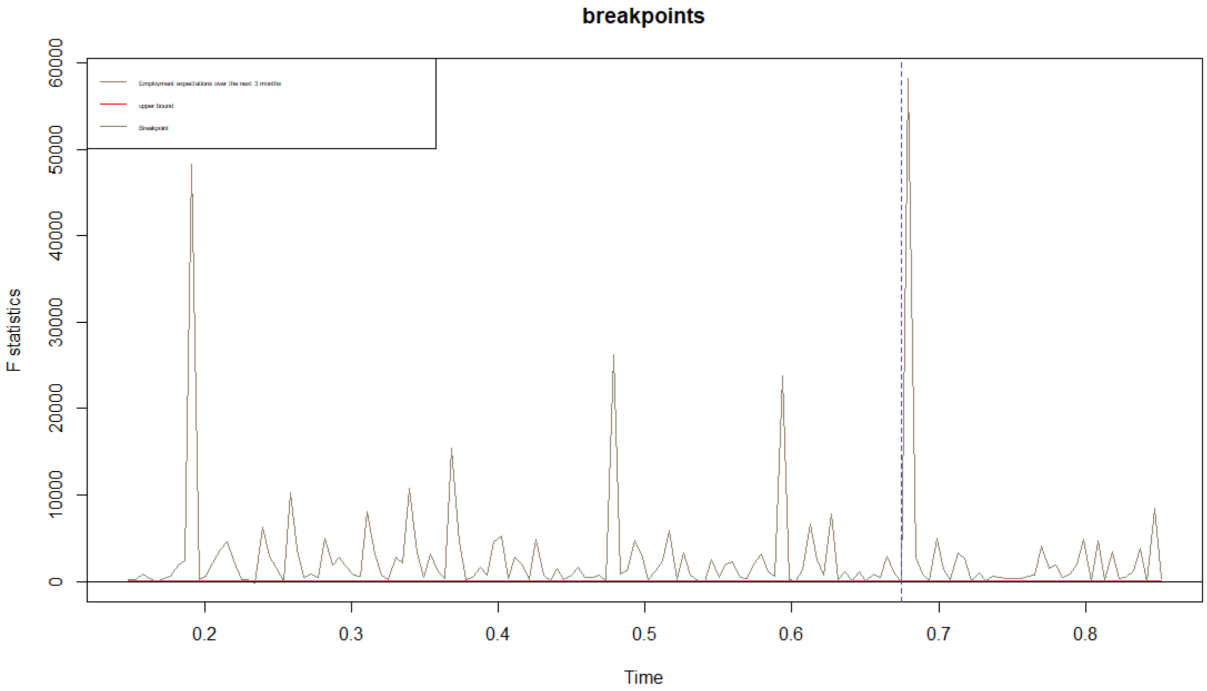


Figure 14.Breakpoint of Consumer Confidence Indicator for Austria.

Finally, Figures 13 and 14 show the structural breakpoints for the variable Consumer Confidence Indicator for the two countries. To make the graphs, we first used the Forward selection approach to identify which hard variables for the Consumer Confidence Indicator variable are statistically significant. The unemployment rate was the single statistically significant variable from hard variables for Belgium, thus we utilized it as an independent variable. As a result, we have the breakpoint shown in Figure 13: the value at position 50. The only hard variable that was statistically significant in Austria, on the other hand, was the Production in Industry variable. As a result, we have the breakpoint shown in Figure 14: the value at position 177.

There are still a lot of different techniques to display data, and not all of them can be discussed here. In the continuation of the analysis, further graphs will be employed in the context of various materials.

As previously said, one of the goals of this study is to do an economic analysis of Belgium and Austria. When there are many variables to examine and an overall estimate of the financial condition is required, this can become very difficult to be accomplished. The goal of this part of the analysis is to obtain a low-dimensional set of features from a large range of variables using Principal Components Analysis (PCA).

## Principal Components Analysis

When faced with a large set of correlated variables, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set.
*Principal component analysis* (PCA) refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach, since it involves only a set of features $x_1$, $x_2$,...,$x_p$ , and no associated response Y . Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization (visualization of the observations or visualization of the variables).PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p features .The first principal component of a set of features $x_1$, $x_2$,...,$x_p$ is the normalized linear combination of the features

$$\mathbf{Z_1 = \varphi_{11}x_{11}+ \varphi_{21}x_{21} + ... + \varphi_{p1}x_p}$$

that has the largest variance. We refer to the elements $\varphi_{11}$,...,$\varphi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\varphi_1$ = ($\varphi_{11}$ $\varphi$ $_{21}$ ... $\varphi$ $_{p1}$)T . We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

After the first principal component $Z_1$ of the features has been determined, we can find the second principal component $Z_2$. The second principal component is the linear combination of $x_1$,...,$x_p$ that has maximal variance out of all linear combinations that are uncorrelated with $Z_1$. The second principal component scores $Z_{12}$, $Z_{22}$... $Z_{n2}$ take the form

$$\mathbf{Z_{i2} = \varphi_{12}x_{i1} + \varphi_{22}x_{i2}+ ... + \varphi_{p2}x_{ip}}$$

where $\varphi_2$ is the second principal component loading vector, with elements $\varphi_{12}$, $\varphi_{22}$,...,$\varphi_{p2}$ . And so on with the other components. The maximum number of components we can make is the same as the number of variables.

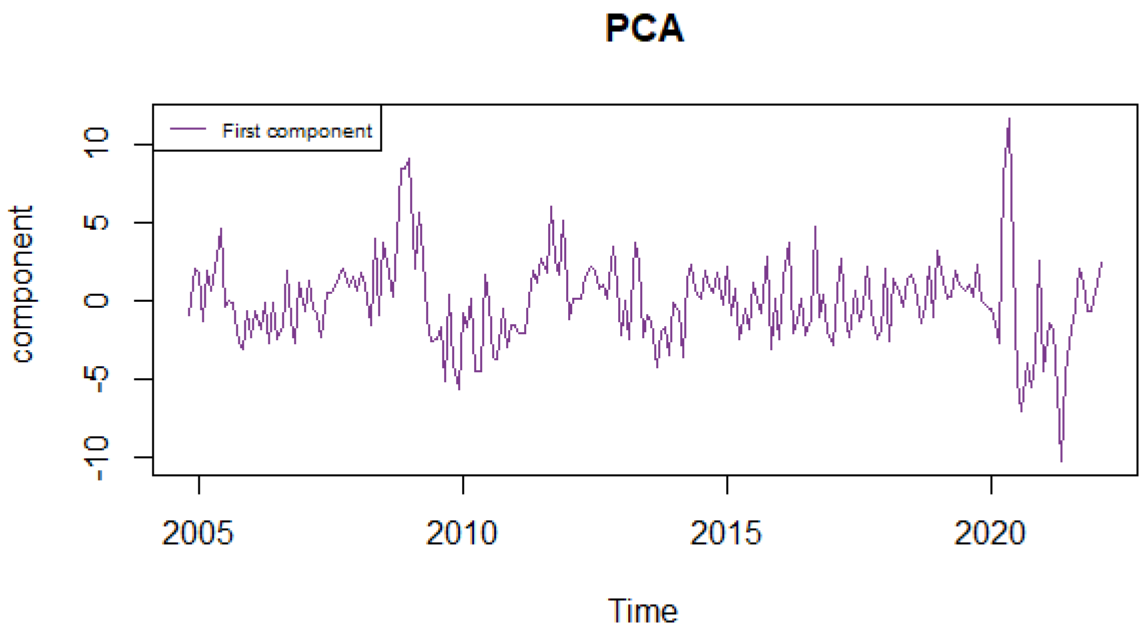


Figure15. The first principal component for Belgium.

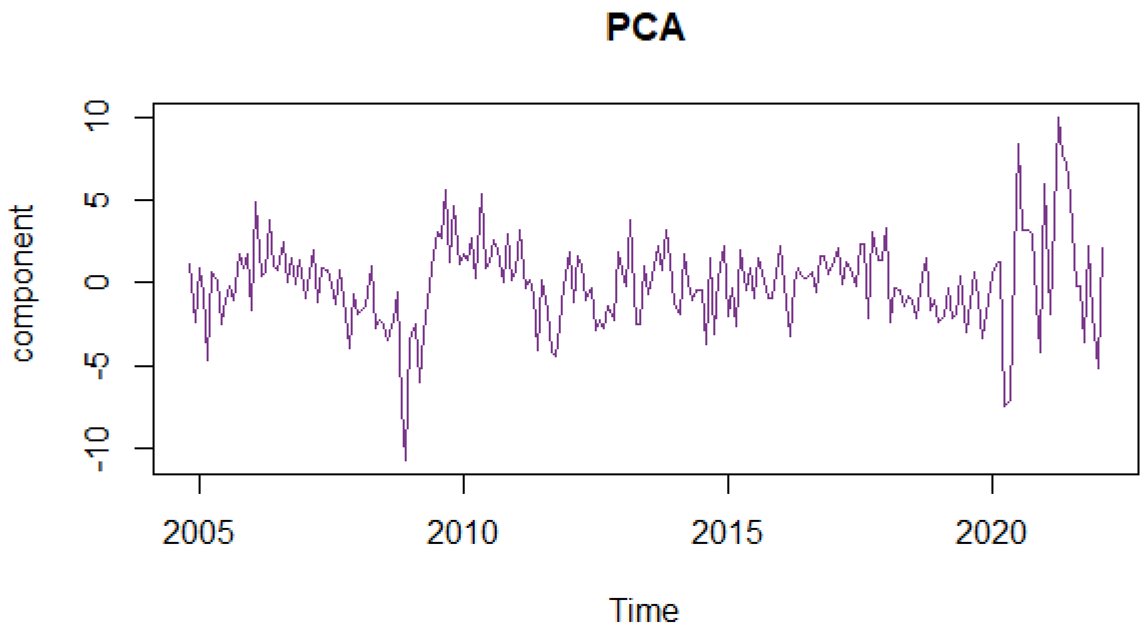


Figure16. The first principal component for Austria.

We can see how the first principal component changes over time in Figure 15 and 16 for both countries. Starting with Belgium, we can see in the graph that the first dip, as seen in the graph, occurs between 2008 and 2012, and is related to the Belgian financial crisis of 2008–2009. Two of the country's largest banks, Fortis and Dexia, began to face major issues, with the value of their stocks plunging, exacerbated by the global financial crisis of the bank. Bailouts, bank sales or nationalizations, bank guarantees and deposit insurance extensions were all used by the government to address the problem. The coronavirus outbreak in Belgium in 2020 is the second significant drop we observe. The coronavirus (COVID-19) epidemic has had a direct impact on Belgium's economy. For example, in 2020, the tourism industry in Flanders and the Brussels-Capital Region is expected to lose 1.7 billion euros in income. Belgium's government balance as a percentage of GDP fell by 11 percent in that year, while GDP per capita fell by 0.4 percent.Austria, on the other hand, had a severe decline in the graph during the 2008 financial crisis, which is reflected in the country's GDP, as previously indicated. The country's GDP had been increasing until then, with yearly growth rates reaching 3.5 percent, but growth slowed in 2008. What's interesting about the Austrian graph is that, unlike Belgium, where we notice a decline in the graph when the pandemic first appears in early 2020, it appears to recover fast and does not remain consistently lower. It should be mentioned here that before the time series plots, it appeared that unemployment in Austria did not decline as much as it did in Belgium following the pandemic, which could indicate that the first country had greater pandemic reflexes than the second.

The conclusion we draw from the above is that the first principal component can be seen as a tool for performing "real-time" monitoring of economic activity, the Financial Condition Index or FCI, as it is representative of Belgium's and Austria's financial status throughout time. This index isolates a component of financial conditions uncorrelated with economic conditions to provide an update on financial conditions relative to current economic conditions. It is being measured with the same way as the First Principal Component, and by Figure 15 and 16 we can extract the conclusions that when the Index rises, the overall risk rises too, respectively when it decreases the overall risk decreases as well.The two graphs,17 and 18, of the Financial Condition Index for the two countries are shown below, and as can be seen, they are identical to the graphs of the First Principal Component for the reasons stated.

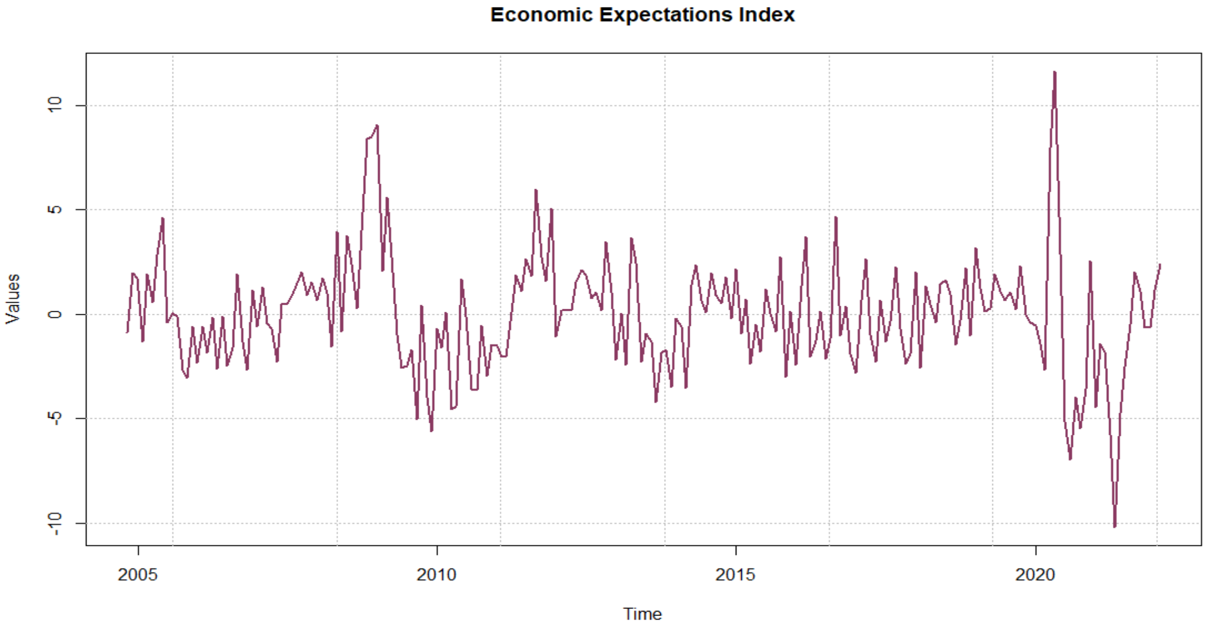


Figure 17. Financial Condition Index for Belgium

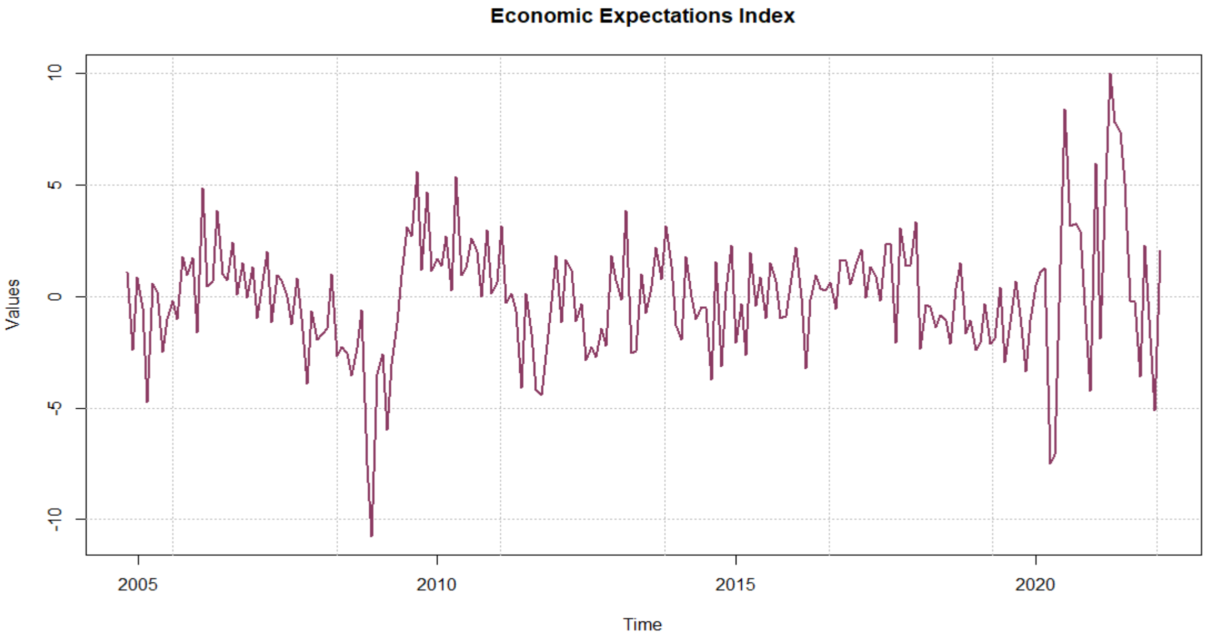


Figure 18. Financial Condition Index for Austria.

We can see how the first principal component behaves over time in Figures 19 and 20 below for both countries. It is added boundaries to tell when a value is out of the average. This is something to be aware of because the variable does not change as it did previously, and something significant could occur. This is backed by the fact that volatility rose during the global financial crisis of 2008–2009, as well as during the coronavirus outbreak in Austria and Belgium in 2020.

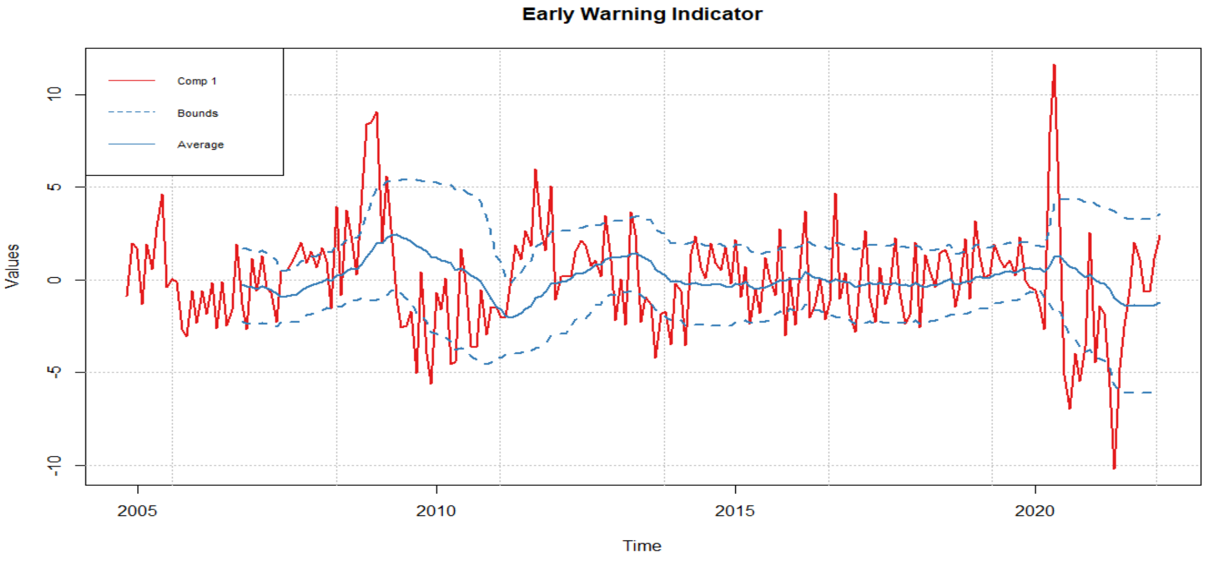


Figure 19. Early Warning Indicator for Belgium

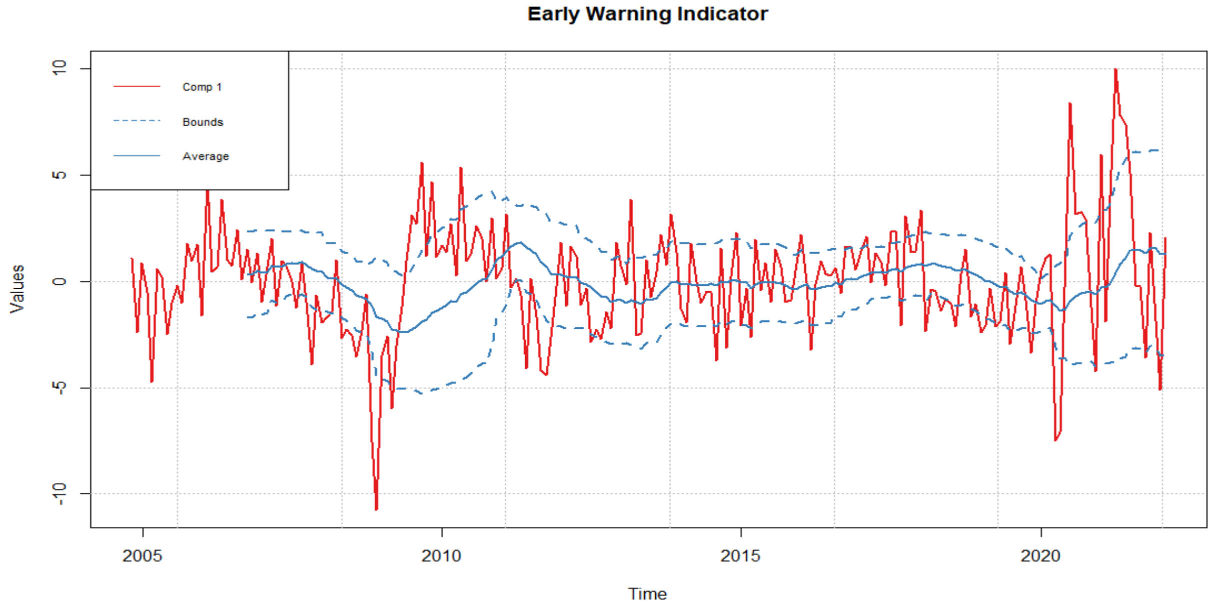


Figure 20. Early Warning Indicator for Austria

| ***Hard Variables*** | **Comp1Belgium** | **\|Comp1\|<br>Belgium** | **Comp1<br>Austria** | **\|Comp1\|<br>Austria** |
|---|---|---|---|---|
| ***Production in industry*** | -0.131 | 0.131 | 0.135 | 0.135 |
| ***Euro area 19 international trade [Exports]*** | -0.095 | 0.095 | 0.109 | 0.109 |
| ***Euro area 19 international trade [Imports]*** | -0.120 | 0.120 | 0.028 | 0.028 |
| ***Production in construction*** | -0.067 | 0.067 | 0.014 | 0.014 |
| ***Unemployment rate*** | -0.019 | 0.019 | -0.018 | 0.018 |
| ***Money market interest rates*** | -0.007 | 0.007 | 0.023 | 0.023 |
| ***Euro yield curves [10 Year Yield]*** | 0.019 | 0.019 | 0.003 | 0.003 |
| ***Turnover and volume of sales wholesale and retail*** | -0.036 | 0.036 | 0.071 | 0.071 |
| ***Euro / ECU [USD]*** | -0.081 | 0.081 | 0.086 | 0.086 |
| ***Spread (10Y-3M)*** | 0.059 | 0.059 | -0.038 | 0.038 |

Table 5.First Component Loadings Hard Variables for Belgium and Austria

| *Soft Variables* | Compo1<br>Belgium | \|Comp1\|<br>Belgium | Comp1<br>Austria | \|Comp1\|<br>Austria |
|---|---|---|---|---|
| *Production development [Industry]* | -0.186 | **0.186** | 0.177 | 0.177 |
| *Production expectations [Industry]* | -0.201 | **0.201** | 0.190 | 0.190 |
| *Employment expectations [Industry]* | -0.184 | **0.184** | 0.199 | 0.199 |
| *Assessment of order-book levels [Industry]* | -0.240 | **0.240** | 0.202 | **0.202** |
| *Assessment of export order-book levels [Industry]* | -0.221 | **0.221** | 0.232 | **0.232** |
| *Assessment of current level of finished products* | 0.123 | 0.123 | -0.145 | 0.145 |
| *Building activity development[Construction]* | -0.151 | 0.151 | 0.117 | 0.117 |
| *Evolution of the current overall order books* | -0.106 | 0.106 | 0.104 | 0.104 |
| *Employment expectations [Constructions]* | -0.129 | 0.129 | 0.058 | 0.058 |
| *Price expectations [Constructions]* | -0.152 | 0.152 | 0.168 | 0.168 |
| *Business activity (sales) development [Retail]* | -0.123 | 0.123 | 0.095 | 0.095 |
| *Volume of stocks currently[Retail]* | 0.091 | 0.091 | -0.049 | 0.049 |
| *Expectations of the number of orders[Retail]* | -0.108 | 0.108 | 0.150 | 0.150 |
| *Business activity expectations[Retail]* | -0.116 | 0.116 | 0.128 | 0.128 |
| *Employment expectations[Retail]* | -0.121 | 0.121 | 0.080 | 0.080 |
| *Business Situation [Services]* | -0.180 | **0.180** | 0.194 | 0.194 |
| *Evolution of Demand [Services]* | -0.126 | 0.126 | 0.200 | **0.200** |
| *Expectation of Demand [Services]* | -0.121 | 0.121 | 0.199 | **0.199** |
| *Evolution of employment[Services]* | -0.105 | 0.105 | 0.031 | 0.031 |
| *Expectation of Employment [Services]* | -0.179 | 0.179 | 0.060 | 0.060 |
| *Euro-zone Business Climate Indicator* | -0.254 | **0.254** | 0.234 | **0.234** |
| *Construction confidence indicator* | -0.149 | 0.149 | 0.106 | 0.106 |
| *Economic sentiment indicator* | -0.317 | **0.317** | 0.336 | **0.336** |
| *Industrial Confidence Indicator* | -0.266 | **0.266** | 0.271 | **0.271** |
| *Retail Confidence Indicator* | -0.158 | 0.158 | 0.136 | 0.136 |
| *Consumer Confidence Indicator* | -0.185 | **0.185** | 0.209 | **0.209** |
| *Financial situation* | -0.034 | 0.034 | 0.042 | 0.042 |
| *Financial situation* | -0.114 | 0.114 | 0.107 | 0.107 |
| *General economic situation* | -0.168 | 0.168 | 0.177 | 0.177 |
| *General economic situation* | -0.173 | 0.173 | 0.217 | **0.217** |
| *Price Trends* | -0.091 | 0.091 | 0.068 | 0.068 |
| *Price Trends* | 0.017 | 0.017 | -0.005 | 0.005 |
| *Unemployment Expectations* | 0.202 | **0.202** | -0.239 | **0.239** |
| *Major purchases* | -0.116 | 0.116 | 0.091 | 0.091 |
| *Savings* | -0.038 | 0.038 | 0.053 | 0.053 |

Table 6.First Component Loadings Soft Variables For Belgium and Austria

Table 5 and 6 shows the loadings of the variables  for the first component for both countries .With bold numbers are the ten variables with the largest influence on the first component; the higher the loading, the larger the influence. The variable, with the highest value of loading is Economic sentiment indicator (ESI).  Economic sentiment indicator (ESI) is a composite indicator produced by the European Commission's Directorate General for Economic and Financial Affairs (DG ECFIN).   Its purpose is to track GDP growth at the Member states, EU, and euro area levels. It stands to reason to have such a large impact on the factor as it concerns GDP, which is a major economic factor in the European economy.

What if we add more Components to the mix? The graphs become smoother as more components are revealed. The fewest number of primary components necessary to explain a considerable degree of variation in the data is chosen. It is possible to conclude from Figures 19 and 20 that two main components are sufficient to explain the variation. In practice, the goal is to remove components until a satisfying pattern emerges, which varies depending on the situation. One can extract as many components as the number of variables, but after the first few major components, there are no noteworthy patterns.

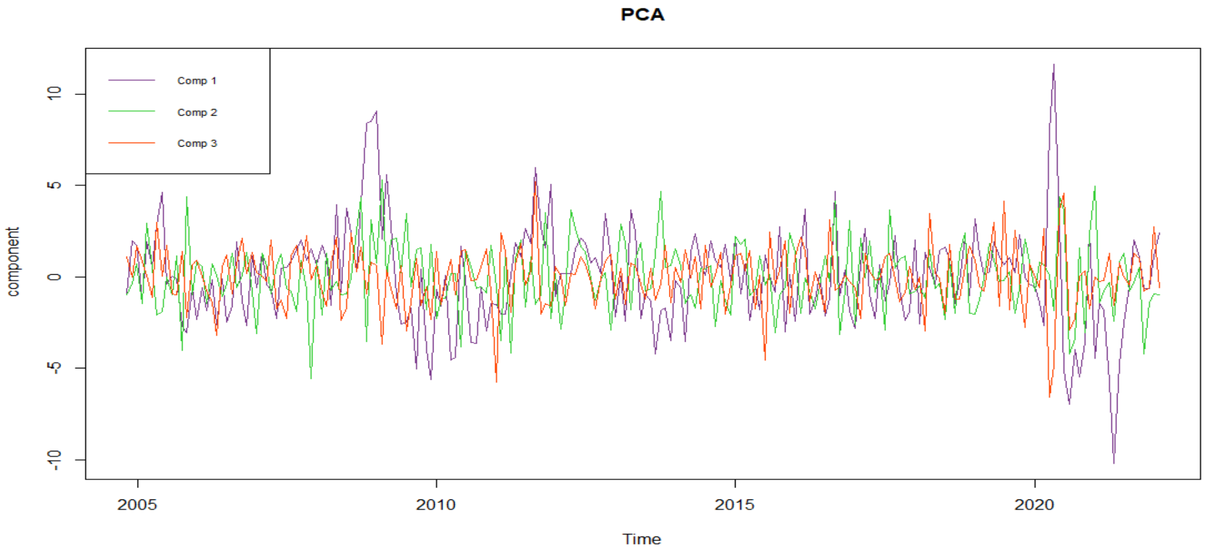


Figure19. The first ,second and third component for Belgium.

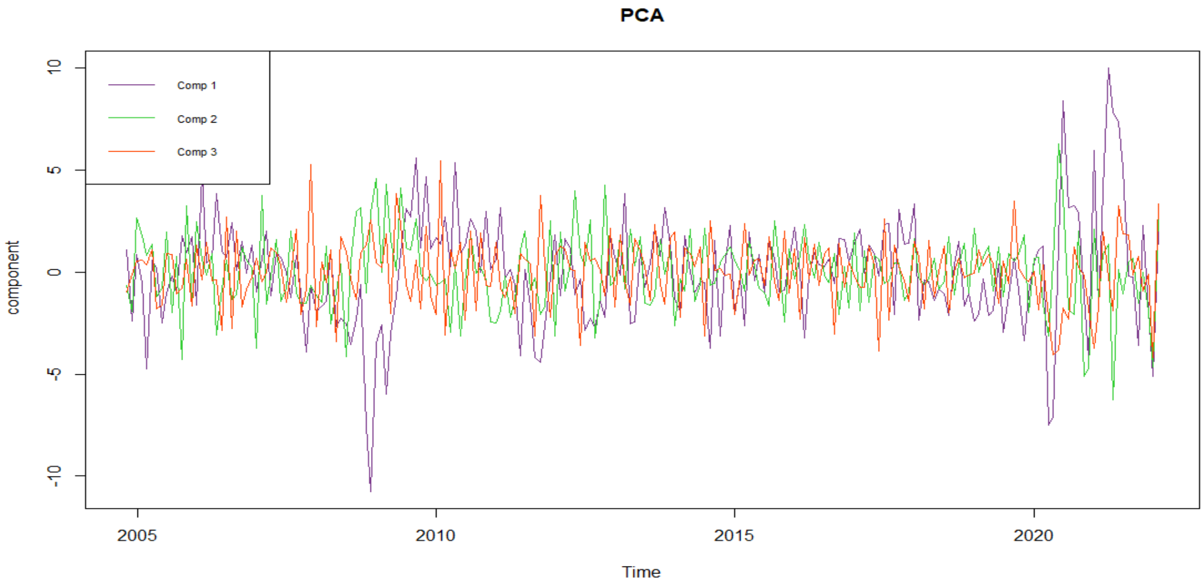


Figure20. The first, second and third component for Austria.

Figure 19 and 20 display the behavior of the first, second, and third components over time for the two countries. We notice that the three components differ from one another, as each component explains a different portion of each variable's variability. The first component explains the most variability, followed by the second, which explains a smaller portion, and finally, the third, which explains an even smaller portion. As a result, the black line has a lot of volatility, while the red line is smoother and the blue line even more so.

We can examine how much each component explains the variability in Figures below. For Belgium and Austria the first component explains 16 % of it, the second component explains 8%, the third and fourth components explain close to 6%, and so on.

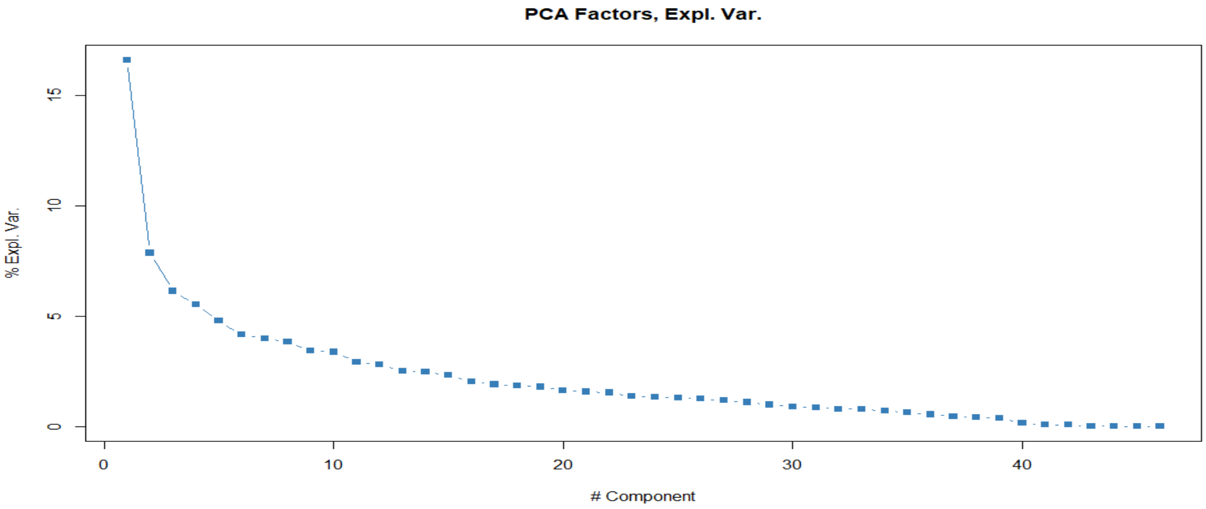


Figure21.How much PCA Factors explain variability for Belgium

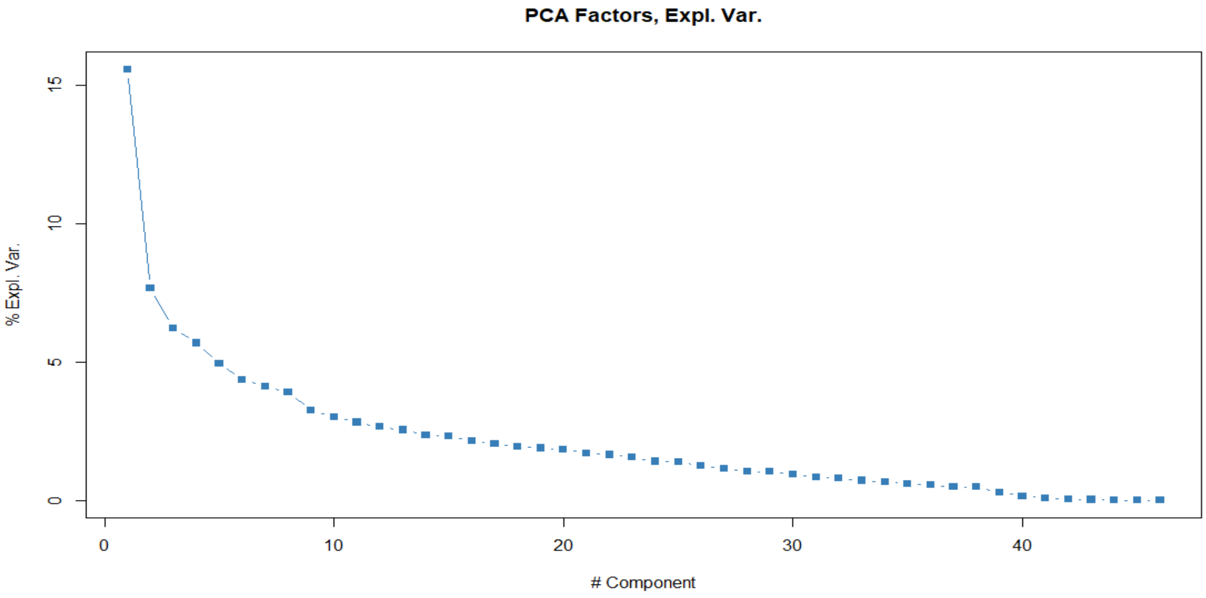


Figure 22.How much PCA Factors explain variability for Austria.

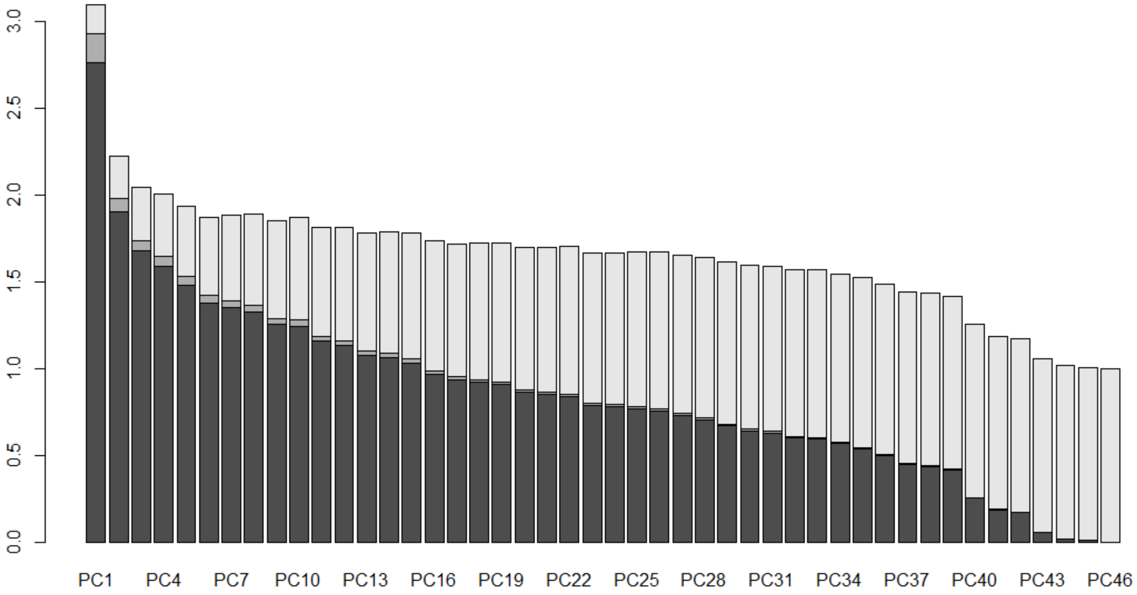


Figure 23.All the components for Belgium .

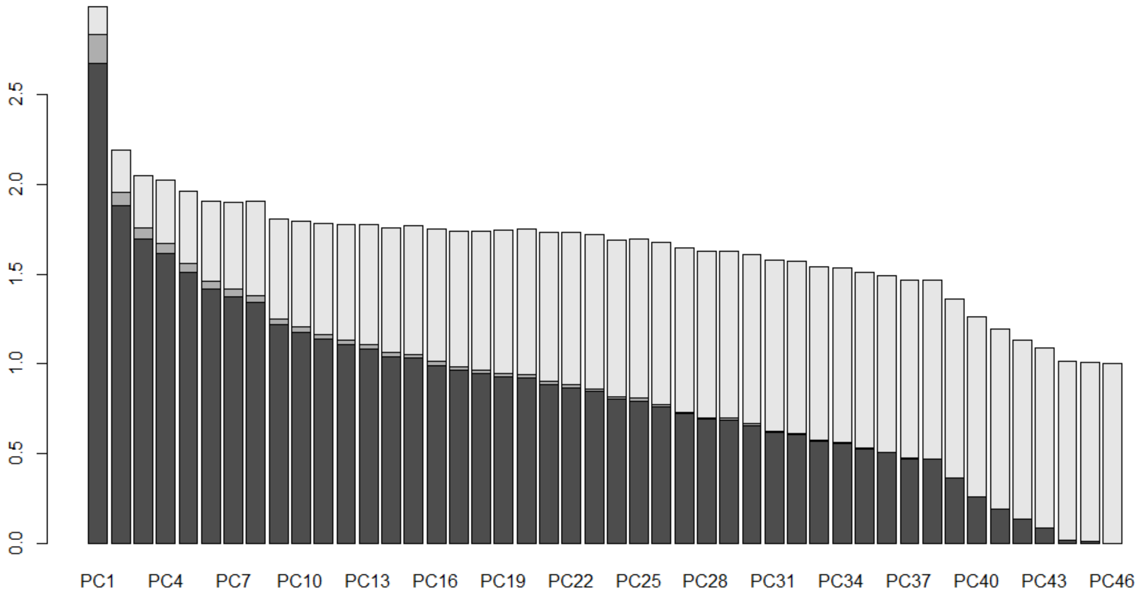


Figure 24.All the components for Austria.

In figures 23 and 24 we observe the Cumulative Proportion of each component separately. Because the intercept alone counts, we have 46 PCAs whereas our variables are 45.

## Partial Least Squares

PCA has many benefits however it has one main disadvantage which is that it is an unsupervised learning method. This means that we extract the combinations of X without taking into account the response Y. To solve this problem, we can use the method of *Partial Least Squares*.

Partial least squares (PLS) is a relatively new method for estimating regression equations, introduced in order to facilitate the estimation of multiple regressions when there is a large, but finite, amount of regressors. The basic idea is similar to Principal Component Analysis (PCA) in that factors or components, which are linear combinations of the original regression variables, are used, instead of the original variables, as regressors. A conceptually powerful way of defining PLS is to note that the PLS factors are those linear combinations of xt, denoted by $Yx_t$,that give maximum covariance between $y_t$ and $Yx_t$ while being orthogonal to each other. Of course, in analogy to PCA factors, an identification assumption is needed, to construct PLS factors, in the usual form of a normalisation.

Assuming for simplicity that $y_t$ has been demeaned and $x_t$ have been normalised to have zero mean and unit variance, a simplified version of the algorithm follows.

1. Set $u_t$= $y_t$ and $v_{i,t}$= $x_i$,t ,i = 1, ...N. Set j = 1

2. Determine the N × 1 vector of indicator variable weights or loadings $w_j$ = ($w_{1j}$… $w_N$j)´ by computing individual covariances: $w_{ij}$= Cov($u_t$, $v_{it}$), i = 1, ...,N . Construct the j-th PLS factor by taking the linear combination given by $w_j$´$v_t$ and denote this factor by f $_{j,t}$ .

3. Regress ut and $v_{i,t}$ , i = 1, ...,N on $f_{j,t}$. Denote the residuals of theseregressions by u´$_t$ and v´ $_{it}$ respectively.

4. If j = k stop, else set $u_t$ = u´$_t$, $v_{i,t}$ = v´$_{i,t}$ ,i = 1, ..,N and j = j + 1 and go to Step 2.

This algorithm makes clear that PLS is computationally tractable for very large data sets. Once PLS factors are constructed $y_t$ can be modeled or forecast by regressing yt on $f_{j,t}$ , j = 1, ..., k.

The estimates of the coefficients β in the regression of $y_t$ on $x_t$ , obtained implicitly via PLS Algorithmand a regression of $y_t$ on $f_{j,t}$, j = 1, ..., k, are mathematically equivalent to

$$\widehat{\beta}_{PLS} = V_k(V_k'X'XV_k)^{-1}V_k'X'y$$

with $V_{k_1} = (X'y \quad X'XX'y \quad \cdots \quad (X'X)^{k-1}X'y), X = (x_1 \cdots x_T)'$ and $y = (y_1 \cdots y_T)'$.

A major difference between PCA and PLS is that, whereas in PCA regressions the factors are constructed taking into account only the values of the $x_t$ variables, in PLS, the relationship between $y_t$ and $x_t$ is considered as well in constructing the factors.Also The main practical difference between PCA and PLS is that PCA often needs more components than PLS to achieve the same prediction error.

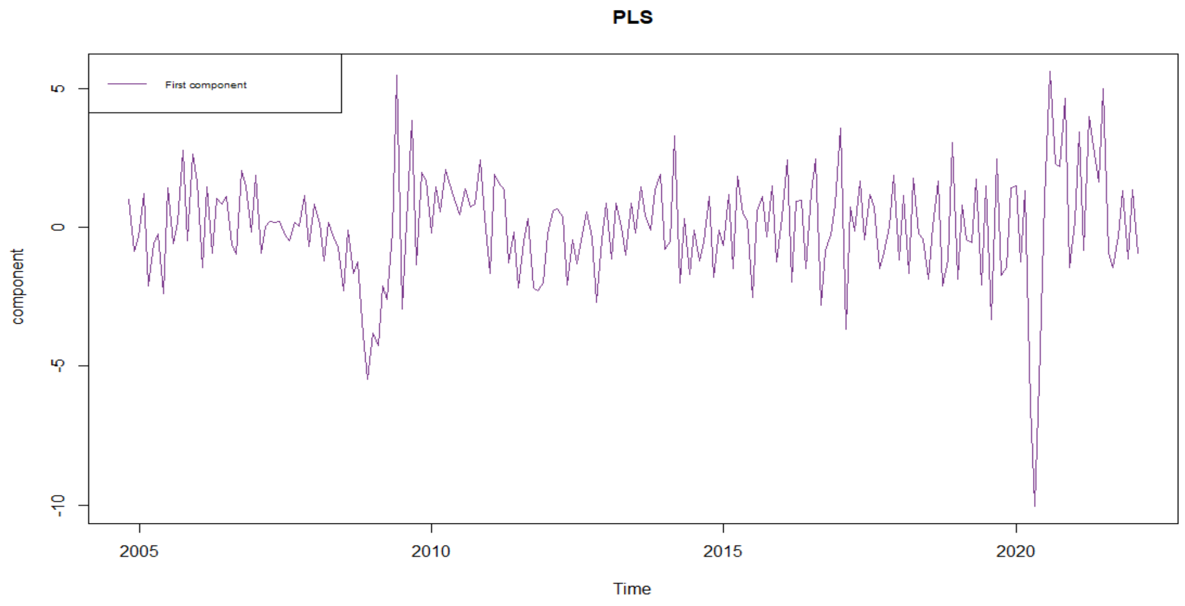


Figure 25. The PLS first principal component for Belgium

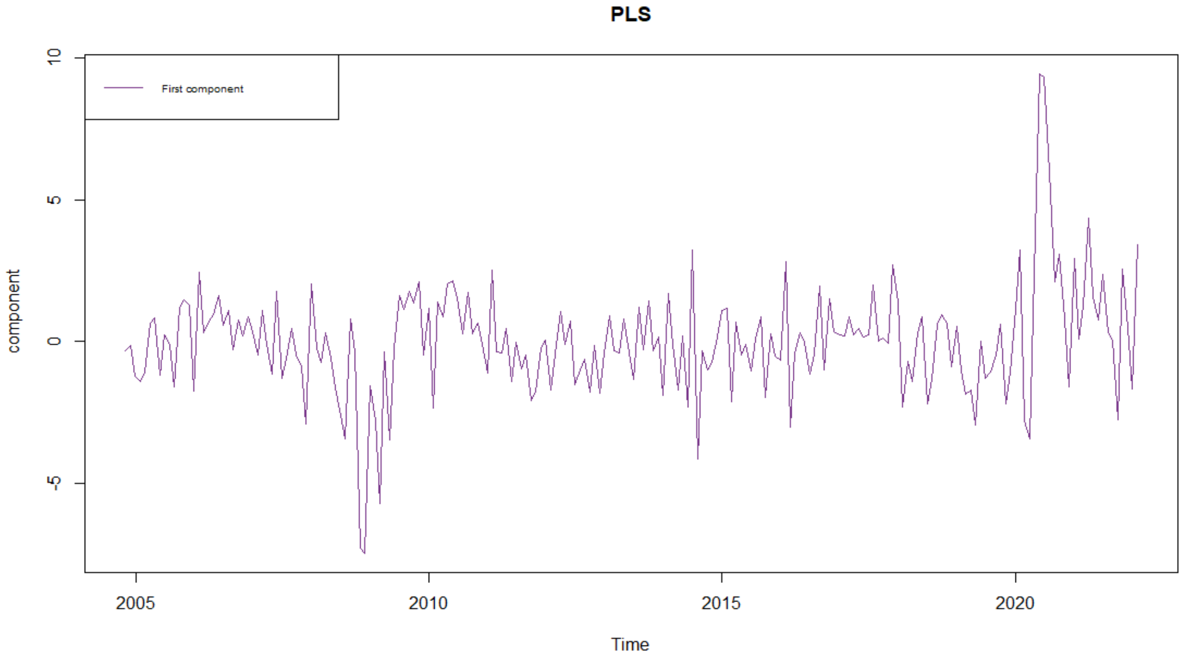


Figure 25. The PLS first principal  component for Belgium

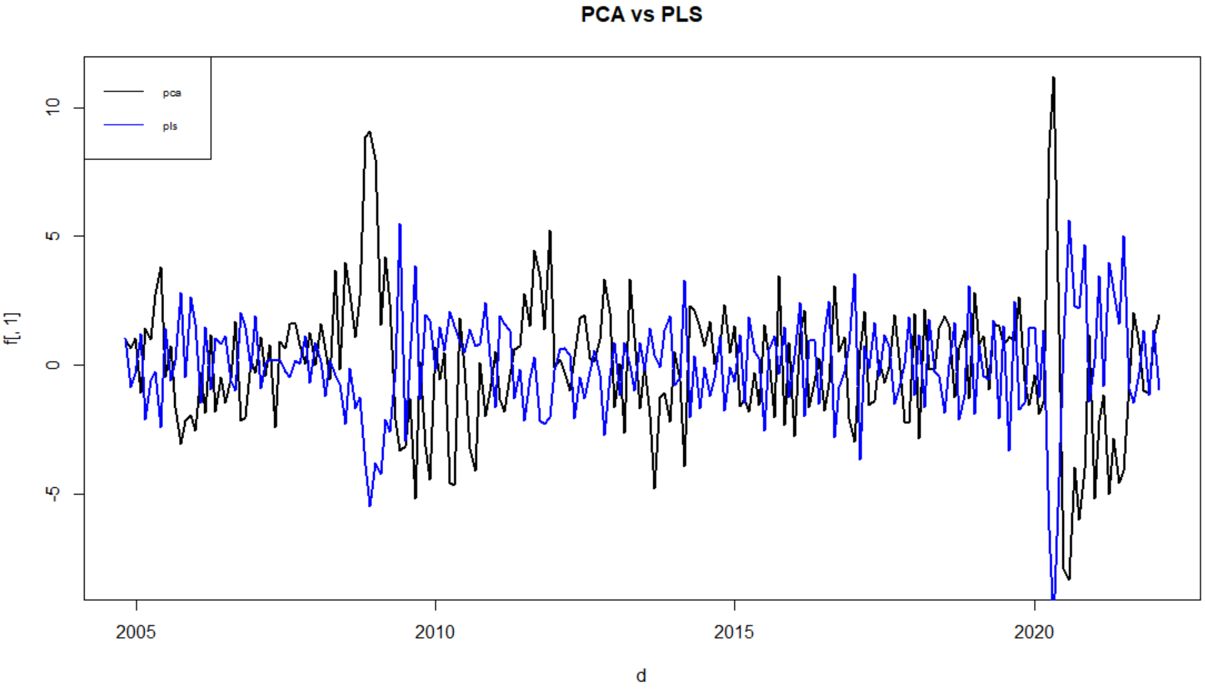


Figure 26. The first principal component PCA vs PLS for Belgium

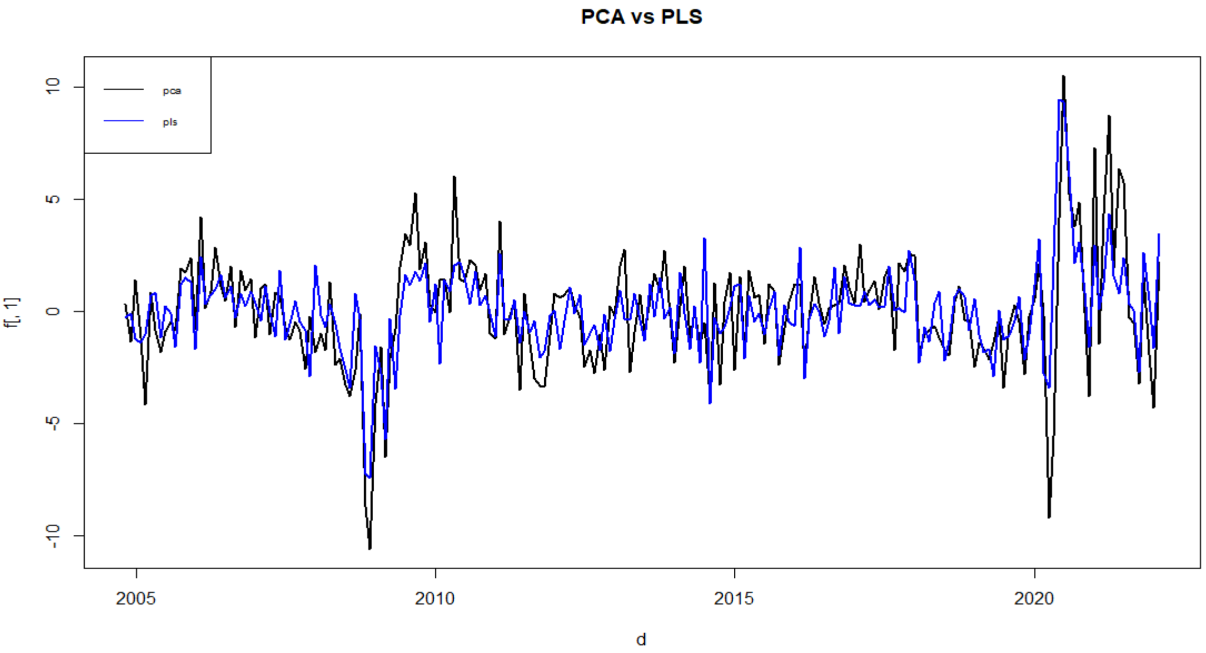


Figure 27. The first principal component PCA vs PLS for Austria

Figures 26 and 27 show how the first principal component of PLS varies from the first principal component of PCA for both countries . It is  observed they are quite different especially for Belgium. T his is explained by the fact that PLS is a supervised learning method and so cannot be used as a financial conditions indicator. Moreover, PLS is attempting to explain a certain Y and will not display real-time information , instead it takes data  and attempts to "guide" it to Y, which in this example is  the variable Industrial Production.

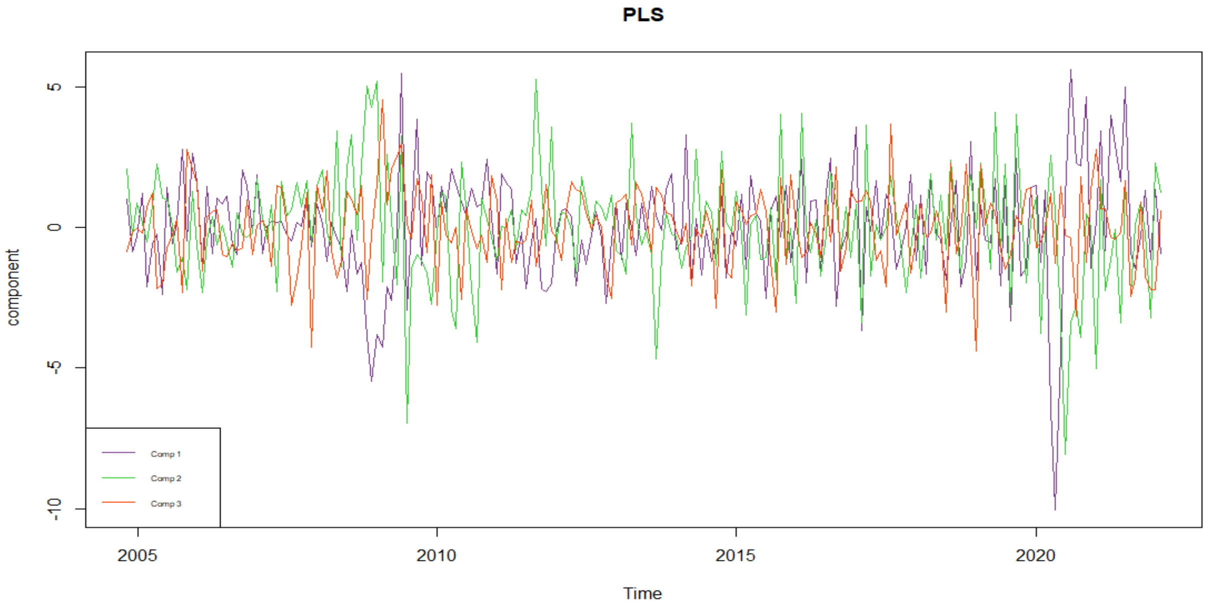


Figure 28. The first ,second and third PLS component for Belgium

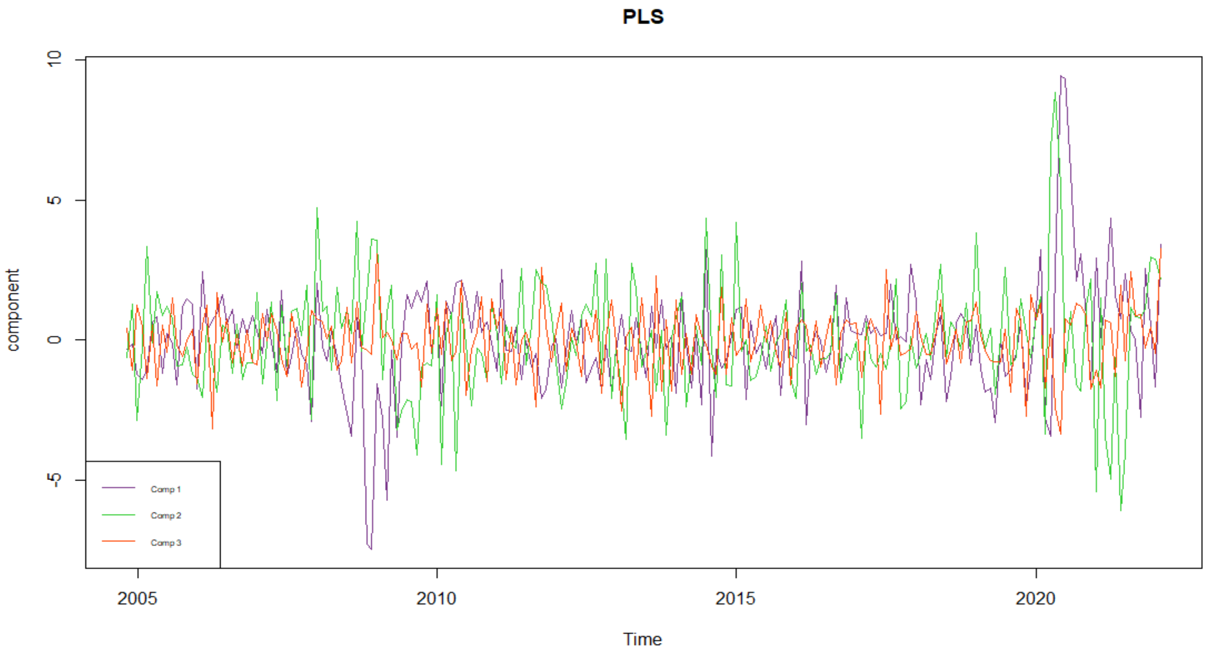


Figure 29. The first ,second and third PLS component for Belgium

The behavior of the first, second, and third components for PLS  through time is visualized in Figure 28 and 29.And again we can see that the three components differ from one another, and that each one is smoother and has less fluctuation than the previous one. This occurs because each component explains a different portion of the variability of each variable.

Let's take it a step further now that we've gathered all of the variables for the creation of PCA. To better comprehend the variables, it is helpful to arrange them together so that the commonalities between them can be seen. The clustering method will be utilized for this grouping.

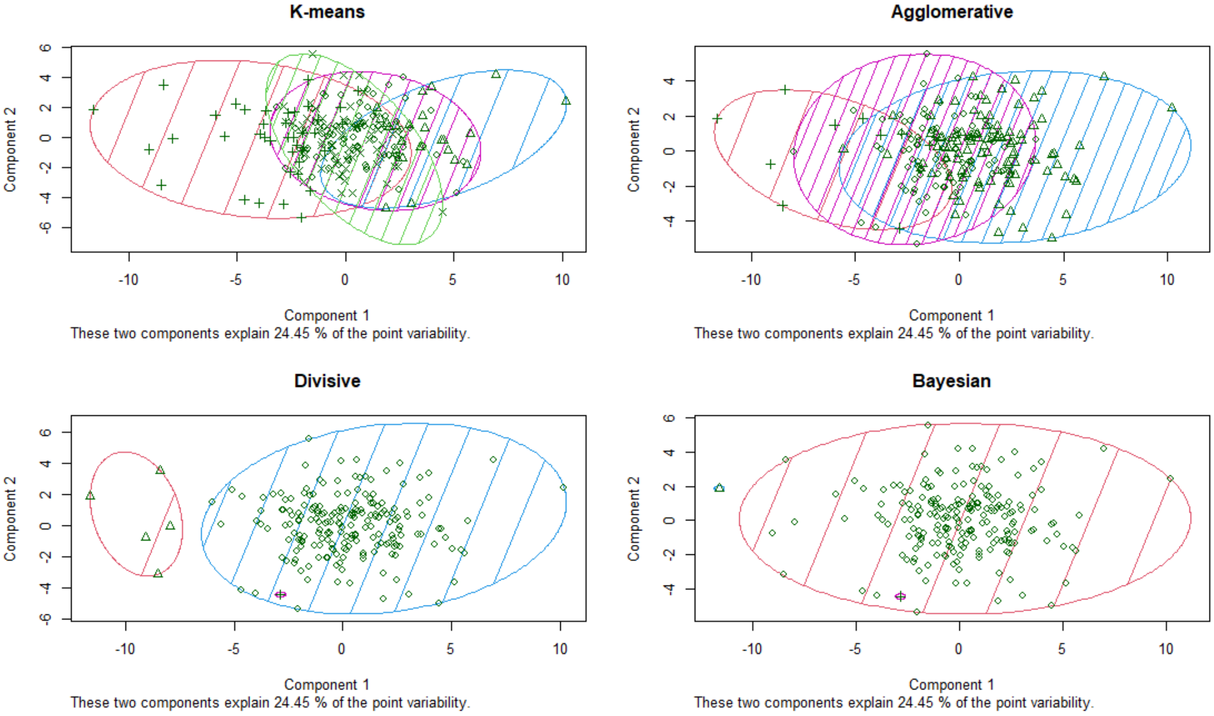
*Clustering* is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated. K-means, agglomerative, Divisive, and Bayesian clustering methods will be used in this study.

*K-means* clustering uses “centroids”, K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it. The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

*Agglomerative Clustering* is a type of hierarchical clustering algorithm. It is an unsupervised machine learning technique that divides the population into several clusters such that data points in the same cluster are more similar and data points in different clusters are dissimilar. The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster.

*The divisive clustering* algorithm is a top-down clustering approach, initially, all the points in the dataset belong to one cluster and split is performed recursively as one moves down the hierarchy. The difference between Agglomerative Clustering and Divisive clustering is that Agglomerative Hierarchical clustering method allows the clusters to be read from bottom to top and it follows this approach so that the program always reads from the sub-component first then moves to the parent. Whereas, divisive uses top-bottom approach in which the parent is visited first then the child.

In a *Bayesian* formulation of a clustering procedure, the partition of items into subsets becomes a parameter of a probability model for the data, subject to prior assumptions, and inference about the clustering derives from properties of the posterior distribution. The difference between K-means Clustering and Bayesian clustering is that K-Means clustering is used to cluster all data into the corresponding group based on data behavior, i.e. malicious and non-malicious, while the Naïve Bayes classifier is used to classify clustered data into correct categories



K-means

Agglomerative

Divisive

Bayesian

Figure 25.Four methods clustering for Belgium

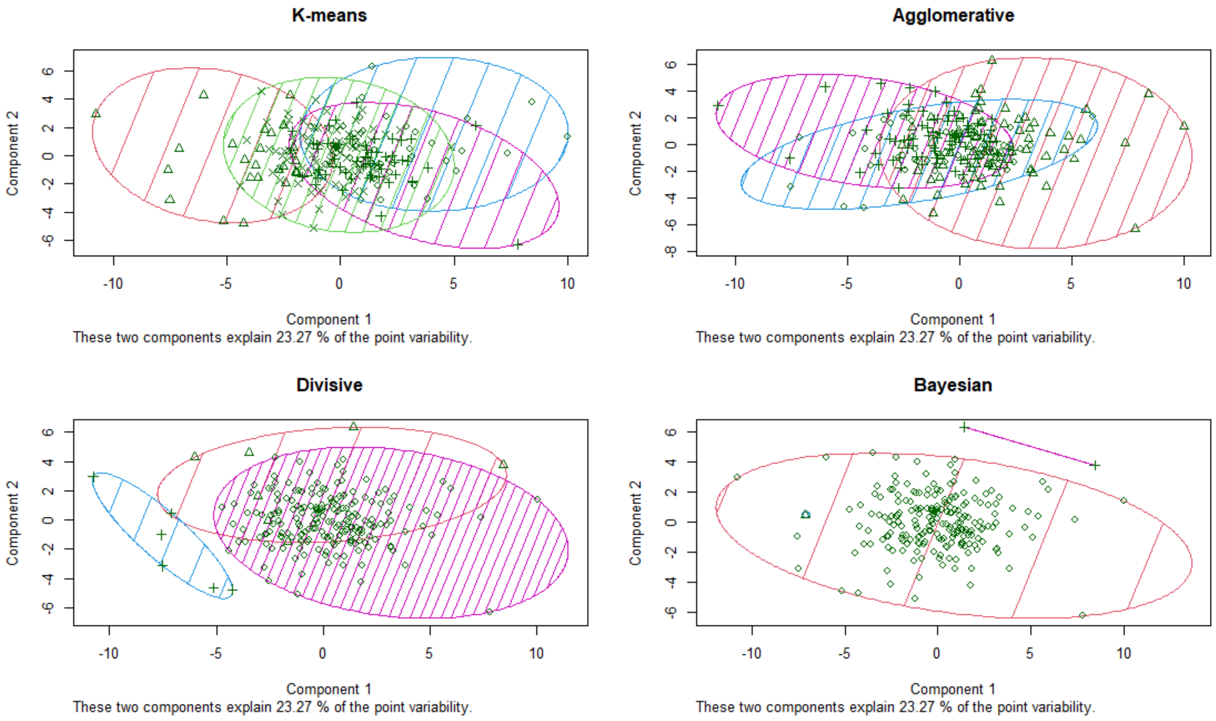


Figure 26.Four methods clustering for Austria

As it appears, in Figures 25 and 26, in both countries, the Divisive method of clustering 45 variables is the most effective. The variable groupings are more distinct. Bayesian is the next most effective method, followed by K-means, and then Agglomerative. The centers are defined to 4 in k-means. This number was chosen since the Elbow technique has already been employed. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. Despite the fact that this method was used to define clusters, it appears that it does not operate.

# Variable Selection

*Variable selection* means choosing among many variables which to include in a particular model, that is, to select appropriate variables from a complete list of variables by removing those that are irrelevant or redundant. The methods we will use to select the variables are Ridge and Lasso.

## Ridge regression

The least squares procedure estimates the $\beta_0$, $\beta_{1,}$..., $\beta_p$ using the values that minimize the value:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge Regression is very similar to Least Squares with the difference that beta rates are estimated by minimizing the value:

$$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

Where $\lambda \geq 0$ is a parameter that must be specified separately. As in Least Squares, Ridge Regression seeks coefficient estimates by reducing the RSS. However, the second term called shrinkage penalty is short when β1, ..., βp is close to zero, and thus has the effect of shrinking βj estimates to zero. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. Unlike Least Squares, which produces only one set of coefficient estimates, Ridge Regression will produce a different set of coefficients for each value of l. The advantage of Ridge regression over Least Squares is in exchanging variance with bias. As λ increases, the flexibility of Ridge regression adjustment decreases, leading to reduced variance but increased bias. In general, in situations where the relationship between response and predictions is close to linear, Least Squares estimates will be low bias but may fluctuate widely. This means that a small change in training data can cause a large change in Least Squares rate ratings, and therefore Ridge regression works best in situations where Least Squares ratings fluctuate widely.

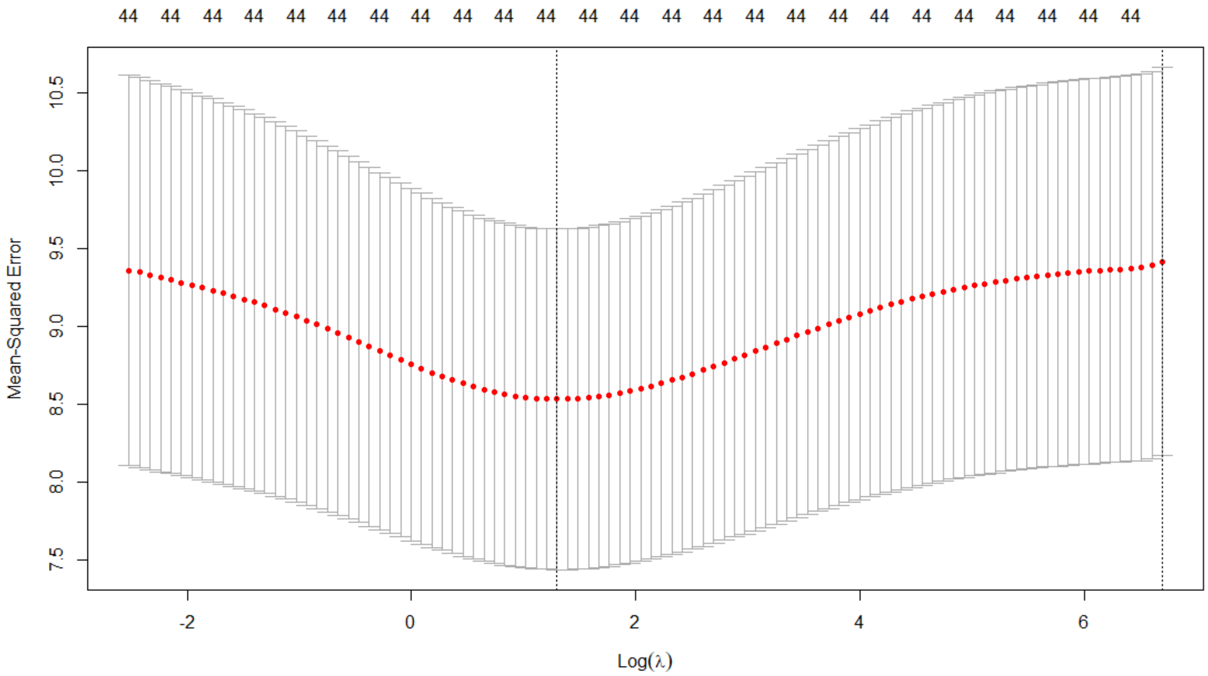


Figure 27.Log (λ) that minimise the MSE with Ridge Regression –Belgium

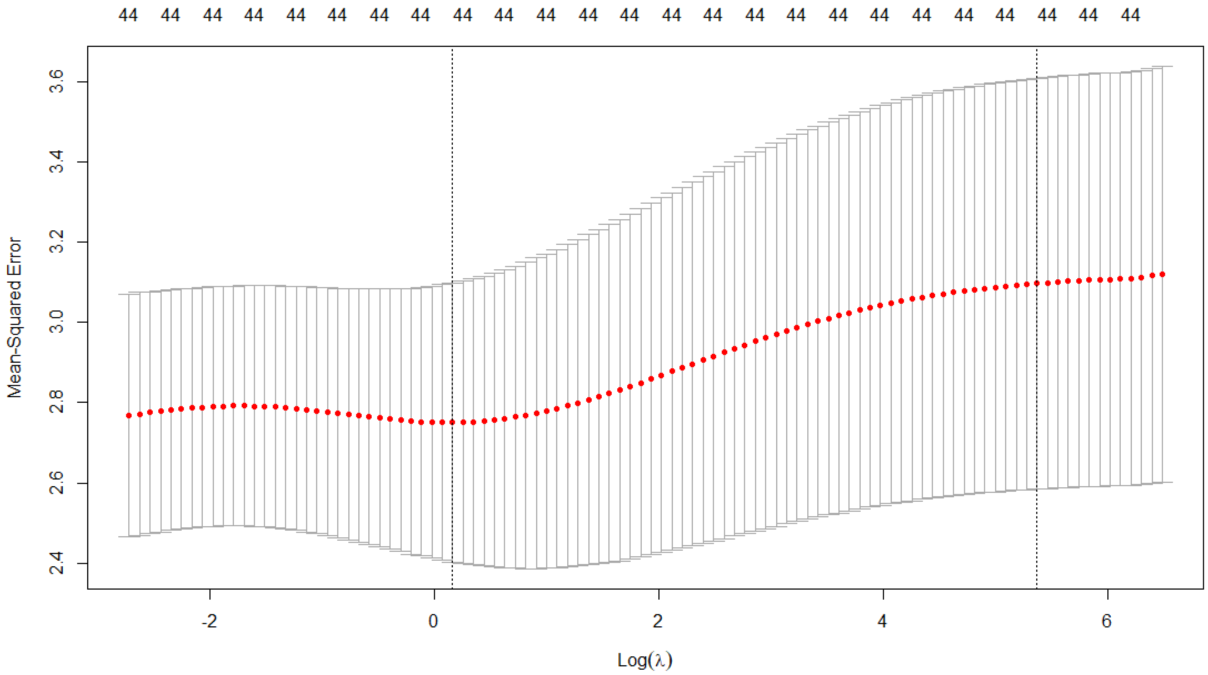


Figure 28.Log (λ) that minimise the MSE with Ridge Regression –Austria

In Figures 27 and 28 is identified the lambda value that produces the lowest mean squared error (MSE) by using leave one out cross-validation. The lowest value of MSE for Belgium is around 8.6, and for Austria it is around 2,8; using the Ridge Regression, we found that these values are given by λ = 3.744 and λ =1.729, respectively. The value of the logarithm that yields the minimum MSE is then given again in the next two graphs. It's also worth noting that as grows, the coefficients go closer to 0 but never reach it.

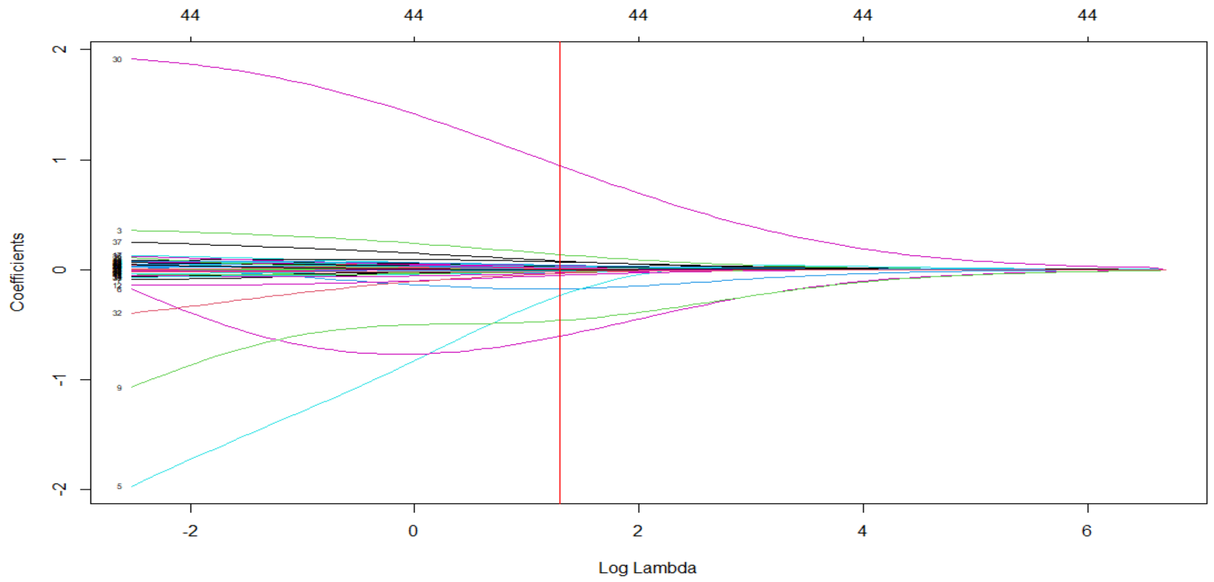


Figure 29. Fit of Log (λ) Ridge Regression –Belgium

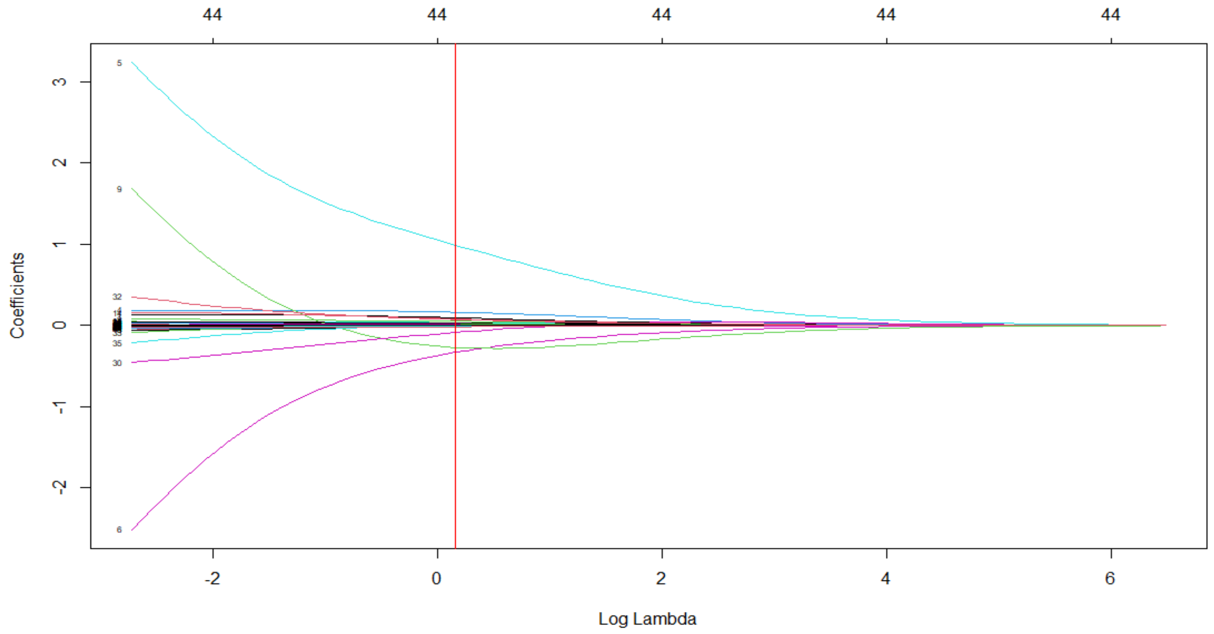


Figure 30. Fit of Log (λ) Ridge Regression –Austria

| *Variable* | *Ridge* | *Ridge \|beta\|* |
|---|---|---|
| **Euro-zone Business Climate Indicator** | 0.921 | 0.921 |
| **Euro yield curves** | -0.584 | 0.584 |
| **Spread** | -0.442 | 0.442 |
| **Money market interest rates** | -0.234 | 0.234 |
| **Production in construction** | 0.133 | 0.133 |
| **Unemployment Rate** | 0.106 | 0.106 |
| **Financial situation over next 3 months** | 0.101 | 0.101 |
| **Evolution of the current overall order books** | 0.071 | 0.071 |
| **Price expectations over next 3 months** | 0.051 | 0.051 |

Table 7.Ridge Coefficients-Belgium

| *Variable* | *Ridge* | *Ridge \|beta\|* |
|---|---|---|
| **Money market interest rates [** | 0,849 | 0,849 |
| **Spread Spread [10Y - 3M ]** | -0,295 | 0,295 |
| **Euro yield curves [10 YEAR YIELD]** | -0,216 | 0,216 |
| **Unemployment Rate** | 0,182 | 0,182 |
| **Euro-zone Business Climate Indicator** | -0,148 | 0,148 |
| **Euro area 19 international trade [EXPORTS]** | 0,090 | 0,090 |
| **Economic sentiment indicator** | 0,066 | 0,066 |
| **Turnover and volume of sales and retail trade** | -0,054 | 0,054 |
| **Production in construction** | 0,049 | 0,049 |
| **Financial situation [Consumer]** | -0,031 | 0,031 |
| **Euro area 19 international trade [IMPORTS]** | 0,020 | 0,020 |
| **Industrial confidence indicator** | 0,019 | 0,019 |
| **Financial situation [Consumer]** | 0,018 | 0,018 |
| **Production expectations months [Industry]** | 0,016 | 0,016 |
| **Expectation of Employment [Services]** | 0,015 | 0,015 |
| **Unemployment Expectations months [Consumer]** | -0,015 | 0,015 |

Table 8.Ridge Coefficients- Austria

Tables 7 and 8 demonstrate how the variables impact the variable Y, i.e. Industrial Production. The first column shows whether the variables affect it positively or negatively, while the second column shows how they effect it in absolute order in descending order. The analyst chooses the criterion selected beta, which in this case is | b | > 0.05 for Belgium and these 9 variables, and | b | > 0.015 for Austria and these 16 variables.

## Lasso regression

Ridge Regression has an obvious disadvantage. The shrinkage penalty will shrink all the coefficients to zero, but will not set any of them exactly to zero (unless λ = ∞). For this reason it will always keep all the variables without distinguishing which ones are important and this should is done by the analyst with the risk of bias. Because we want to avoid this risk and it is preferable to select the variables from the data themselves, the alternative Lasso regression was created. Lasso rates minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

We see that Ridge and Lasso Regression only have the difference that $\beta_{j2}$ has been replaced by y | $\beta_j$ | which also reduces the coefficient value to 0 with the difference that in Lasso Regression the penalty can actually reset some coefficients and therefore select variables. In conclusion Lasso Regression solves the Ridge Regression problem when we have many variables.

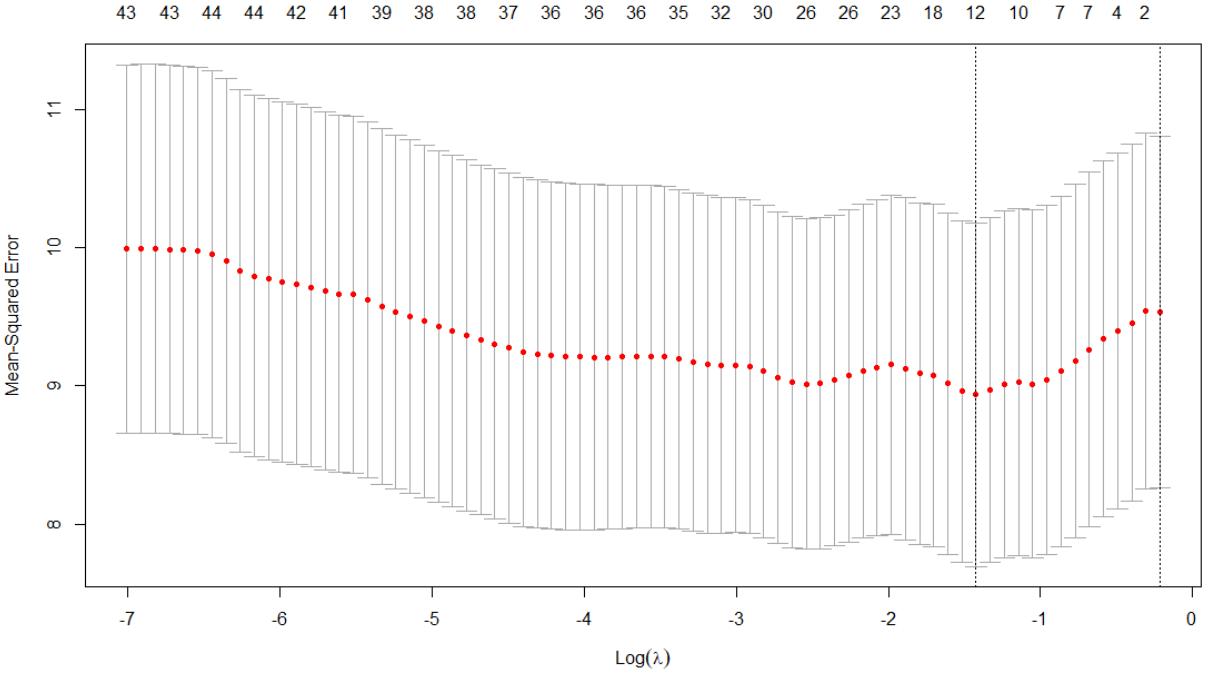


Figure 31.Log (λ) that minimise the MSE with Lasso Regression –Belgium

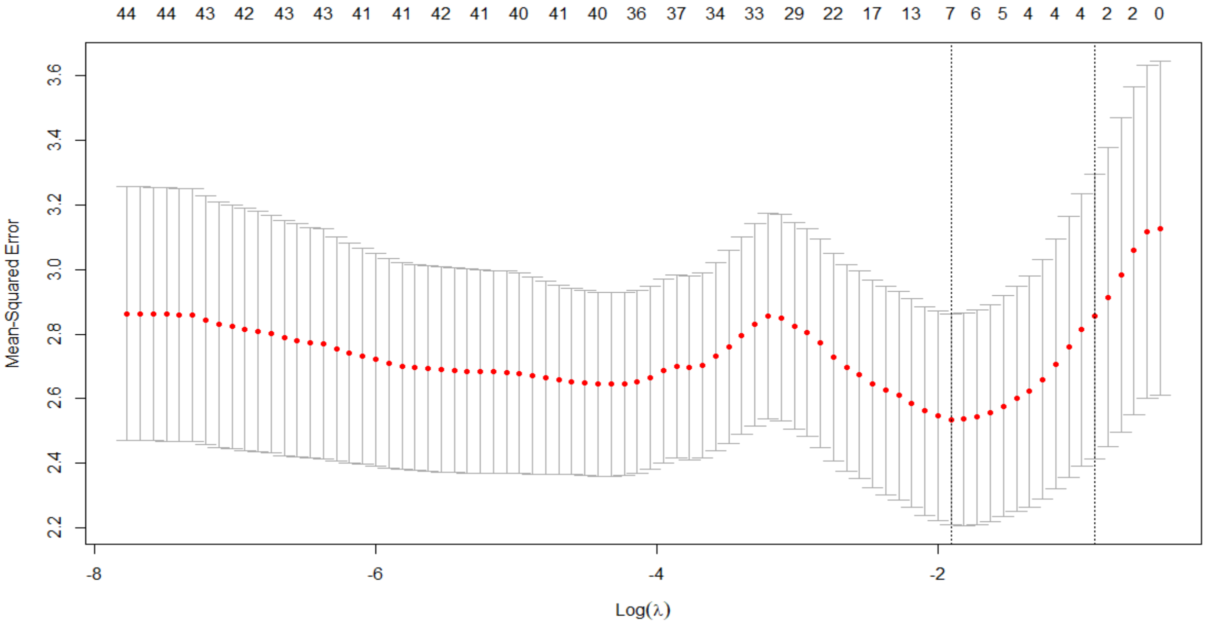


Figure 32.Log (λ) that minimise the MSE with Lasso Regression –Austria

In Figures 31 and 32 is identified the lambda value that produces the lowest mean squared error (MSE) by using leave one out cross-validation. The lowest value of MSE for Belgium is around 9, and for Austria it is around 2,6; using the Lasso Regression, we found that these values are given by λ= 0.169 and λ = 0.0944, respectively. The value of the logarithm that yields the minimum MSE is then given again in the next two graphs In all four graphs, the horizontal axis shows that, unlike the Ridge in Lasso, certain coefficients are zeroed, resulting in a reduction in total.

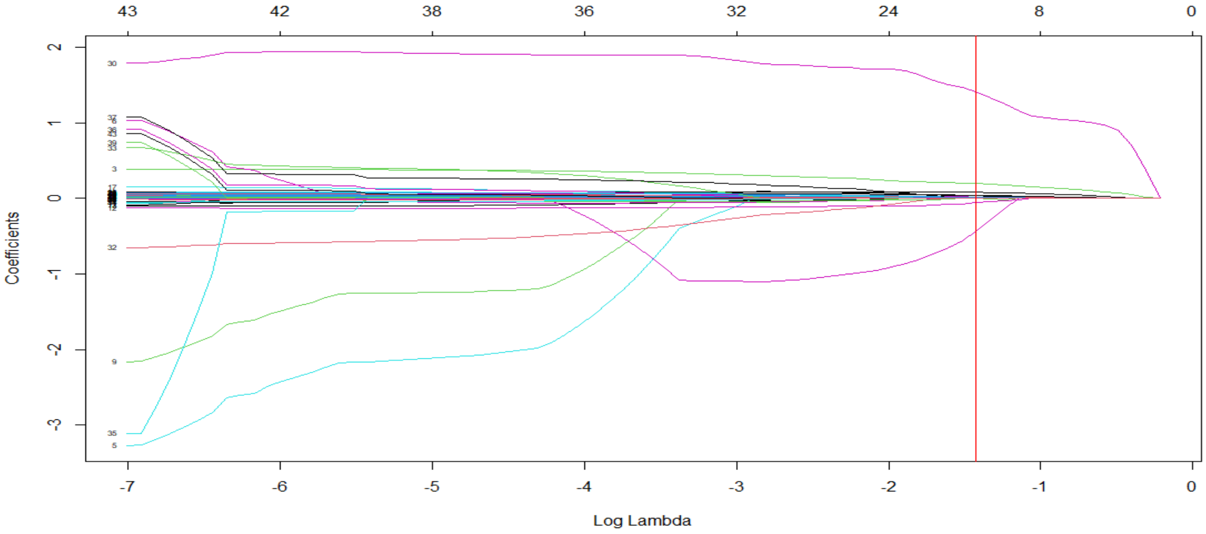


Figure 33. Fit of Log (λ) Lasso Regression –Belgium

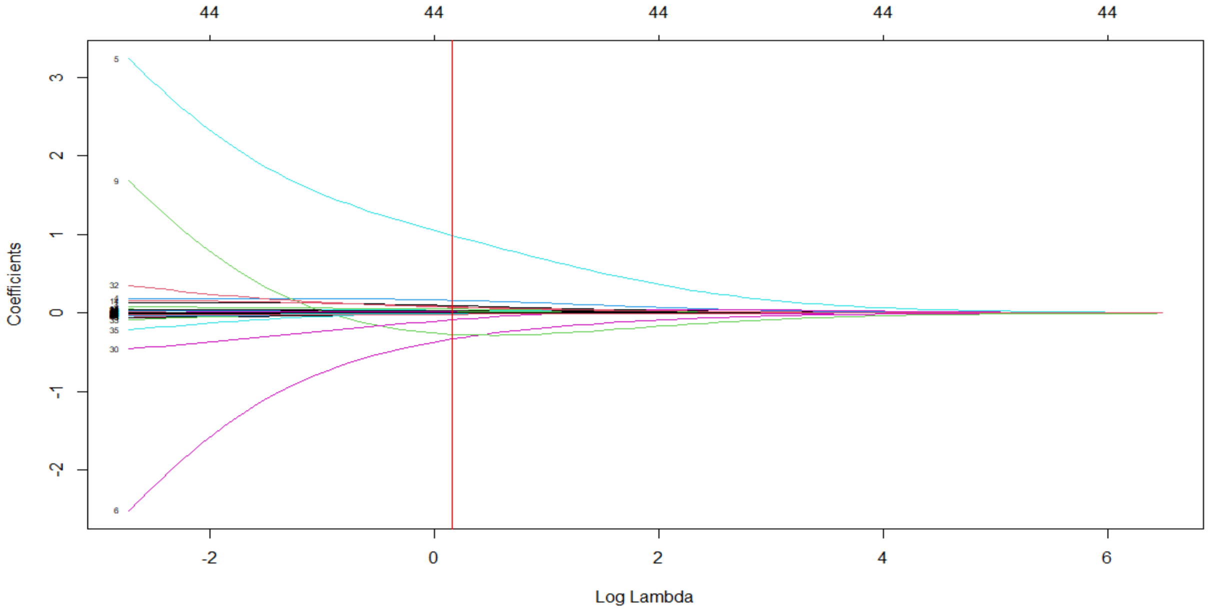


Figure 33. Fit of Log (λ) Lasso Regression –Austria

| *Variables* | *Lasso* | *Lasso \|beta\|* |
|---|---|---|
| **Euro-zone Business climate Indicator** | 1.516 | 1.516 |
| **Euro yield curves** | -0.425 | 0.425 |
| **Production in construction** | 0.207 | 0.207 |
| **Evolution of the current overall order books** | 0.109 | 0.109 |
| **Financial situation over next 12 months** | 0.100 | 0.100 |
| **Employment expectations over next 3 months** | -0.056 | 0.056 |
| **Euro area 19 international trade[imports]** | 0.047 | 0.047 |
| **Employment expectations over next 3 months[retail]** | 0.042 | 0.042 |
| **Assessment of order-book levels[industry]** | 0.039 | 0.039 |
| **Production expectations over the next 3 months** | -0.036 | 0.036 |
| **Evolution of employment over the past 3 months** | 0.034 | 0.034 |
| **Assessment of export order books levels** | 0.034 | 0.034 |
| **General economic situation over the next 12 months** | -0.032 | 0.032 |
| **Business Situation over the past 3 months** | 0.024 | 0.024 |
| **Price expectations over the next 3 months** | 0.016 | 0.016 |
| **Business activity development over the past 3 months** | 0.011 | 0.011 |
| **Expectations of Employment over the next 3 months** | 0.005 | 0.005 |
| **Savings over the next 12 months** | -0.001 | 0.001 |

Table 9.Lasso Coefficients -Belgium

| ***Variable*** | ***Lasso*** | ***Lasso \|beta\|*** |
|---|---|---|
| **Money market interest rates [3 MONTH YIELD]** | 0,627 | 0,627 |
| **Economic sentiment indicator** | 0,168 | 0,168 |
| **Euro area 19 international trade [EXPORTS]** | 0,156 | 0,156 |
| **Unemployment Rate** | 0,086 | 0,086 |
| **Turnover and volume of sales in wholesale and retail trade** | -0,065 | 0,065 |
| **Production in construction** | 0,060 | 0,060 |
| **Retail Confidence Indicator** | -0,019 | 0,019 |
| **Expectation of Employment over the next 3 months [Services]** | 0,019 | 0,019 |
| **Soft Production development observed over the past 3 months [Industry]** | -0,014 | 0,014 |
| **Employment expectations over the next 3 months [Industry]** | -0,013 | 0,013 |
| **Euro area 19 international trade [IMPORTS]** | 0,012 | 0,012 |
| **Price expectations over the next 3 months [Construction]** | -0,011 | 0,011 |
| **Financial situation over the last 12 Months [Consumer]** | -0,004 | 0,004 |
| **Business activity expectations over the next 3 months [Retail** | -0,001 | 0,001 |

Table 10.Lasso Coefficients –Austria

Through the Lasso method, it shows that the variables affect the variable Y, ie Production in Industry, in Tables 9 and 10. The first column shows whether they have a positive or negative impact, while the second shows how they have an absolute value impact in descending order. While there are some common variables (Euro-zone Business Climate Indicator, Euro yield curves, Production in construction, Evolution of the current overall order books, Price expectations over the next 3 months), there are many others that did not exist in Ridge or have been replaced by other variables in a different order. This distinction derives from the fact that, unlike Ridge, Lasso does not contain analyzer bias, and the variables are produced from the data themselves. We would have ignored numerous factors and incorrectly considered some variables relevant because of the criterion set if we had used the Ridge variables.

In conclusion, when we have a limited number of variables and can utilize them all, Ridge Regression is preferable than Lasso Regression, which is preferable when we have a huge number of variables and must choose which ones will be beneficial.

# Forecasting

Forecasting is the act of analyzing and mining data in order to predict what will happen in the future. Prediction is concernedwith future certainty; forecasting looks at how hidden currents in the present signal possible changes in direction for companies, societies, or the world at large. Thus, the primary goal of forecasting is to identify the full range of possibilities, not a limited set of illusory certainties. There are three basic types of forecasting: qualitative techniques, time series analysis and projection, and causal models. The dependent variable Y is the Production in Industry variable in the forecasting methods, and the recursive estimation method is the preferred estimation method.

We will make predictions on the immediately preceding part of the analysis, i.e. the Ridge and Lasso Regression, in this section of the analysis. The beta, intercept, and minimum λ estimations, as well as the forecasts for the two methods, will be shown below.

| *Coefficient* | *Belgium* | *Austria* |
|---|---|---|
| **b1** | 0.067 | 0.0179 |
| **b2** | 0.089 | -0.0281 |
| **b3** | 0.107 | -0.0028 |
| **b4** | 0.037 | 0.1162 |
| **b5** | 0.020 | 0.0533 |
| **b6** | 0.020 | 0.0596 |
| **b7** | -0.016 | -0.0318 |
| **b8** | 0.016 | 0.0700 |
| **b9** | -0.011 | 0.0002 |
| **b10** | 0.060 | -0.0813 |
| **b11** | 0.049 | 0.0442 |
| **b12** | 0.109 | -0.0429 |
| **b13** | 0.051 | 0.0368 |
| **b14** | 0.023 | 0.1616 |
| **b15** | 0.015 | 0.0266 |
| **b16** | 0.044 | -0.0268 |
| **b17** | 0.021 | -0.0039 |
| **b18** | 0.050 | 0.0897 |
| **b19** | 0.034 | -0.0146 |
| **b20** | -0.001 | 0.0903 |
| **b21** | -0.058 | -0.0520 |
| **b22** | -0.010 | 0.0605 |
| **b23** | 0.050 | 0.015906 |
| **b24** | 0.025 | -0.0135 |
| **b25** | -0.015 | -0.02473 |
| **b26** | 0.026 | -0.04102 |
| **b27** | 0.069 | -0.09045 |
| **b28** | -0.078 | -0.00671 |
| **b29** | -0.036 | 0.01383 |
| **b30** | 0.065 | 0.066965 |
| **b31** | 0.043 | 0.059048 |
| **b32** | 0.050 | 0.109638 |
| **b33** | 0.028 | 0.047741 |
| **b34** | 0.034 | 0.012308 |
| **b35** | 0.079 | 0.04381 |
| **b36** | 0.057 | 0.007762 |
| **b37** | 0.015 | 0.019986 |
| **b38** | 0.030 | -0.07085 |
| **b39** | 0.086 | 0.082266 |
| **b40** | 0.008 | 0.025696 |
| **b41** | 0.008 | 0.023917 |
| **b42** | -0.082 | 0.105359 |
| **b43** | 0.012 | -0.02676 |
| **b44** | 0.013 | 0.111097 |

Table 11.Ridge beta Estimation for both countries

The beta estimates for the two methods are shown in Tables 11 and 12. Ridge Regression and Lasso Regression betas represent the differences between Y (t) and X(t-1). According to the two tables, in the Ridge method, all betas are distinct from 0 even if they are very close, whereas in the Lasso method, many betas appear to be zero. This is supported by the fact that, as described in the preceding section of the analysis, the Lasso method can select variables and eliminate them, whereas in Ridge this does not happen because no threshold has been established. This also occurs when estimating coefficients.

| *Coefficient* | *Belgium* | *Austria* |
|---|---|---|
| **b1** | 0.191 | 0 |
| **b2** | 0.230 | 0 |
| **b3** | 0.355 | 0 |
| **b4** | 0 | 0.151 |
| **b5** | 0 | 0 |
| **b6** | 0 | 0 |
| **b7** | 0 | 0 |
| **b8** | 0 | 0 |
| **b9** | 0 | 0 |
| **b10** | 0.112 | -0.073 |
| **b11** | 0 | 0 |
| **b12** | 0.382 | 0 |
| **b13** | 0.000 | 0 |
| **b14** | 0 | 0.329 |
| **b15** | 0 | 0 |
| **b16** | 0 | 0 |
| **b17** | 0 | 0 |
| **b18** | 0.051 | 0.108 |
| **b19** | 0 | 0 |
| **b20** | 0 | 0.100 |
| **b21** | -0.038 | 0 |
| **b22** | 0 | 0 |
| **b23** | 0.090 | 0 |
| **b24** | 0 | 0 |
| **b25** | 0 | 0 |
| **b26** | 0 | 0 |
| **b27** | 0.181 | -0.042 |
| **b28** | -0.336 | 0 |
| **b29** | -0.229 | 0 |
| **b30** | 0.170 | 0 |
| **b31** | 0.046 | 0 |
| **b32** | 0 | 0.095 |
| **b33** | 0 | 0 |
| **b34** | 0 | 0 |
| **b35** | 0 | 0 |
| **b36** | 0.195 | 0 |
| **b37** | 0 | 0 |
| **b38** | 0 | 0 |
| **b39** | 0.322 | 0.016 |
| **b40** | 0 | 0 |
| **b41** | 0 | 0 |
| **b42** | -0.148 | 0.071 |
| **b43** | 0 | 0 |
| **b44** | 0 | 0.095 |

**Table 12.Lasso beta estimations for both countries**

Table 13 displays the intercept estimate first, then the minimum λ mentioned and earlier for the two methods, and lastly the forecast. The formula is used to make the forecast is forecast=Intercept+beta*X(Standardised).For Belgium the Ridge Regression forecast is 0.102 and for Lasso Regression forecast is -0.605. For Austria the Ridge Regression forecast is 0.217 and for Lasso Regression forecast is 0.086.

| | *Ridge* | | *Ridge* | *Lasso* | | *Lasso* |
|---|---|---|---|---|---|---|
| *Country* | *Intercept* | *Ridge λ* | *Forecast* | *Intercept* | *Lasso λ* | *Forecast* |
| **Belgium** | 0.142 | 11.765 | **0.102** | 0.138 | 0.145 | **-0.605** |
| **Austria** | 0.382 | 2.523 | **0.217** | 0.381 | 0.166 | **0.086** |

**Table 13.Estimations and forecasts for both countries**

The forecasts for PCA and PLS for the two countries are shown in Tables 14 and 15. Table 14 shows the predictions for PCA (1), PCA (2), PCA (3), PCA (4), and PCA (5), i.e. the first five factors. Table 15 likewise shows the PLS (1), PLS (2), PLS (3), PLS (4), and PLS (5), which are the PLS of the first five factors.

| *Country* | *PCA(1)* | *PCA(2)* | *PCA(3)* | *PCA(4)* | *PCA(5)* |
|---|---|---|---|---|---|
| **Belgium** | 0.671 | 0.437 | 0.488 | 0.596 | 0.617 |
| **Austria** | 0.151 | 0.043 | 0.092 | 0.057 | 0.057 |

Table 14.PCA(1)-PCA(5) both countries

| *Country* | *PLS(1)* | *PLS(2)* | *PLS(3)* | *PLS(4)* | *PLS(5)* |
|---|---|---|---|---|---|
| **Belgium** | 0.754 | 0.442 | 0.465 | -0.083 | 0.088 |
| **Austria** | 0.195 | 0.264 | 0.593 | 0.573 | 0.642 |

Table 15.PLS(1)-PLS(5) both countries

## Forecasting Evaluation

What criteria do we use to assess a forecasting model? What forecasting ability does a model have? In the final section of this analysis, we will attempt to answer these primary questions. This will be accomplished by developing a number of models and comparing them based on a set of criteria. What we're looking for in a forecasting model  now is whether the independent variables can predict the dependent, not how they explain it. The following are the models: AR(1),AR(2),Naïve, Linear Regression ESI, Linear Regression,Ridge, Lasso, PCA(1), PCA(2),PCA(3),PLS(1),PLS(2),PLS(3), Average of all models, Average PCAs, Average PLSs, Average MAE.

Let's have a look at these criteria.

## Mean Absolut Errors (MAE)

The simplest measure of forecast accuracy is called *Mean Absolute Error* (MAE). MAE is simply, as the name suggests, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. Once we have computed ($T^{OUT} - h$) forecasts for each model, we evaluate the forecasting performance using the mean absolute error

$$MAE_{j,h} = \frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} \left| e_{j,t} \right| ,$$

where $e_j$ is the out-of-sample forecast error (in levels) for model j. All our tables present the MAE relative to an AR(1), which serves as our benchmark.

## Root Mean Square Forecast Error (RMSFE)

Root *Mean Square Forecast Error* (RMSFE) are computed as a ratio between the RMSFE of the SWBF/SWBGG model and the RMSFE of the SW model. Hence values greater than one indicate that the SW model has a lower RMSFE than the alternative model featuring financial frictions. Once we have computed ($T^{OUT} - h$) forecasts for each model, we evaluate the forecasting performance using the mean absolute error and the root mean squared forecast error statistics, defined as:

$$RMSFE_{j,h} = \left( \frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} e_{j,t}^{2} \right)^{\frac{1}{2}} ,$$

where $e_j$ is the out-of-sample forecast error (in levels) for model j. All our tables present the MAE and RMSFE relative to an AR(1), which serves as our benchmark.

## Sign Success Ratio (SSR)

Finally, to evaluate the directional forecasts of each model we use the *Sign Success Ratio* (SSR) which is defined as the proportion of instances that the direction of the forecasts from each model is the same to the direction of the actual values. This is given by:

$$SSR_j = \frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} I \left[ sgn(\Delta y_{t+h}) = sgn(\widehat{y}_{t+h} - y_t) \right] ,$$

where sgn denotes the sign operator and I is an indicator variable which takes the value 1 if the signs are equal and 0 otherwise.

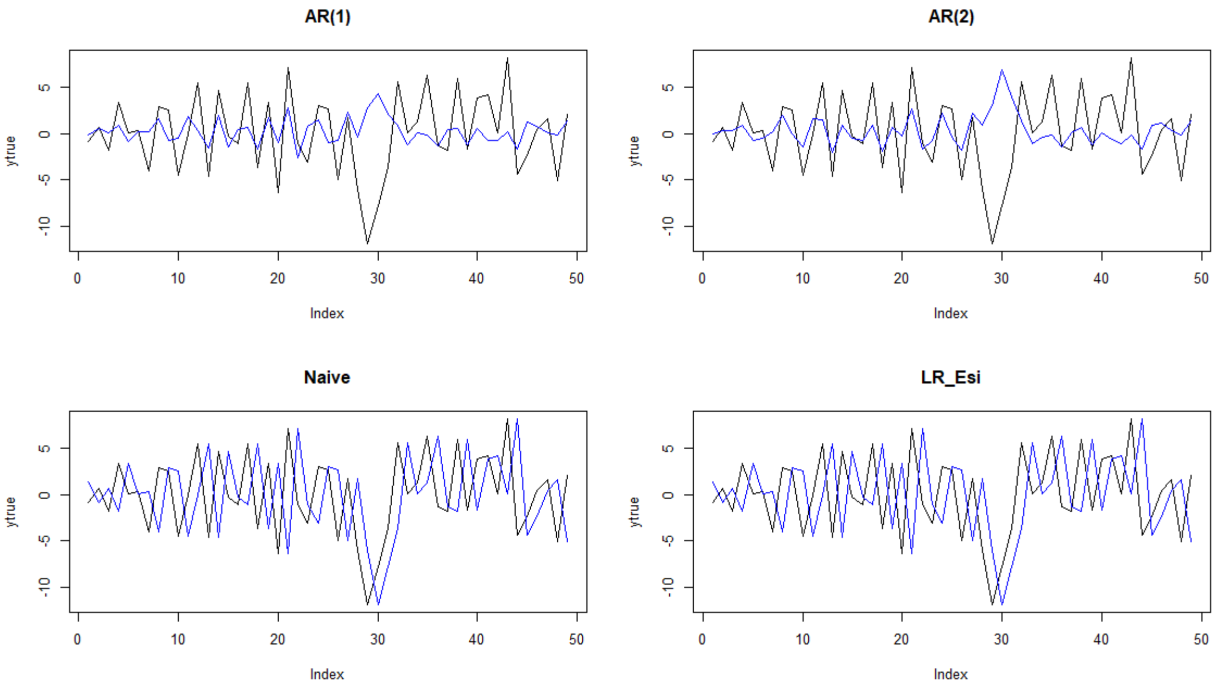


Figure 33.AR(1),AR(2),Naïve ,LR_ESI Forecast VS Reality –Belgium

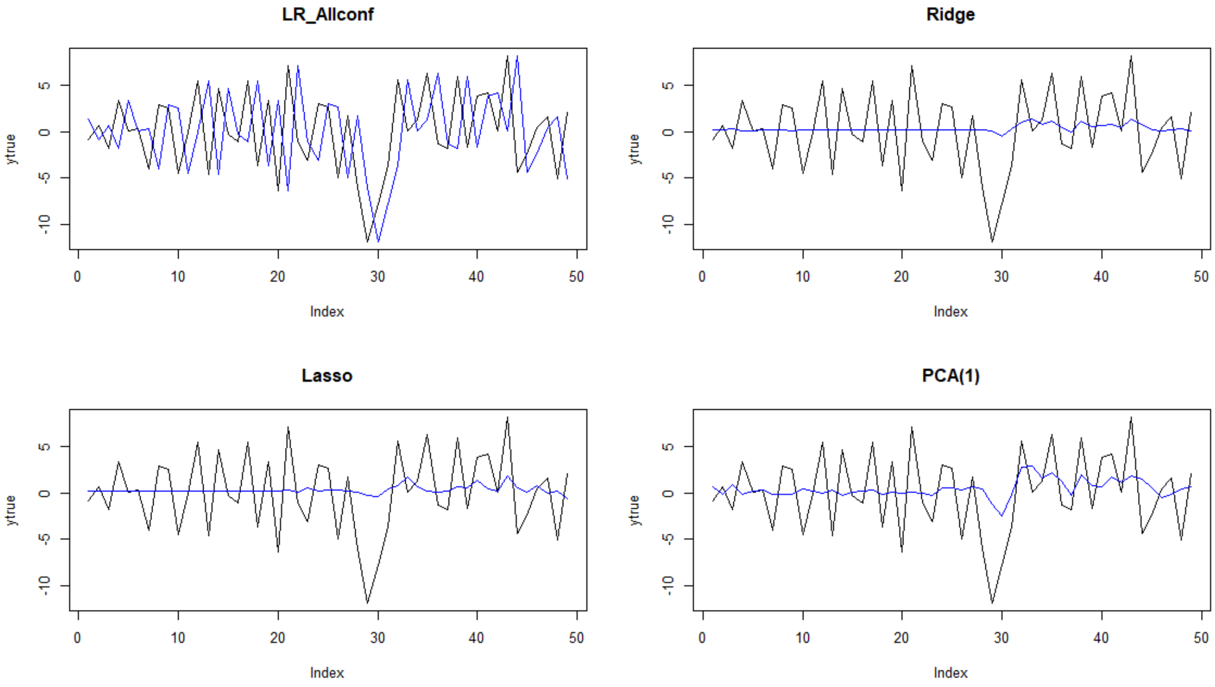


Figure 34. LR_Allconf, Ridge Lasso PCA(1) Forecast VS Reality –Belgium

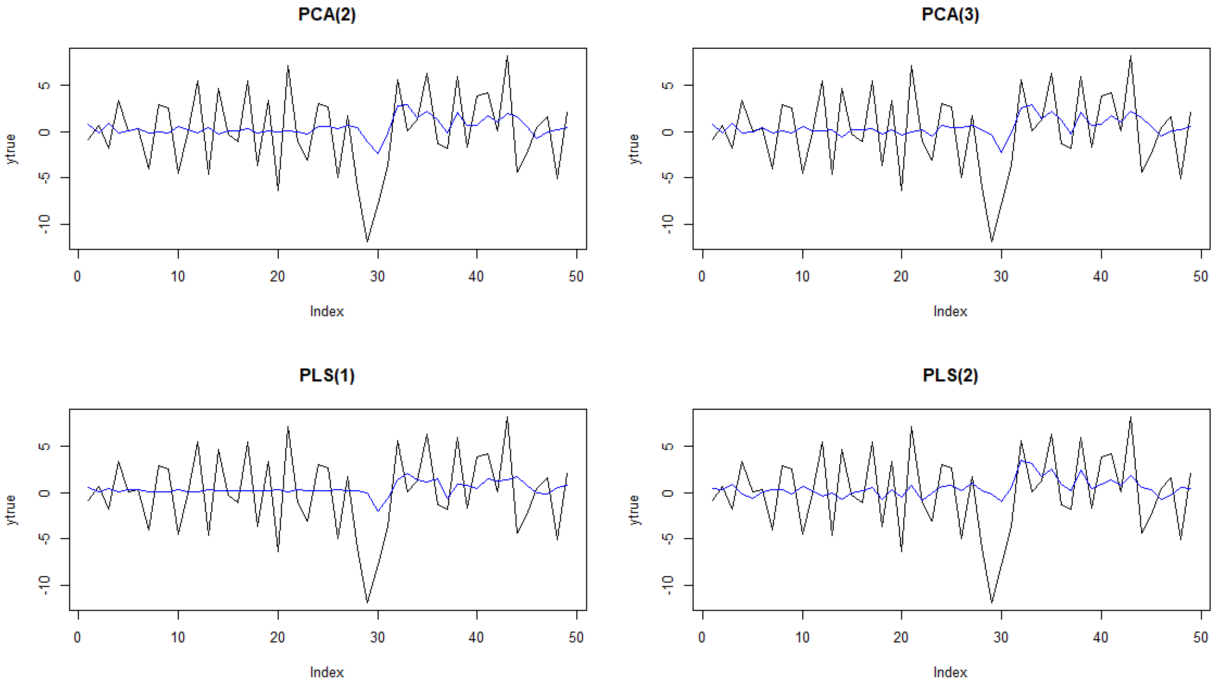


Figure 35. PCA(2) ,PCA(3),PLS(1),PLS(2) Forecast VS Reality –Belgium

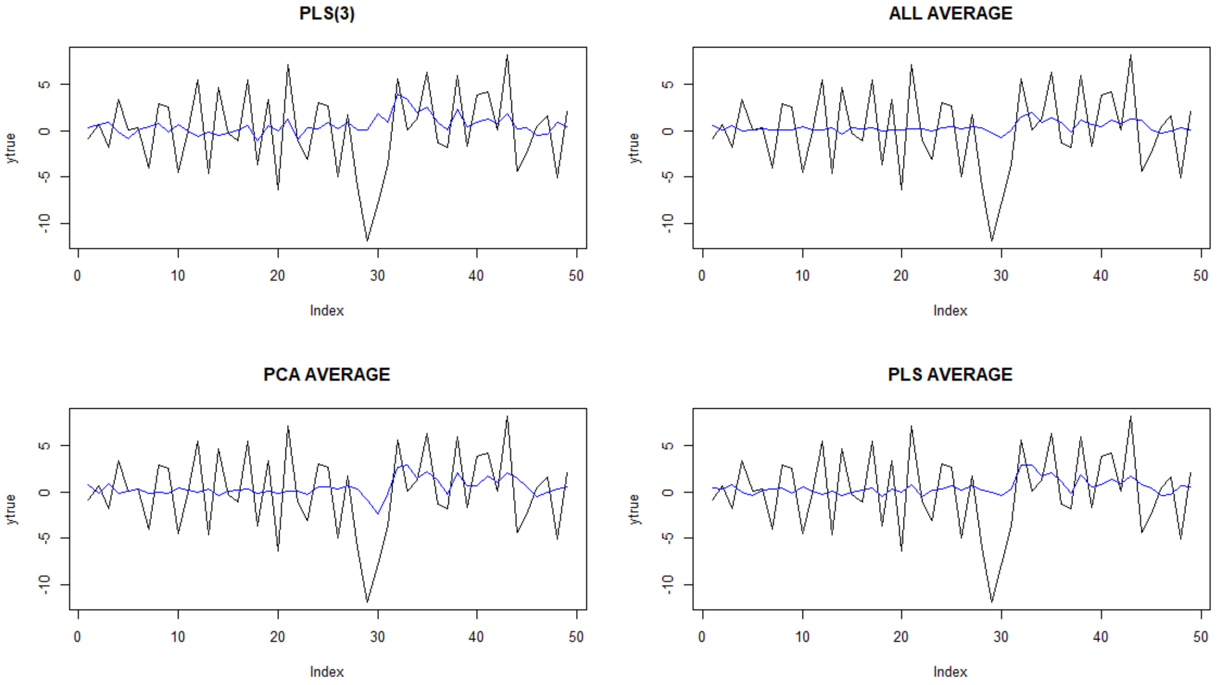


Figure 36. PLS(3),All Average ,PCA Average,PLS Average Forecast VS Reality –Belgium

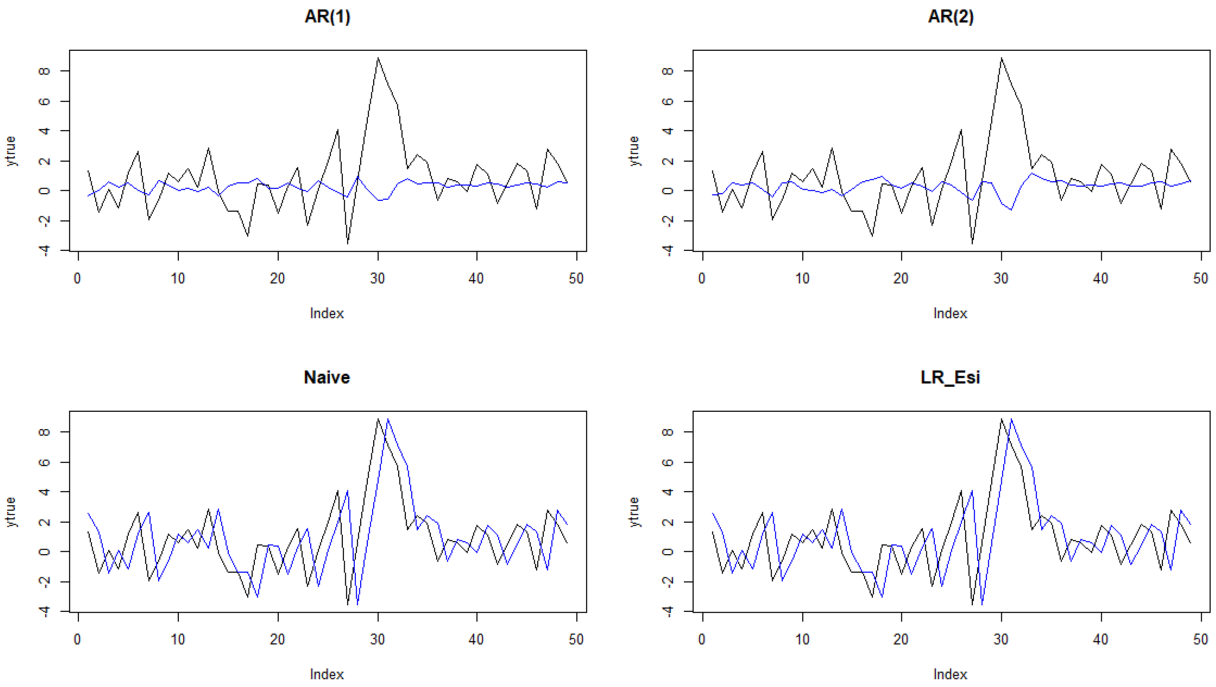


Figure 37. LR_Allconf, Ridge Lasso PCA(1) Forecast VS Reality –Austria

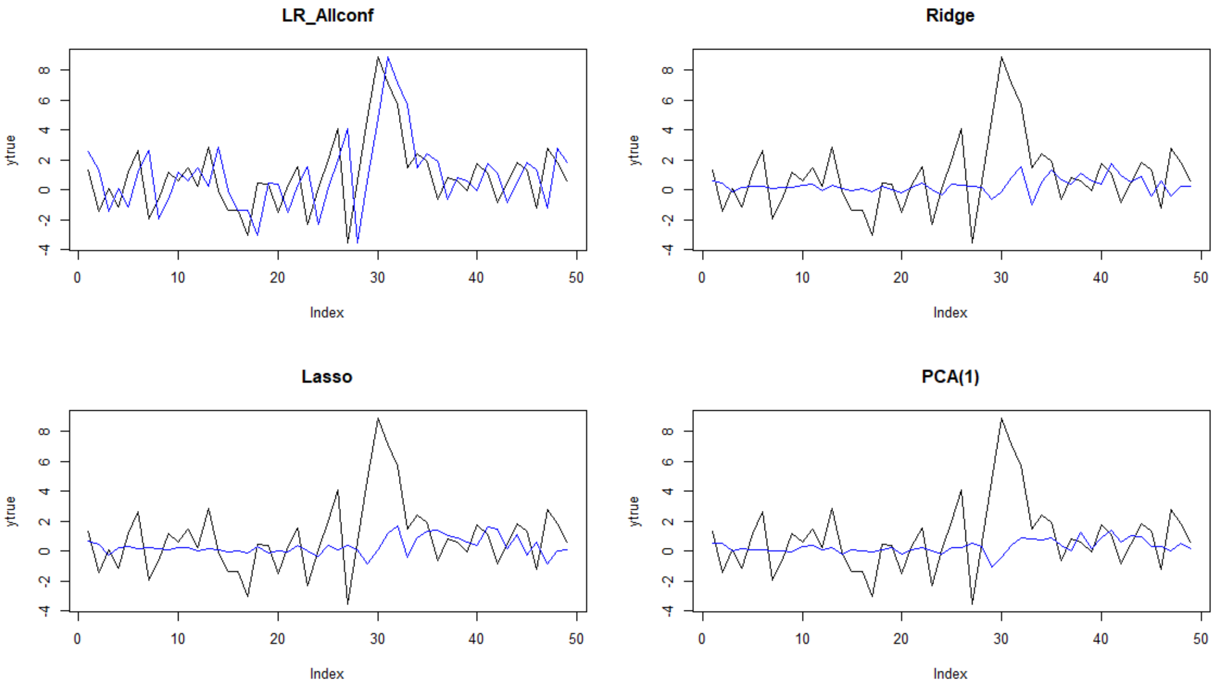


Figure 38. LR_Allconf, Ridge Lasso, PCA(1) Forecast VS Reality –Austria

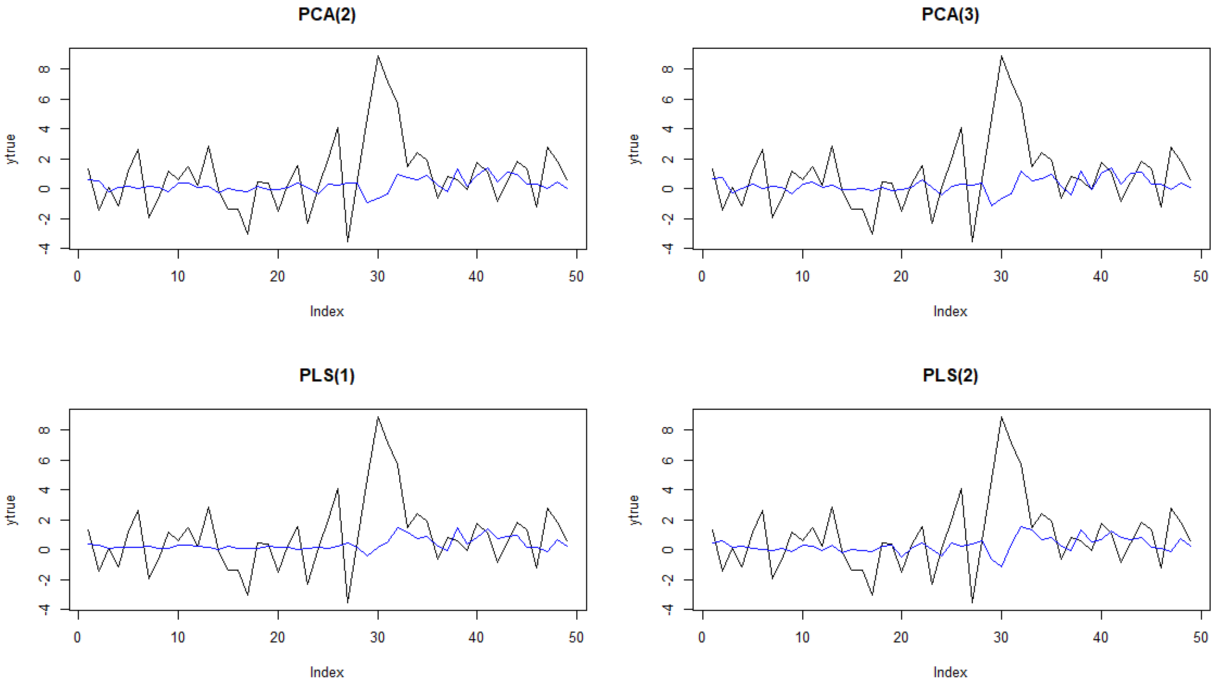


Figure 39 .PCA(2) ,PCA(3),PLS(1),PLS(2) Forecast VS Reality –Austria

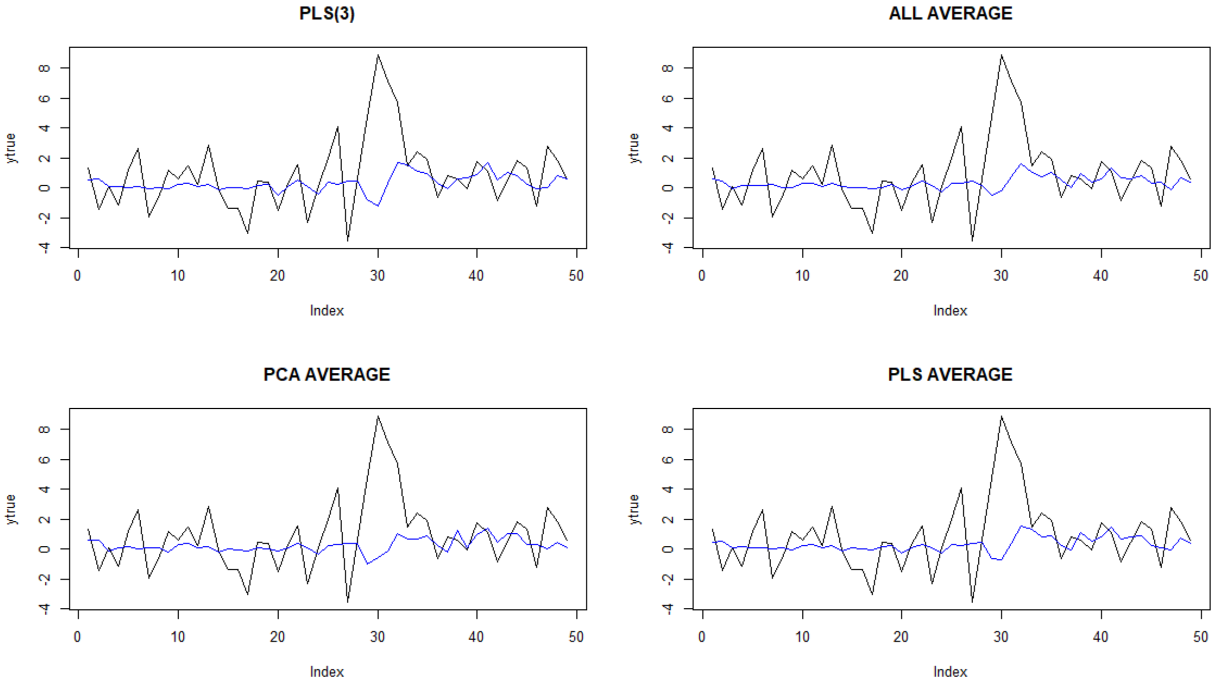


Figure 40 .PLS(3),All Average ,PCA Average, PLS Average Forecast VS Reality –Austria

The black line in the eight figures above depicts what actually happened, while the blue line depicts the prediction. While the two lines do not converge in most models in both countries, we notice that the models in Belgium and Austria appear to fall too far apart and are nearly identical, i.e. the naïve and two linear models.

| ***Model*** | ***SSR Belgium*** | ***SSR Austria*** |
|---|---|---|
| **AR(1)** | 0.653 | 0.673 |
| **AR(2)** | 0.633 | 0.694 |
| **Naive** | 0.388 | 0.551 |
| **LR_ESI** | 0.490 | 0.694 |
| **LR_AllConf** | 0.633 | 0.694 |
| **Ridge** | 0.571 | 0.592 |
| **Lasso** | 0.531 | 0.571 |
| **PCA(1)** | 0.551 | 0.694 |
| **PCA(2)** | 0.551 | 0.592 |
| **PCA(3)** | 0.551 | 0.612 |
| **PLS(1)** | 0.571 | 0.653 |
| **PLS(2)** | 0.531 | 0.633 |
| **PLS(3)** | 0.510 | 0.714 |
| **AVG-ALL** | 0.551 | 0.653 |
| **AVG-PCA** | 0.551 | 0.612 |
| **AVG-PLS** | 0.531 | 0.633 |
| **AVG-MAE** | 0.571 | 0.633 |

**Table 16.SSR both Countries**

Table 16 indicates how much the prediction was correct if the price dropped or increased. In Belgium, you can see the model that has largely succeeded; what the price has done is AR (1). In Austria, however, the models appear to have fared better in forecasting direction, with the AR (2) models, the two Linear Regression models, and the PCA (1) all performing well. The optimum model option is this larger SSR if the purpose is to see if the dependent variable will rise or not, rather than what value it will get.

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | **0.653** | 3.225 | 18.472 | 4.298 |
| AR(2) | 0.633 | 3.207 | 20.130 | 4.487 |
| Naive | 0.388 | 5.333 | 38.935 | 6.240 |
| LR_ESI | 0.490 | 3.381 | 17.742 | 4.212 |
| LR_AllConf | 0.633 | 3.329 | 16.306 | 4.038 |
| Ridge | 0.571 | 3.300 | 16.545 | 4.068 |
| Lasso | 0.531 | 3.319 | 16.580 | 4.072 |
| PCA(1) | 0.551 | 3.223 | **15.095** | **3.885** |
| PCA(2) | 0.551 | 3.225 | 15.109 | 3.887 |
| PCA(3) | 0.551 | 3.216 | 15.234 | 3.903 |
| PLS(1) | 0.571 | 3.325 | 16.420 | 4.052 |
| PLS(2) | 0.531 | **3.180** | 15.258 | 3.906 |
| PLS(3) | 0.510 | 3.228 | 16.338 | 4.042 |
| AVG-ALL | 0.551 | 3.321 | 16.428 | 4.053 |
| AVG-PCA | 0.551 | 3.220 | 15.138 | 3.891 |
| AVG-PLS | 0.531 | 3.241 | 15.828 | 3.979 |
| AVG-MAE | 0.571 | 3.261 | 16.134 | 4.017 |

Table 17. Forecast Statistics–Belgium

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | 0.673 | 1.737 | 6.378 | 2.525 |
| AR(2) | 0.694 | 1.714 | 6.583 | 2.566 |
| Naive | 0.551 | 1.999 | 6.035 | 2.457 |
| LR_ESI | 0.694 | 1.671 | 5.853 | 2.419 |
| LR_AllConf | 0.694 | 1.614 | 6.349 | 2.520 |
| Ridge | 0.592 | 1.694 | 5.784 | 2.405 |
| Lasso | 0.571 | 1.721 | 5.816 | 2.412 |
| PCA(1) | 0.694 | 1.655 | 6.019 | 2.453 |
| PCA(2) | 0.592 | 1.684 | 6.280 | 2.506 |
| PCA(3) | 0.612 | 1.678 | 6.291 | 2.508 |
| PLS(1) | 0.653 | 1.633 | 5.598 | 2.366 |
| PLS(2) | 0.633 | 1.633 | 6.073 | 2.464 |
| PLS(3) | **0.714** | **1.594** | 6.040 | 2.458 |
| AVG-ALL | 0.653 | 1.613 | **5.577** | **2.362** |
| AVG-PCA | 0.612 | 1.672 | 6.188 | 2.487 |
| AVG-PLS | 0.633 | 1.616 | 5.879 | 2.425 |
| AVG-MAE | 0.633 | 1.612 | 5.630 | 2.373 |

Table 18.Forecast Statistics-Austria

In tables 17 and 18 is presented the forecast statistics for the out-of sample period 12/31/2017 to 12/31/2021(49 out of sample observations) for both countries. For Belgium, it appears that the AR (1) model has the greatest score, the PLS (2) model has the lowest MAE, and the PCA model has the lowest MSFE and RMSFE (1). What's interesting is that only the last two statistical indicators agree on which model is the best. In Austria, however, the SSR and MAE appear to concur on the most appropriate model, as the PLS (3) produces both the highest SSR and the lowest MAE. As in Belgium, the latter two criteria agree with each other, resulting in the All Average model being the best model. Below are some charts showing some of the better models taken from this table.

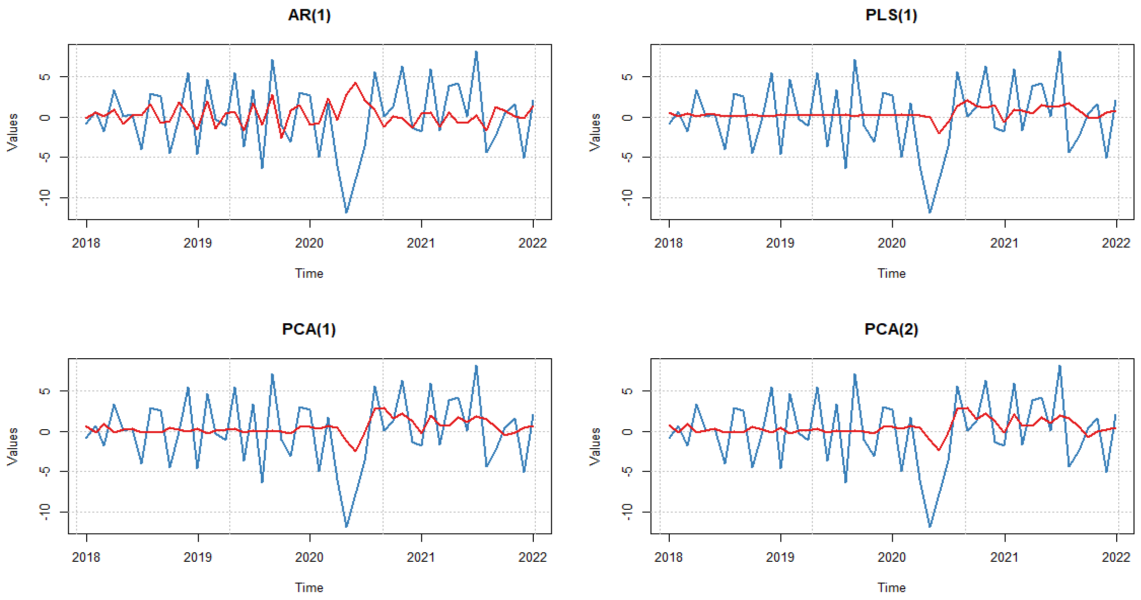


Figure 41.Best models for Belgium

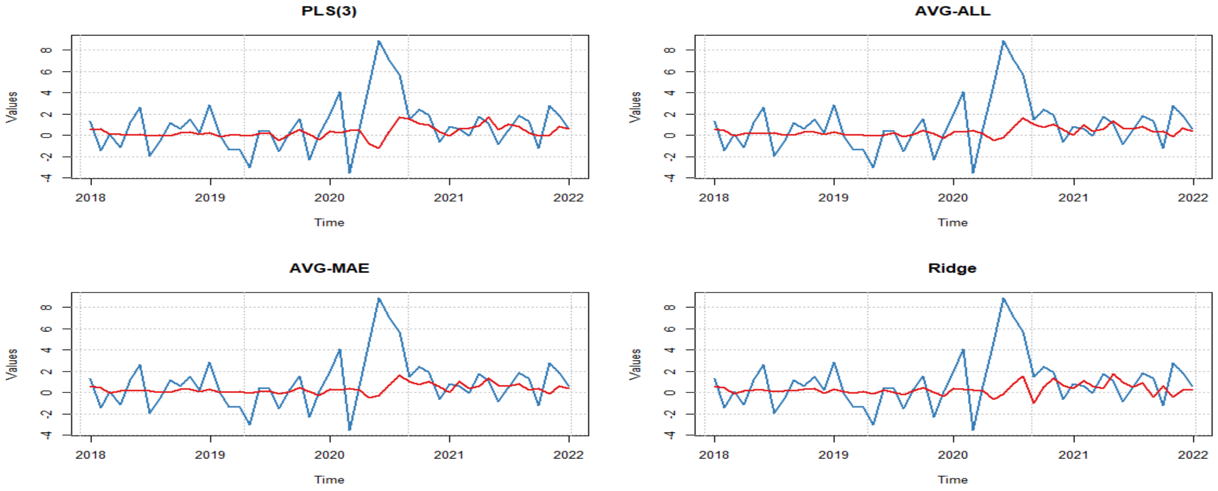


Figure 41.Best models for Austria

We'll look at how forecast error changes as you go deeper into the horizons. When the temporal horizon shifts, the statistical indicators shift as well, which means the best model for each criterion is likely to shift as well.

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | 0.519 | 2.971 | 12.760 | 3.572 |
| AR(2) | **0.667** | 2.903 | 12.299 | 3.507 |
| Naive | 0.519 | 3.538 | 20.144 | 4.488 |
| LR_ESI | 0.519 | 2.937 | 12.475 | 3.532 |
| LR_AllConf | 0.556 | **2.885** | **11.559** | **3.400** |
| Ridge | 0.407 | 2.979 | 12.705 | 3.564 |
| Lasso | 0.519 | 2.961 | 12.692 | 3.563 |
| PCA(1) | 0.481 | 2.962 | 12.656 | 3.557 |
| PCA(2) | 0.481 | 2.958 | 12.466 | 3.531 |
| PCA(3) | 0.444 | 3.009 | 12.722 | 3.567 |
| PLS(1) | 0.481 | 2.992 | 12.624 | 3.553 |
| PLS(2) | 0.444 | 2.966 | 12.543 | 3.542 |
| PLS(3) | 0.481 | 2.989 | 12.696 | 3.563 |
| AVG-ALL | 0.481 | 2.924 | 12.255 | 3.501 |
| AVG-PCA | 0.444 | 2.976 | 12.609 | 3.551 |
| AVG-PLS | 0.444 | 2.981 | 12.607 | 3.551 |
| AVG-MAE | 0.481 | 2.928 | 12.281 | 3.504 |

Table 19.Forecast Statistics for h=3 –Belgium

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | 0.673 | 1.737 | 6.378 | 2.525 |
| AR(2) | 0.694 | 1.714 | 6.583 | 2.566 |
| Naive | 0.551 | 1.999 | 6.035 | 2.457 |
| LR_ESI | 0.694 | 1.671 | 5.853 | 2.419 |
| LR_AllConf | 0.694 | 1.614 | 6.349 | 2.520 |
| Ridge | 0.592 | 1.694 | 5.784 | 2.405 |
| Lasso | 0.571 | 1.721 | 5.816 | 2.412 |
| PCA(1) | 0.694 | 1.655 | 6.019 | 2.453 |
| PCA(2) | 0.592 | 1.684 | 6.280 | 2.506 |
| PCA(3) | 0.612 | 1.678 | 6.291 | 2.508 |
| PLS(1) | 0.653 | 1.633 | 5.598 | 2.366 |
| PLS(2) | 0.633 | 1.633 | 6.073 | 2.464 |
| PLS(3) | **0.714** | **1.594** | 6.040 | 2.458 |
| AVG-ALL | 0.653 | 1.613 | **5.577** | **2.362** |
| AVG-PCA | 0.612 | 1.672 | 6.188 | 2.487 |
| AVG-PLS | 0.633 | 1.616 | 5.879 | 2.425 |
| AVG-MAE | 0.633 | 1.612 | 5.630 | 2.373 |

Table 20.Forecast Statistics for h=3 =Austria

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| **AR(1)** | 0.531 | 3.361 | **17.681** | **4.205** |
| **AR(2)** | 0.510 | 3.369 | 17.765 | 4.215 |
| **Naive** | 0.510 | 4.700 | 36.494 | 6.041 |
| **LR_ESI** | **0.571** | 3.364 | 17.767 | 4.215 |
| **LR_AllConf** | 0.408 | 3.479 | 18.855 | 4.342 |
| **Ridge** | 0.531 | 3.361 | **17.681** | 4.205 |
| **Lasso** | 0.531 | 3.361 | **17.681** | 4.205 |
| **PCA(1)** | 0.490 | 3.377 | 17.863 | 4.226 |
| **PCA(2)** | 0.449 | 3.381 | 18.016 | 4.245 |
| **PCA(3)** | 0.469 | 3.395 | 18.177 | 4.263 |
| **PLS(1)** | 0.490 | 3.381 | 17.888 | 4.229 |
| **PLS(2)** | 0.490 | 3.375 | 18.154 | 4.261 |
| **PLS(3)** | 0.531 | **3.346** | 18.671 | 4.321 |
| **AVG-ALL** | 0.551 | 3.359 | 18.236 | 4.270 |
| **AVG-PCA** | 0.490 | 3.385 | 18.009 | 4.244 |
| **AVG-PLS** | 0.531 | 3.367 | 18.193 | 4.265 |
| **AVG-MAE** | **0.571** | 3.363 | 18.118 | 4.257 |

Table 21.Forecast Statistics for h=6 -Belgium

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| **AR(1)** | 0.694 | 1.691 | 5.890 | 2.427 |
| **AR(2)** | 0.694 | 1.691 | 5.889 | 2.427 |
| **Naive** | 0.612 | 2.450 | 10.944 | 3.308 |
| **LR_ESI** | 0.694 | **1.682** | 5.846 | **2.418** |
| **LR_AllConf** | 0.653 | 1.731 | 6.073 | 2.464 |
| **Ridge** | 0.694 | 1.692 | **5.888** | 2.427 |
| **Lasso** | **0.714** | 1.700 | 5.943 | 2.438 |
| **PCA(1)** | 0.592 | 1.720 | 6.014 | 2.452 |
| **PCA(2)** | 0.612 | 1.720 | 6.013 | 2.452 |
| **PCA(3)** | 0.633 | 1.717 | 6.093 | 2.468 |
| **PLS(1)** | 0.673 | 1.695 | 5.923 | 2.434 |
| **PLS(2)** | 0.633 | 1.714 | 6.028 | 2.455 |
| **PLS(3)** | 0.612 | 1.713 | 6.064 | 2.463 |
| **AVG-ALL** | 0.633 | 1.707 | 5.941 | 2.437 |
| **AVG-PCA** | 0.612 | 1.718 | 6.038 | 2.457 |
| **AVG-PLS** | 0.633 | 1.704 | 5.995 | 2.448 |
| **AVG-MAE** | 0.653 | 1.698 | 5.927 | 2.435 |

Table 22.Forecast Statistics for h=6 -Austria

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | 0.531 | 3.362 | 17.633 | 4.199 |
| AR(2) | 0.531 | 3.363 | 17.632 | 4.199 |
| Naive | 0.388 | 5.598 | 45.004 | 6.708 |
| LR_ESI | **0.633** | **3.329** | **17.217** | **4.149** |
| LR_AllConf | 0.531 | 3.457 | 18.467 | 4.297 |
| Ridge | 0.510 | 3.492 | 18.754 | 4.331 |
| Lasso | 0.449 | 3.517 | 18.986 | 4.357 |
| PCA(1) | 0.449 | 3.395 | 17.532 | 4.187 |
| PCA(2) | 0.429 | 3.428 | 18.020 | 4.245 |
| PCA(3) | 0.551 | 3.475 | 18.519 | 4.303 |
| PLS(1) | 0.490 | 3.418 | 18.005 | 4.243 |
| PLS(2) | 0.490 | 3.378 | 17.407 | 4.172 |
| PLS(3) | 0.408 | 3.419 | 17.712 | 4.209 |
| AVG-ALL | 0.449 | 3.500 | 18.812 | 4.337 |
| AVG-PCA | 0.510 | 3.429 | 17.969 | 4.239 |
| AVG-PLS | 0.469 | 3.401 | 17.679 | 4.205 |
| AVG-MAE | 0.449 | 3.467 | 18.476 | 4.298 |

Table 23.Forecast Statistics for h=12 -Belgium

| *Model* | *SSR* | *MAE* | *MSFE* | *RMSFE* |
|---|---|---|---|---|
| AR(1) | 0.694 | 1.696 | 5.890 | 2.427 |
| AR(2) | 0.694 | 1.696 | 5.889 | 2.427 |
| Naive | 0.612 | 2.559 | 11.864 | 3.444 |
| LR_ESI | 0.694 | 1.702 | 5.955 | 2.440 |
| LR_AllConf | **0.714** | **1.683** | **5.887** | **2.426** |
| Ridge | 0.694 | 1.696 | 5.892 | 2.427 |
| Lasso | 0.673 | 1.707 | 5.933 | 2.436 |
| PCA(1) | 0.694 | 1.712 | 5.917 | 2.432 |
| PCA(2) | 0.694 | 1.722 | 5.944 | 2.438 |
| PCA(3) | 0.653 | 1.737 | 6.080 | 2.466 |
| PLS(1) | 0.694 | 1.714 | 5.931 | 2.435 |
| PLS(2) | 0.694 | 1.707 | 5.908 | 2.431 |
| PLS(3) | 0.694 | 1.702 | 5.943 | 2.438 |
| AVG-ALL | 0.673 | 1.693 | 5.974 | 2.444 |
| AVG-PCA | 0.673 | 1.723 | 5.975 | 2.444 |
| AVG-PLS | 0.694 | 1.708 | 5.926 | 2.434 |
| AVG-MAE | 0.694 | 1.691 | 5.955 | 2.440 |

Table 24.Forecast Statistics for h=12 -Austria

What we see in the above tables is that when we move to longer horizons, the SSR for all models falls, but the MAE, MSFE, and RMSFE increase. In other words, as h increases, predicting models become less successful for both countries. Also worth noting is that when h = 12, the 4 criteria for the most suited model agree for both Belgium and Austria, and that both countries' models are Linear Regression models.

# BIBLIOGRAPHY

- James, G., Witten, D., Hastie, T., Tibshirani, T. (2013). An Introduction to Statistical Learning with Applications in R. Springer, New York
- Helland (1990)
- Giannone, D., Reichlin, L., Small, D. (2008). “Nowcasting: The real-time informational content of macroeconomic data”. Journal of Monetary Economics