

**UNIVERSITY OF TARTU**

**Millenium Data – A Quantitative Exploratory Analysis Using  
Regularization Models on the UK GDP Growth Through Multi  
Frequency Data**

Master Thesis

Alumni: Pedro Marco Melo Silva

Supervisor: Luca Alfieri

2023

This Master Thesis has been written independently. The ideas and data taken from other authors or other sources have been fully referenced.

## **Abstract**

This study analyzes the dataset “A millennium of macroeconomic data for the UK” provided by the Bank of England and, through regularization models (Lasso, Ridge and Elastic Net), explores the multitude of time series to be checked for relevance to the yearly GDP growth throughout the selected period of the UK history. This combination is used to identify which variables and models are the most relevant to explain the GDP performance throughout the analyzed period and to be applied to forecast it. A comparison to an autoregressive benchmark model takes place to understand which model has the lowest root mean square errors (RSMES) in order to test for accuracy and Diebold-Mariano for significant differences between the regularization models and the AR. As a result, for this dataset analyzed, the regularization models have performed the best, especially the Elastic Net.

**Keywords:** *Time Series, Elastic Net, Lasso, Ridge, UK GDP Forecasting*

**CERCS code:** S180

## 1. INTRODUCTION

Understanding the relationship between cause and effect between factors is a basic human trait. This curiosity is what drove humanity to ask good questions and strive to seek the right answers. Good questions are the engine for which science unravels the world surrounding us and it is no different in the realm of modern economics, which started with Adam Smith's observations on social and macroeconomic dynamics laid the foundation for liberalism and other theories economic theories and explanations on cause and consequences of humans and countries' dynamics through his *magna opus* "The Wealth of Nations" (1776).

This thesis stems from acknowledging the power of asking good questions to foster intellectual growth and as a result, it relies on the innovative approach brought by Kuznets (1934) which introduced metrics to measure a country's national income that, later on, became known as the Gross Domestic Product (GDP). By taking the GDP as a measurement of a country's wealth, the research conducted in this thesis tackles the author's curiosity on understanding which factors are the most relevant to explaining the development of this metric on a yearly basis. This analysis will be expanded in a way that it will encompass a comparison between Regularization Models (Ridge, Lasso, and Elastic-Net) to an autoregressive model AR (5) applied to a forecast of up to three horizon periods.

The novelty of this thesis relies on the exploration of the dataset made available by the Bank of England (2017), called A Millennium of Macroeconomic Data for the UK. Such dataset provides with information from over a thousand years of data, reaching different categories inside the economic scope of their work done and it's time frequency diversity, which this paper accommodates properly, especially the multifrequency present on this dataset, which at first glance, seem incompatible and would require a bridging method to be applied in order to uniform all the frequencies from the time series to be used in the analysis. Converting the higher frequency time series to the lower one, resulting in the investigation of weekly, monthly, and yearly data. On total, 56 (fifty-six) economic variables were selected from the vast pool provided, including the responsive variable for analysis, the yearly UK GDP. After analysis of the dataset, it was concluded that, to use the most amount of period compatible variables, the optimal year range for analysis in this given dataset has been between the years of 1870 to 2006.

The theoretical literature used for this paper relies on the work of economists and thinkers and their perspective on Business Cycles and GDP growth. According to Bylund (2022), Mises puts the entrepreneur as the main drive for the market process, but in this thesis, it is explored if there is enough magnitude for the variables related to productivity and innovation generation, key components of entrepreneurship, and analyze how relevant these variables are to explain the yearly GDP of the UK and to be used to forecast it. In Huerta Soto's work (2009 & 2010), it is acknowledged entrepreneurship related variables as drivers for economic growth but it shows concern for politicians that act on expanding the upcycle of an economy through loose monetary policies, creating distortions and misallocation of capital in the economy, specially to sectors that are more capital intensive.

By knowing how both perceive the main drivers for the GDP growth, in this paper it is investigated how much, it is possible to connect their two positions and analyze if, government expenses, which are villainized by Huerta-Soto could generate a leverage large enough to make the relevance of productivity and innovation bigger than government related expenses. These and other observations will be tested for relevance by machine learning regularization models.

The empirical literature in this paper relies on papers of forecasts done by researchers that have done regularization tests on other datasets. Such as Zhang et al. (2019) that investigates the performance of a forecasting model for oil prices forecasting and Anesti et al. (2021) that has a similar approach applied to monthly British GDP on a monthly basis.

After having identified which are the most relevant variables, to be considered for the yearly UK GDP, a forecast will take place with all regression models and compared to the benchmark, an Auto Regressive Model of order 5 - AR (5), using the root mean Square Error (RMSE) to assess which of the methods is the most accurate one.

By investigating the variables, this paper has concluded that Regularization Models, such as Lasso, Ridge, and Elastic Net yield better results than an Autoregressive Model AR (5). Although not in all the models and all the horizons. It has also been possible to detect which variables are the most relevant to the models, and which of these models yield the best results, being the Elastic Net the one that balances the best outcomes from the Ridge and Lasso and provides with the most accurate forecasting available in this paper. Such results

will be connected to the literature and further explanation on the goals achieved and considerations will be outlined on the fourth and fifth sections.

This paper has been structured in a way that can assertively guide the reader through this research while understanding the references and technique used. The Section 2 of this thesis presents the Literature Review, divided into theoretical background, which explains the theoretical approach the paper used as a backbone for the thesis; and empirical literature, focusing on the technical literature content from the models employed and mentioned throughout this paper. In Section 3 there are three subsections, divided by a research design that register the steps taken to generate the results needed; a detailed approach on the data employed; and the technical methodology for the machine learning models used in this paper. Section 4 provides the quantitative and qualitative results of the data. Section 5 is the conclusion of this paper, which summarizes the findings of the paper and sets up the pace for further analysis in the future.

## **2. LITERATURE REVIEW**

### **Theoretical Background**

According to Schwab (2021), the context in which the economist Simon Kuznets was involved fostered his curiosity. In the midst of the Great Depression of 1929, Kuznets tried to understand the world around him. Kuznets (1934) in his report to the National Bureau of Economic Research releases his work on National Income and produced the Gross National Product (GNP). A metric to measure how much of goods and services from American companies based in the US and abroad had been produced in the period between 1929-32.

Still according to Schwab (2021), by the time the Bretton Woods agreement came to place in 1944, the Gross Domestic Product (GDP), an adapted version of GNP focused solely on the sum of all goods and services generated inside the borders of a given country.

This approach is still prevalent and used as the norm in measuring the economic output of a country. And, according to the Office of National Statistics (ONS) (2022), there are three ways to measure the GDP: through the output (which are the goods and services produced in the economy), through the expenses made (capital allocated through business and expenses made by the State and households) and income (profits coming from businesses, wages coming from households and receivables from taxes by the government).

Due to the GDP metric being a novelty of the 20<sup>th</sup> century, to have an estimate of the economy size in periods before its conception, Thomas & Dimsdale (2017), in their Bank of England have conducted a thorough approach on researching various historical bibliographies to be detailed on the Section 3 of this thesis.

As mentioned by Oppers (2002), some theories might sound as old answers for the same old problems. The one mentioned in his paper embraces the individual subjectiveness, as an economic agent that by expressing its preferences and determining what is the utility of a good or service, generates a perception of value and the creation of prices (Huerta Soto, 2010). But its subjective approach ends right there, and the focus is leaned towards the productivity and innovation generation, objective and tangible elements as engines that back the well-functioning of the economy. Positioning itself completely in the opposite direction to the current economic system based on fractionary reserves and active control of the central

government in the matter of issuing currency and interfere in the interest rate as deemed convenient.

The objectiveness of the theory brought by Hayek (1996b) has been built on the vision of the capital-using production process provenient from savings due and postponed consumption time preference, brought by Eugen von Böhm-Bawerk's work, *Capital and Interest* (1884) (Von Mises & Ebelin, 1996). Huerta-Soto (2010) when citing Hayek (1996b), mentions that according to Bawerk's work, the increase in saving represses consumption and therefore, all other things constant, makes the relative price of consumer goods drop. This results in the "Ricardo Effect", consisting in the increase of demand for investment goods fomented by the increase in real wages, which happens, all other things constant, due to the decrease in prices of consumer goods generated by savings. Due to this fact, it is seen a relative increase in companies' profits in the furthest stages of consumption, which the products tend to have a higher value in a context of shrinking interest rates, due to the increase of savings. The result of such combined effects is the enlargement of the productive structure, which becomes more capital-intensive thanks to the financing allowed to the resources saved in larger amounts.

In contrast to the assumption posed before, Huerta Soto (2010), mentions that, when credit expansion and deposits through fiduciary means are done, a growth in the monetary supply, decrease in interest rate and enlargement of the economic productive structure takes place. Still according to Huerta Soto thoughts (2010), with more capital available economic agents will be more likely to take risks and invest into sectors that once were not that viable or were just too much capital intensive. There is a proportion of the capitalists' expenses and their new consumption capability. As a result, this distortion caused in the market creates new jobs and economic growth is fostered.

According to Huerta Soto (2009), the sectors that will benefit the most during the beginning of such monetary expansion are the ones the furthest from consumption, also known as capital intensive, such as factories, ship ports, airports, and the construction sectors. Still following Huerta Soto's (2009) logic, the development of such projects due to abundance of money increases the demand for productive factors, which since could not keep up to the growth pace of the new market trends are inflated.

The Business Cycle theory was first coined by Clément Juglar (1862), after observing that economic activity tended to follow a regular pattern of upswings and downswings, lasting approximately 10 years (Dal-Pont Legrand & Hagemann, 2007). His work has been complemented by many different economists after him, resulting in different versions of this same concept, that has its phases of prosperity, recession, depression, and recovery triggered by different factors. The Austrian Business Cycle Theory (ABCT), started with Ludwig von Mises, and later complemented by Friederich von Hayek, and their subsequent followers are emphatic on their critics of the fractionary reserve-based banking system and the ability of the central bank to expand the money supply at will. Hoppe et al. (1998) states that any contractual agreement involving individuals that simultaneously own the same thing (or alternatively, the same thing as simultaneously owned by more than one person) is an objective fraud. The fractionary reserve system, although not directly involved in the expansion of the monetary base (M0), has the power to multiply the broad money (M1 and M2) and affect the way money is distributed in society. In addition, according to Garrison (2004), the theory developed by Mises and Hayek deals with unsustainable boom, which due to cheap-credit policies coming from the central bank it can lead the economy to a growth path that is incongruent with the economic reality present at the moment, making it in have synergies with Soto's statement on the topic.

Among the many variables observed in the data set provided by the Bank of England, it is especially important to bear in mind the ones relevant for the GDP growth according to the mentioned authors in the literature review. In this paper, an analysis regarding the weight of these variables for the GDP growth will be conducted and evaluated if they have significant weight for the mentioned dependable variable throughout the use of the Regularization Models applied to the forecasting intended.

### **Empirical Literature on GDP Forecasting and regularization models**

As technical references for papers that have utilized a similar approach for forecasting to validate and investigate an assumption the following works were used throughout the development of this thesis.

In the work of Zhang et al. (2019), he uses the Lasso and Elastic Net approaches, in order to forecast the future oil prices. His motivation to use the Lasso method relies on the assumption

that rational investors should only focus on the most powerful predictors and ignore all the variables. Due to its characteristics of setting the less relevant explanatory variable coefficients to zero (or close to it), these models attend exactly what he is aiming for and can be used as a reference for the development of this thesis. The conclusion found in this paper reports that, when compared to the other models they used for forecasting, including Ridge, it was found that Lasso was the best performing model, while the Elastic Net stayed behind for a small difference. This result makes clear that his assumption is correct, and investors are better off only focusing on a few variables to predict the oil price, since there is no significant marginal benefit with adding any of the other variables to the model.

The importance of adopting multi-frequency data is due to the reason brought by Striaukas et al. (2020), he explains that some economic variables are only released to the public in higher frequencies other than the yearly time frames. Such assumptions can be verified on the dataset used during this thesis, and in many monthly and quarterly reports that are released for the common public, such as unemployment.

The use of “bridge” models are a recurrent alternative (Baffigi et al. 2004, as cited in Siliverstovs (2017)) to address to such disparity in time series that need to be analyzed without losing much information and oversimplifying the Regularization Model by excluding variables that are not timely aligned. The most common use of these bridge models is to convert high frequency data, through averaging out the values of a high frequency data frequency to fit a lower one. The aggregation method is one bridge method, and this has been used by Petropoulos (2014) as a way to lead substantial improvements in terms of his forecasting performance.

There is a work that holds some similarity to this paper, and it should be referenced here. It was done by Anesti et al. (2021) and had the goal of forecasting the monthly British GDP on a monthly basis, we could state that, in the data chosen for her forecasting, she created a high dimensional dataset while selecting the variables from many different sources. This gave her more robustness in her Lasso regression, which performs, on par to the Elastic Net. This paper can be used as a reference to verify if the loss of information generated by the aggregation methods used in this thesis to convert the high frequency data into a lower one will generate significant different results, also, due to the limitations of the dataset used in this thesis (restricted to the Millennium Data provided by the Bank of England), it does not

possesses a high dimensional framework that could benefit the LASSO regression (and the other regularization models) as much as this one.

The code used as reference for the development of this thesis regarding the regularization forecasting models processes relied on the work done by Medeiros et al. (2016) and their inflation forecast using Lasso and Elastic Net.

### **3. METHODOLOGY AND DATA**

In the Methodology and Data section, it is outlined the methodology and data employed in this thesis. To make a thorough study on the Millenium Data provided by the Bank of England, and a comparative forecasting model, first, it briefly mentioned the research design of the study while addressing each step taken to reach the results. Second, it is explained which specific variables are being used, their characteristics and which data treatment they went through. Lastly, it will be explained how the models adopted work in further detail and explain how they were conducted for the dataset to be analyzed.

#### **3.1. RESEARCH DESIGN**

This thesis brings a quantitative design to understand how different variables with different time frequencies can correlate to the UK GDP through the thousand years of data provided by the Bank of England (BoE). The main objective for this work is identifying which variables are these and using them to create a forecast to be compared to a benchmark. The following steps were adopted throughout the research:

1. Through the dataset provided by the BoE, identifying which variables had compatible years. This is crucial, because, although the method employed heavily benefits from a high-dimensional data, it is still pivotal for the data to be run that the data desired to be analyzed should have information in similar periods. After observing the dataset, the years that provide the most occurrence of information are between 1870 and 2006. Therefore, this is the timeframe used for the research.
2. While selecting the data variables, it is also considered their timeframe. On the data available, there are variables which are not in the same frequency and therefore, some aggregation method, where, in this work, it transforms higher frequency data into lower frequency should be applied.
3. After selecting which data should be used, stationarity tests have been conducted to test if they were ready for a regression and forecasting. For the non-stationary time series, differencing methods were applied.

4. Finally, the regressions are done and compared. An AR (5) model is run for the yearly UK GDP, as a benchmark, while the Lasso, Ridge and Elastic Net models will be run to identify which of the variables analyzed on the curated dataset of the BoE are the most relevant to the yearly UK GDP growth at different horizons and through different window methods, while subsequently a forecast will be taken to be compared to the benchmark.

### **3.2. DATA AND ANALYSIS**

The data used and analyzed in this study has been provided by the Bank of England, through the work provided by Thomas & Dimsdale (2017). As mentioned in the Literature review, the GDP is a metric coined in the 40s, therefore, authors needed to estimate what this value would be for the period before this metric became the norm for measuring economic output.

As mentioned by Thomas & Dimsdale (2017), it is possible to measure the GDP through expenditure, income, and output approaches. The following methods are a guide to the works he used as reference for the estimates of the yearly UK GDP at Market Prices, used for this thesis.

For the years from 1948 onwards, the source for their yearly GDP has been the Office of National Statistics (ONS). For the years between 1920 and 1948, a balanced estimate was made by Sefton & Weale (1995), where they relied on Feinstein's (1972) GDP estimates, made by the implementation of the three mentioned approaches. Further, they used a least-squares model to assign weights to the three estimates.

Between the periods of 1841 to 1913, there were some steps to be taken to estimate the UK GDP of this period. To start, for the 1841 to 1870 frame, output-based estimates for the UK were created with the data from Broadberry et al (2015) and Andersson & Lennard (2016) for the British Islands and Ireland. Through an annual chain-linking of constant price estimates, it was provided a new output-based volume series for the UK. Having this output-based estimate, a compromise measure took place, combining Feinstein's expenditure and income estimates, which were benchmarked to known points. From 1870 to 1913 the work of Solomou and Weale (1991) with balanced estimates for GDP from has been used to expand the series to

1913, which had its data extrapolated to 1920 through Feinstein's compromise estimate for the UK, found in Mitchell (1988). For more detailed information on the way this data has been treated, checking the complete work of Thomas & Dimsdale (2017) would be of great educational value.

During this thesis, the time series gathered by the author to forecast the dependent variable, given the criteria outlined in the Research Design, were fifty-five (55).

The Table 1 below presents all the names, time frequency and categorization of the independent variables:

**Table 1** – Presentation of Dataset utilized.

<b>Time series</b>	<b>Category</b>	<b>Time Frequency in the original dataset</b>
Real Consumption	National Accounts - 1	Yearly
Real Investment	National Accounts – 1	Yearly
Stock building Contribution	National Accounts – 1	Yearly
Real Government Consumption of Goods and services	National Accounts – 1	Yearly
Export Volumes	National Accounts – 1	Yearly
Import Volumes	National Accounts – 1	Yearly
Population	Labour, Capital, and Productivity - 2	Yearly
Employment	Labour, Capital, and Productivity - 2	Yearly
Unemployment	Labour, Capital, and Productivity - 2	Yearly
Average Weekly Hours Worked	Labour, Capital, and Productivity - 2	Yearly
Capital Services whole economy	Labour, Capital, and Productivity - 2	Yearly

TFP Growth	Labour, Capital, and Productivity - 2	Yearly
Labour productivity per head	Labour, Capital, and Productivity - 2	Yearly
Labour productivity per hour	Labour, Capital, and Productivity - 2	Yearly
Labour share whole economy excluding rents	Labour, Capital, and Productivity - 2	Yearly
Spliced Wholesale Producer Price Index	Wages and Prices - 3	Monthly
GDP Deflator at Market Prices	Wages and Prices - 3	Yearly
Export Prices	Wages and Prices - 3	Yearly
Import Prices	Wages and Prices - 3	Yearly
Terms of Trade	Wages and Prices - 3	Yearly
Earnings per Head	Wages and Prices - 3	Yearly
Oil prices	Wages and Prices - 3	Yearly
Consumer price index	Wages and Prices - 3	Yearly
Consumer price inflation	Wages and Prices - 3	Yearly
Real consumption wages	Wages and Prices - 3	Yearly
Bank Rate	Financial Markets - 4	Yearly
Spliced Series for Discount Rate on Prime Short-Term Paper 1718-2005	Financial Markets - 4	Monthly
Spliced Bank Deposit Rate Series	Financial Markets - 4	Monthly
Long Term Consols Yield 1753-2015 Corrected for Goschens' Conversion Issues	Financial Markets - 4	Monthly
Spliced Monthly Share Price Index Weighted by Market Capitalisation	Financial Markets - 4	Monthly
Monthly Corporate Bond Yields 1854-2015	Financial Markets - 4	Monthly
Monthly Corporate Bonds Spreads 1854-2016	Financial Markets - 4	Monthly
Exchange Rate	Financial Markets - 4	Monthly

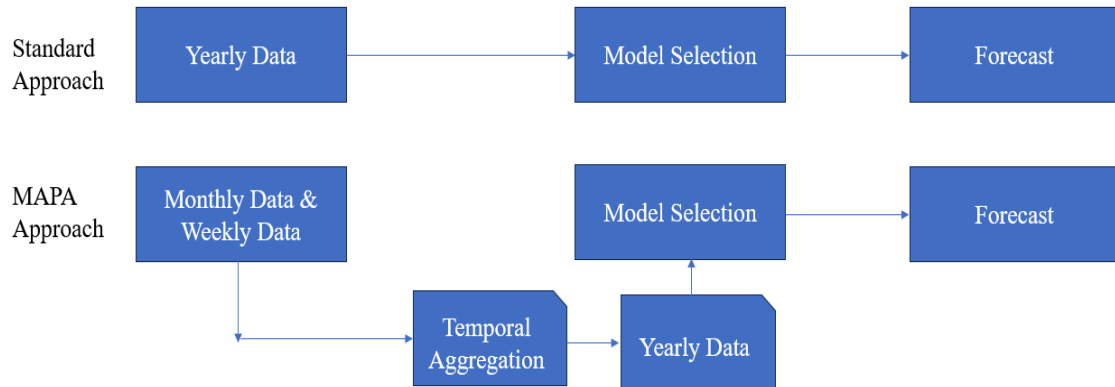
Government Securities	Financial Markets – 4	Weekly
Discounts Advances and Other Securities	Financial Markets – 4	Weekly
Total Government Securities	Financial Markets – 4	Weekly
Other Securities	Financial Markets – 4	Weekly
Real ERI	Financial Markets – 4	Yearly
House Price Index	Financial Markets – 4	Yearly
Notes and Coin in Circulation 1870-2015	Money and Credit- 5	Monthly
Monetary base M0 -1870-2015	Money and Credit- 5	Yearly
Reserve of notes and coin	Money and Credit- 5	Weekly
Public Deposits	Money and Credit- 5	Weekly
Other Deposits and Accounts	Money and Credit- 5	Weekly
Bankers Deposits	Money and Credit- 5	Weekly
Total Liabilities	Money and Credit- 5	Weekly
Memo Fiduciary issue	Money and Credit- 5	Weekly
Notes in Circulation	Money and Credit- 5	Weekly
Notes held by the Bank in Banking Department	Money and Credit- 5	Weekly
Broad Money	Money and Credit- 5	Yearly
Public Sector Net Lending Borrowing of GDP	Fiscal- 6	Yearly
UK Public sector debt	Fiscal- 6	Yearly
Trade deficit	Trade - 7	Yearly
Current account	Trade - 7	Yearly
Current Account Deficit Including Estimated Non-Monetary Bullion Flows Prior to 1946	Trade - 7	Yearly

On total, there are thirty-three (34) yearly variables, nine (9) monthly and twelve (12) weeklies. These data have been selected considering the already categorization done by the Bank of England when this dataset was released. Only the weekend time series were not labeled by them. This allowed the author to determine which label suits them

the best. The categorization for such variables will be useful to address, through a Ridge model, which categories hold the most significant variables to the model.

The weekly dataset happens to be mostly regarding financial operations, more precisely banking and financial markets. According to Biondi (2018) there is an undeniable connection between banking, credit, and money, therefore, for the variables related to banking operation, they were assigned the label “Money and Credit”, while the ones related to securities were assigned as “Financial Markets”.

After selecting the variables to be analyzed on the forecast processes, a transformation process takes place. The most usual way to deal simultaneously with daily, monthly, and yearly variables is to average out the higher frequency variables, and therefore reducing it, to homogenize their sampling frequency with respect to the lowest frequency one. According to Petropoulos & Kourentzes (2014), this approach known as Multiple Aggregation Prediction Algorithm (MAPA) can be illustrated with the following diagram.



**Figure 1** – Comparison between Standard Approach and MAPA.

The MAPA approach applied to the data studied will result in the temporal aggregation of the higher frequency data. Therefore, the higher frequency time series, to be converted to low frequency will be. To transform monthly data (1 year = 12 months), we define the total amount of months in the data set used (from 1870 to 2006), which is 1644, divided by the number of months in a year (12). This gives us

all the monthly data converted by year. The mentioned process generates 137 aggregated observations that will match the required yearly frequency.

The same is done for the weekly data but considering the nuances relevant to the higher frequency characteristic of this time series. In one year, there are 52 weeks, which generates 7124 time series that are aggregated by using simple summation relevant to the number of weeks in a year. Resulting in the same 137 observations needed for the low frequency data.

To prepare the data to be used for forecasting, the dataset has also been tested for stationarity through the Autocorrelation Function (ACF) test and scaled and double checked by the Augmented Dickey-Fuller test, as also mentioned in Anesti et al. (2021) in her Autoregression analysis. Due to the differencing process, resulting in a slightly shorter data dimension. There were variables that had to be double differencing, resulting in the observed data now having been reduced to its starting point at the year 1872, culminating in a total of 135 observations to be used in the forecast. In an Auto Regressive (AR) model, the data must be stationary to provide a constant mean and variance over time, since the model assumes that the future behavior of the time series will follow the same patterns that happened in the past.

The scaling process happens for the purpose of providing a fair comparison between the different present variables, since regularization techniques penalize large coefficients, and mitigates the impact of outliers in the dataset. The scaling process adopted in this thesis is the Standardization method, made by subtracting the mean and dividing by the standard deviation of each time series to be analyzed. Resulting in the standard deviation of every time series being 1 and mean equal to zero. (Hastie et al., 2016). According to Giglio & Brown (2018), when the predictors are scaled, their products and the partial residual are effectively measuring the correlation. On the other hand, when the predictors are mean-centered, the product ends up measuring the covariance. Due to sensitivity to variance, in this thesis it was chosen to use scaled data to be worked with.

The ADF test, for the UK GDP, variable that will be used in the Autoregressive model, AR (5), after reached stationarity is:

**Table 2 - Augmented Dickey-Fuller test for UK GDP.**

---

Dickey-Fuller: -5.4137	Lag Order: 5	p-value: 0.01
------------------------	--------------	---------------

---

After testing the lag order of 5, compared to the lag order of 1, it is found that an AR (5) presents a lower RMSE than an AR (1), a reference initially considered for analysis since this was the approach taken by Anesti et al. (2021). Also, although the AR(1) possesses a slightly lower AIC value, when conducting a Ljung -Box test to have check the hypothesis of the residuals being white noises, the AR (5) model accepts the null hypothesis more emphatically, while the AR (1) is too close from rejecting the null hypothesis.

**Table 3 - Comparison Table AR (1) and AR (5).**

---

<b>AR (1)</b>	AIC: 369.24	Box-Ljung test: 0.946	RMSE: 0.938
<b>AR (5)</b>	AIC: 369.98	Box-Ljung test: 0.7941	RMSE: 0.9122

---

Due to the time series structure to be used in the regularization models, the data model is divided in two. The first half of the observations corresponds to the data model to be trained and the other half to the model to be tested. This is done to improve the performance of the regularization models, since they have the best performance when applied to a high dimensional dataset. By using this approach, the researcher of this thesis transforms the dataset to be used on the regularization models closer to a high dimensional environment, where the number of observations (p) is greater than the number of variables (n), as mentioned in Emmert-Streib & Dehmer (2019).

### **3.3. METHODOLOGY**

In statistics, the Ordinary Least Squares model (OLS), also referred as Linear Regression, is the most common way to estimate linear models. The main purpose of this model is minimizing the residual sum-of-squares of the dependent variable by

selecting coefficients for the independent variables that minimize this value (Hasti et al, 2016).

If the OLS assumptions<sup>1</sup> are met for a linear regression, this method would be enough to explain the relation between variables and make predictions assertively. When someone has to make sense out of real world data, these assumptions are hard to be met, thus, having the opportunity to apply models that work as a supplement to the OLS are the effective ways to address the shortcomings of the Linear Regression (Doreswamy & Vastrad, 2013).

Overfitting is the most prevalent problem that occurs when using an OLS model which will lead it to a bad performance. This phenomenon happens when there are too many independent variables, and they are highly correlated with one another (also called multicollinearity). When multicollinearity takes place, the OLS estimator will have a low bias, although it will have a large variance (Emmert-Streib et al., 2019). Such big variances will result in the production of bad estimates as its byproduct. According to Kaufman (2013), another problem that can be generated is heteroskedasticity, which causes biased and inefficient estimates of the OLS model<sup>1</sup> standard errors.

When faced with this issue, there are two ways to solve it:

1. Reducing the number of explanatory variables:

Reducing the complexity of the OLS model by not including variables to the model. Such action would defeat the purpose of this work. The idea behind this paper is testing which are the best variables to explain the GDP growth of the UK in the studied period.

2. Applying Regularization Models:

These regression methods apply a shrinkage system which reduces the importance of the least relevant features in the dataset. In this example given, would be the ones with a high variance.

---

<sup>1</sup>The OLS assumptions are, according to Schmelzer et al. (2023): Error term having a conditional mean of zero; independently and identically distributed data; and absence of large outliers. It is also possible to add to it the addition made by Stock & Watson (2020), where they also listed not having perfect multicollinearity.

Due to the limitations of the presented model, this thesis will take advantage of Regularization models, which can be considered an enhanced version of an OLS model and an effective way to extract meaning out of the data available. Models such as: Ridge, Lasso and Elastic Net are the ones being chosen for this thesis due to their shrinkage and variable selection nature. Precisely what is needed for the purpose of identifying the most significant variables that can explain the GDP development of the UK throughout the years. In the following sections of this thesis, there will be introduction of these regularization models used and their strengths.

### **3.3.1 Least Absolute Shrinkage and Selection Operator (LASSO) Regularization (L1)**

According to Medeiros et al. (2016), the Lasso model, also known as L1 was first proposed by Tibshirani (1996), and the idea behind this method is shrinking less relevant coefficients in a regression to zero through a penalization process. This method is widely adopted in econometrics and forecasting, just as in the crude oil forecasting prices brought by Zhang et al.(2019).

Regularization models work as a complement to the OLS model, the difference lies in the application of the lambda penalization applied to the model. According to Medeiros et al. (2016), the Lasso estimator can be defined as the following (in its Lagrangian form):

$$\hat{\beta} = \arg \arg || Y - X\beta ||^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (I)$$

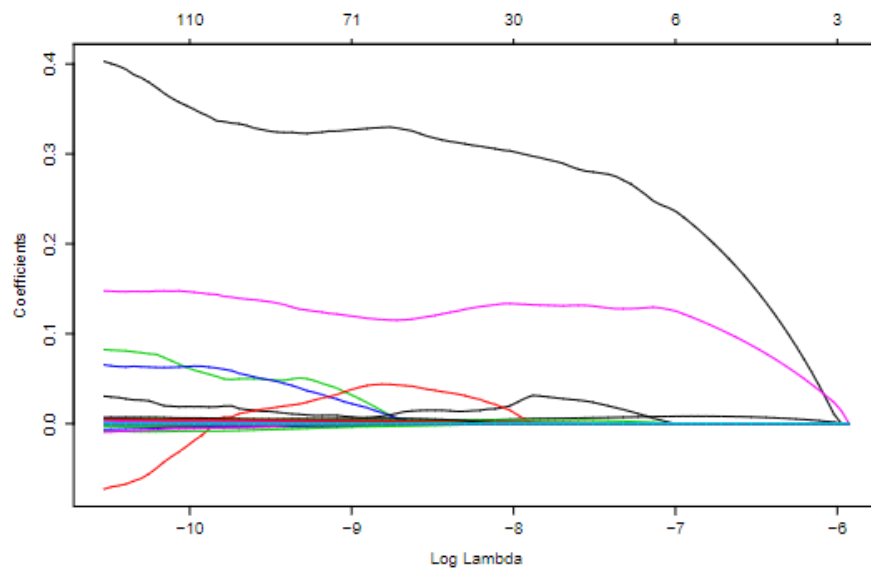
Where  $\hat{\beta}$  is the LASSO estimate,  $\beta$  stands for the vector of parameters, Y is the correspondent independent variable to be analyzed, X in the P x n data matrix<sup>2</sup> (covariate matrix) and  $\lambda$  is the shrinkage parameter for coefficients (tuning parameter). This is a constraint on the model parameters that shrinks the regression coefficients to zero, that is by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value ( $\lambda$ ).

---

<sup>2</sup> P are the covariates and n are the number of observations

Still according to Medeiros (2016), the selection of the shrinkage parameter has to be chosen between the interval of the shrinkages parameters between the largest penalty parameter that delivers the same result an OLS model (which means, it has not shrinkage); and the smallest penalty parameter that results in a model only with a constant term.

In the example provided by Medeiros (2016), we can observe that, in his inflation forecasting example, as the shrinkage parameter is increased, the number of variables and the size of the coefficients decrease.



**Figure 2** – Example of a Shrinkage parameter on a Lasso Model.

On the top of the chart, it is possible to see the number that have not been taken out from the model throughout the increase of the shrinkage parameter.

Such a regularization method is applied in the work presented by Zhang et al. (2019), as the most effective method on his paper to make variable selection. The objective of their paper is achieving parsimony in their dataset through variable selection that shrinks to exactly zero the variables that do not present relevance to the model.

In addition, Emmert-Streib et al.(2019) in his paper, brings the need for the  $\lambda$  parameter, which can be accomplished through a cross-validations process. In the section 3.3.4 of this work, it will be explained why this thesis' author has chosen to follow (Wang et al., 2007) approach for the tuning parameters.

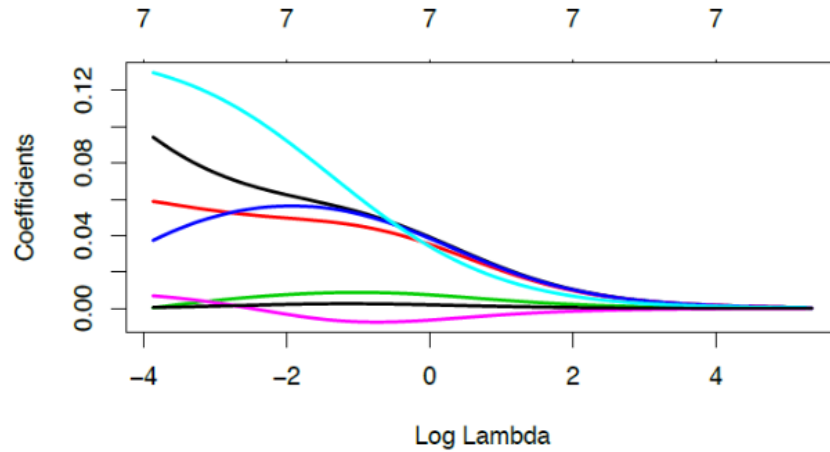
### 3.3.2. Ridge Regularization (L2)

According to the work of Hoert et al. (1970), pioneers in the regularization methods, and assuming that regularization methods are enhanced versions of the OLS model it is correct to state that the Ridge Regression (L2), shares the same equation structure as the lasso one LASSO, although the difference lies on the penalization section of the equation:

$$\hat{\beta} = \arg \arg || Y - X\beta ||^2 + \lambda \sum_{j=1}^p |\beta_j|^2 \quad (\text{II})$$

According to Anesti et al. (2021) this model works with the penalization of the residual sum of squares (RSS) with the sum of squared coefficients. Such a penalization results in variables that have coefficients with inferior contribution to the model to be shrunk substantially and approaching zero, but never becoming zero.

Such behavior can be seen in the example provided by Emmert-Streib et al.(2019) below, where the shrinkage parameter is increased and, as consequence, all the variables lean towards zero, although it never reaches it.



**Figure 3** – Example of a Shrinkage parameter on a Ridge Model.

Still following the line of thought of Emmert-Streib et al.(2019), a ridge regression can reduce the variance present on the dataset, at the cost of an increasing bias as a byproduct. The accuracy prediction of the model can be increased and it has the best performance in situations where the OLS estimates have a high variance and when there is not a high-dimensional dataset. By acknowledging this fact, it is expected that the ridge regression comes to have a good performance in the dataset selected for this paper.

### 3.3.3. Elastic Net

As mentioned by Altebany (2021) in quotation to Friedman et al. (2010), the Elastic Net method is a combination of the Lasso and Ridge. Through this model, shortcoming from both models can be solved by staying somewhere between these two, this results in a powerful tool for variable selection. The Elastic Net method was first introduced by Zou & Hastie (2005).

By considering the equation (I) and replacing the previous vector of parameters ( $\beta$ ) for the variable  $P_{\alpha}$ , representing the Elastic Net penalty, therefore, we have:

$$\hat{\beta} = \arg \arg || Y - X\beta ||^2 + \lambda P_{\alpha}(\beta) \quad (\text{III})$$

This new variable represents the combination of both Lasso and Ridge Regularization, where, as explained in the paper written by Emmert-Streib et al. (2019), in this combination of both models,  $\alpha$  represents a parameter which is going to balance them.

$$\begin{aligned}
 P_{\alpha}(\beta) &= \alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \\
 &= \sum_{j=1}^p \alpha \beta_j^2 + (1 - \alpha) |\beta_j|
 \end{aligned} \tag{IV}$$

The closer  $\alpha = 1$ , the more the Elastic Penalty will lean towards a Ridge model, while when the opposite applies, it means that the model is moving towards a Lasso.

Still following the thought of Emmert-Streib et al. (2019), while the ridge regression is optimal when independent variables are abundant and all of them will have non-zero coefficients, even when they are collinear. In a hypothetical case where two of these variables might have the same estimates of coefficients and, or are multicollinear, the Ridge will shrink them both, while Lasso will tend to choose one, while dropping the other from the model.

### 3.3.4. Tuning Parameters

According to Wang et al. (2007), after obtaining the initial estimator, it is needed to proceed with the selection of the tuning parameters in the iterative process to complete the whole algorithm. Cross-validation is a common way to find the optimal parameter for the lambda (shrinkage parameter), but, in their quotation to Shao (1997), if there is a true model among the candidates and it has a finite dimension, a BIC can be useful and yield good results. By following the steps of Medeiros et al. (2016) and the interpretation given of Wang et al. (2007), the BIC is a reliable alternative for this thesis' data parsimony, since it is close to be high dimensional and lagged factors are being used here.

## 4. RESULTS

In this topic the results extracted from the data analysis will be presented and divided in two subtopics. The first one shows a comparison between the benchmark and methods applied to the dataset and the interpretation of these results. On the second, a qualitative approach is brought to explain why the independent variables selected for the best performing model are relevant to understand its relations to the dependent variable.

### 4.1. COMPARISON OF REGULARIZATION METHODS AND BENCHMARK

In this section, the root mean Squared Error (RSME) results for the forecast are posed against the benchmark, providing its relative result. The lower the table result (on Tables 4 and 5) are, the better is the model compared to the AR (5), since the formula for this relative result is the regularization model RMSE divided by the AR (5) RMSE values. The RMSE metric is being used for comparison since it is a reliable and widely accepted source of accuracy for forecasting models.

**Table 4** - Expanding Window Results Relative RMSEs compared to AR (5).

	<b>Horizon 1</b>	<b>Horizon 2</b>	<b>Horizon 3</b>
<b>LASSO (alpha = 1)</b>	93.1	104.8	111.3
<b>RIDGE (alpha = 0)</b>	90.8	91.3	91.4
<b>ELASTIC-NET (alpha = 0.5)</b>	82.3**	82.9***	82.8***

**Table 5** - Fixed Window Results Relative RMSEs compared to AR (5).

	<b>Horizon 1</b>	<b>Horizon 2</b>	<b>Horizon 3</b>
<b>LASSO (alpha = 1)</b>	105.5	126.2	131.0
<b>RIDGE (alpha = 0)</b>	90.4	91.1	91.2
<b>ELASTIC-NET (alpha = 0.5)</b>	82.0**	93.5	96.6

According to the results from these models, the best result is the Elastic-Net forecast for a horizon period of 1 (one), although there is only a minor advantage when using the Fixed Window approach, instead of the Expanding Window. The Expanding Window method performs better on the Horizons 2 (two) and 3 (three) for the Elastic-Net forecast.

Through the work of Elliot and Timmermann (2016) on economic forecasting, it is the standard for this field to evaluate and compare forecasting model performances with the use of different forecasting windows and the use of the Diebold-Mariano test. According to the authors, for fixed estimation window, the estimation of parameters is fixed to an initial data sample of a certain observation range that does not get updated and incorporated to the training set for each subsequent forecasted value in the future. For the expanding window case, the window the data is trained on gets progressively bigger with each new prediction made and added to the dataset to be used for prediction. These predictions, on total, are done gradually one-step ahead at a time (West & McKraken, 1998).

The Diebold-Mariano test (Diebold & Mariano, 1995) is used in econometrics to compare if two forecasting models are significantly different from one another. After running the test, for all the models used in this thesis, it is shown that, only the Elastic Net model from Fixed and Expanding windows on the first horizon can reject the null-hypothesis for the test on a 5% significance level, which for this test it is the hypothesis of the two forecasting methods being accurately similar.

**Table 6** – Diebold-Mariano (DM) Test: Elastic Net on Horizon 1.

DM = 2.2158	Forecasting Horizon = 1	Loss function power = 2	p-value = 0.03016
-------------	----------------------------	----------------------------	-------------------

On Tables 4 and 5, for the \* marks it means the statistical significance level, for the DM test, where \* is for the 1% significance level; \*\* for 5% significance level and \*\*\* for 10%.

An acceptable tuning parameter for these models have been achieved by using the BIC. According to Medeiros et al. (2015), such an approach is used to seek the best hyper parameters and it delivers superior results when compared to cross-validation methods in terms of accuracy, and computational efficiency. Something crucial when running many different models in many different windows as done in this paper.

The poor performance of the Lasso model compared to the benchmark can be explained through its better performance being attached to a high dimensional data. Since the BoE data selected for this project, after being divided between testing and predicting datasets, comes close, but does not reach full high dimensionality, therefore, it is understandable why this model performed worse than the Ridge and Elastic Net, as expected (Emmert-Streib et al., 2019). This resulted in an excessive penalization, where most data were shrunk to 0, except for the Public Deposits. This highly sparse result generated by the Lasso model is also a result of strong correlated variables (Giglio & Brown, 2018), which has been confirmed after running a VIF test for multicollinearity. The influence of the Lasso regularization model is not completely neglected, since the high performance of Elastic Net relies on half of its structure related to Lasso model's influence upon it.

The Ridge model is the second-best performer, which means that, although not a proper method for variable selection, it shows the degree of relevance of the data studied in this thesis to explain UK GDP yearly growth. By using the most adequate Ridge model of the Table 5 on the Horizon = 1, it is possible to extract the most

relevant independent variables for the model. The Table 7 presented below ranks the five most important data points according to their magnitude:

**Table 7 – Relevance Rank of Independent Variables.**

<b>Independent Variable</b>	<b>Coefficient</b>	<b>Categories</b>
Other Deposits and Accounts	-1.33125E-34	Money and Credit- 5
Discounts, advances, and other securities	-6.00541E-35	Financial Markets – 4
Other Securities	1.925E-35	Financial Markets – 4
Public Deposits	-1.7996E-35	Money and Credit- 5
UK Public sector debt	-1.411E-35	Fiscal - 6

In contrast to regression models, a negative coefficient attributed by the Ridge model to the independent variables, does not always represent a negative relationship with the dependent variable. The model emphasizes the magnitude of the coefficients, and that is why in this rank it is possible to see coefficients with varied signs.

The reason why the Elastic Net performed the best is due to the fact of its combination of L1 and L2 penalties. While the Ridge kept all the 55 coefficients shrunk, but never turning them to 0, it never got the efficiency of totally getting rid of the irrelevant ones that the Elastic Net reached. With this optimal solution, the most relevant variables, in order of importance, were: **Public Deposits; UK Public Sector Debt; and Real Government Consumption.** The categories for these respective variables are Money and Credit – 5; Fiscal – 6; and National Accounts – 1, respectively.

Although not specifically designed for grouping selection, due to the combination of L1 and L2 penalties, this model happens to present relevant properties for variable

selection, since it presents a flexible sparsity, where, when faced with multicollinearity between variables, the Elastic Net set the coefficients of some of these features to zero, effectively feature selecting more representatives from relevant groups analyzed when compared to the Lasso model (Tutz & Ulbricht, 2009).

## **4.2. INVESTIGATION ON RESULTS**

When considering the variables selected by the Elastic Net model, we can find endorsements in the literature on why these independent variables are deemed as relevant to explain the UK GDP yearly growth and its better performance on forecasting when compared to the AR (5) model.

- **Public Deposits**

- This variable relates to the weekly balance sheet of the banking department of the Bank of England. Although public deposits are commonly referred to as deposits made by the public in a certain institution, this is not the case. The Bank of England, as a central bank, does not operate in this way. The Public Deposits referred by Thomas & Dimsdale (2017) is the amount of government money held by this institution.
- The relevance of the public deposits lies on the Government's use of the central bank to perform various financial transactions, according to the Bank of England's information (Bank of England, 2023), it can be considered to be working as their own private bank and treasury's agent, resulting in operations that may be but are not only restricted to: risk management; banking for financial sector firms; and also wholesale banking services to the Government.
- By providing the UK Government with large-scale, financial resources at minimum credit risk, the UK Government can, not only finance their operations, but also expand the broad money supply (M1) and spend capital in investments they perceive as a priority to foster growth, a widely adopted way to warm up the economy in times of economic

downturn, a common approach used by economist and policy decision makers more aligned to Keynesian ideas (Jahan et al, 2014).

- **UK Public Sector Debt**

- This yearly variable represents the number of national debts, and according to Thomas & Dimsdale (2017) it has been done excluding public sector banks and includes estimated capital value of terminable annuities. In addition to Huerta Soto's (2010) assumption on the positive relation between Public Sector Debt and GDP, Checherita-Westphal & Rother (2010) reveals a concave relationship in the debt-to-GDP ratio, and its inflexion point would be roughly between 90-100% of the GDP.
- According to Hills & Thomas (2010), in the UK of the XX century there were many events that fomented the use of monetary and fiscal policies as alternatives to stabilize the economy, such as the Great Depression and the two world wars. In the Appendix B it is possible to visualize a chart that illustrate such dynamic by presenting the GDP of the UK combined with the Debt-GDP ratio of the country.

- **Real Government Consumption**

- This yearly variable relates to the government consumption of goods and services. It consists of government expenditures that benefit the population, such as national defense, public school education, and infrastructure repairs and construction (U.S. Bureau of Economic Analysis (BEA), 2019).
- As mentioned by Jahan et al.(2014), government consumption, or as mentioned by them "government purchases" are ways which the government may use to foster economic growth and balance. It is possible to see that during the following decades (50 to 80) Keynesian policies increased in popularity and were widely adopted, generating consumption growth in families and companies through fiscal and monetary policies incentives (Hills & Thomas, 2010). On Appendix C, it is possible to see the growth the Real UK Government Consumption,

compared to the Real Consumption Wages through the analysis period of this thesis. It is shown that, after the wars, there is a constant up trend on these variables. It is important to highlight the UK Government Consumption reaching, in the 70s, levels superior to the ones found in the First World War and the constant consumption growth of families throughout the whole period of analysis. The data available confirm the argument brought by the Hills & Thomas (2010).

## 5. CONCLUSIONS

This thesis has completed its purpose of exploring a thousand years of data provided by the Bank of England, filtering the ones that could be reasonably be used on this thesis. As a result, it provided an investigative analysis on the most meaningful variables to explain the yearly GDP growth in the UK through regularization models. The approach adopted in this thesis shows that conventional aggregation methods can work for regularization methods and perform better than an autoregressive model (AR (5)).

When compared to the work done by Anesti et al. (2021), it is possible to understand that, the work presented here did not lose relevance for not being supplied by a dataset as high dimensional as the one used on that work, or having an uniform frequency of data (it was only used monthly time series in that case). Another difference worthy to be pointed out between this thesis and that work is the type of data being used as predictors for the independent variable. In their model, the Lasso regression verified provides a better performance than the Ridge, making it verifiable empirically on this thesis that when the number of observations is bigger than the predictors, and high correlations between predictors also take place, the performance of Lasso tends to be worse than a Ridge regression (Emmert-Streib & Dehmer, 2019).

The conclusion that can be taken from the dataset used in this thesis for these models is that, when looking at the results provided by the best Ridge model on Table 7 and Elastic Net, all the data come from categories related to direct government influence acting as factors that mostly explains the yearly UK GDP.

This is well aligned to what the literature review brings mentioned by Huerta Soto (2009), and as a result, no variable related to entrepreneurship (the Category 2 - Labour, Capital, and Productivity) has enough magnitude for the sparse regularization models consider them as relevant (Lasso and Elastic Net), nor to be in the ranking of the five best performers on the Ridge model. The author chose to pay attention to this category due to entrepreneurs being economic agents who identify business opportunities, allocate resources, and drive innovation. Therefore, productivity related variables are a result of entrepreneurial actions that constantly seeks ways to produce goods and services better in order to get an edge over the competition.

Such a scenario makes the author reach the conclusion that, with the use of the Bank of England's dataset, it is verified that there is no leverage gained through government expenses and interference that can make any of the observed variables have a higher magnitude than these factors.

On the other hand, when applied the Elastic Net, the assumptions brought by Huerta Soto (2009) about the benefit of capital expansion and its effects on further economic stages, represented by the variable Real Government Consumption of Goods and Services, are proven to be of great relevance to the model and a relevant predictor to be used for the yearly UK GDP forecasting. Since the investments made by the government tend to go into capital intensive sectors, it was expected that unemployment rate; TFP (total factor productivity); and overall working and productive labor hours would benefit from it, since there is more economic activity. These expectations did not come to reality and they did not have a sound performance to the model.

Such a conclusion leaves the possibility of further investigation. Other variables can be added to the dataset so it can become high dimensional, creating a better performance for the Lasso model. Also, the Midas-sgLasso regression can be a good addition to the other three regularization methods brought in this paper. According to Marsilli (2014) the MIDAS method formally conceptualized by Ghysels et al. (2007) and Andreou et al. (2010) offers an economical weighting framework to manage distributed lags, enabling the explanation of low-frequency variables by utilizing higher frequencies without the need for aggregation procedures, like the one used in the development of this paper, which although efficient and useful, simplifies the higher frequency values through its process. The combination of sg-LASSO and MIDAS used by Striaukas et al. (2020) is also a group variables, where the most relevant variables from the most relevant groups are selected for the model, providing a similar structure of result to the Elastic Net used in this thesis, that takes advantage of being the intersection of two models.

Finally, it is also possible to do further investigations on the search of other predictors that have a bigger magnitude than government related ones for the GDP growth. By detecting such variables, it is possible to do an approach similar to the one done by Zhang et al. (2019) where there are only a few variables that are relevant enough to pay attention to the oil price prediction and trend. By understanding which variables can be leveraged by public expenses

and government interference, one can position themselves to have an investment performance above the market.

## REFERENCES

1. Thomas, R and Dimsdale, N (2017) "A Millennium of UK Data", Bank of England OBRA dataset, <http://www.bankofengland.co.uk/research/Pages/onebank/threecenturies.aspx>
2. Schwab, K. (2021). Stakeholder Capitalism: A Global Economy that Works for Progress, People and Planet. John Wiley & Sons.
3. What is GDP and how do we measure it? - Office for National Statistics (2022, December 12). Retrieved 4 January 2024, from <https://blog.ons.gov.uk/2022/12/12/what-is-gdp-and-how-do-we-measure-it/#:~:text=Three%20Measures%20of%20GDP&text=There%20are%20three%20distinct%20ways,income%20and%20government%20tax%20take>
4. Dal-Pont Legrand, M., & Hagemann, H. (2007). Business cycles in Juglar and Schumpeter [Post-Print]. HAL. <https://econpapers.repec.org/paper/haljournal/halshs-00454505.htm>
5. Huerta Soto, J. A Escola Austríaca: Mercado e Criatividade Empresarial: 2. Ed. São Paulo: Instituto Ludwig von Mises Brasil, 2010.
6. HUERTA SOTO, J. Money, Bank Credit, and Economic Cycles: 2. Ed. Auburn: Ludwig von Mises Institute, 2009
7. Hayek, F. Obras Completas de F. A. Hayek: Unión Editorial, Madrid, 1996b.
8. Hans-Hermann Hoppe, with Jörg Guido Hülsmann and Walter Block, "Against Fiduciary Media," in the Quarterly Journal of Austrian Economics, vol. 1, no. 1, pp. 21–22.
9. Murray N. Rothbard, The Mystery of Banking, 2nd ed. (Auburn, Alabama: Ludwig von Mises Institute, 2008), p. 98.
10. Biondi, Yuri. "Banking, Money and Credit: A Systemic Perspective" Accounting, Economics, and Law: A Convivium, vol. 8, no. 2, 2018, pp. 20170047. <https://doi.org/10.1515/acl-2017-0047>
11. Baffigi, A., R. Golinelli, and G. Parigi (2004). Bridge models to forecast the euro area GDP. International Journal of Forecasting 20 (3), 447–460.
12. Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. Econometric Reviews, 26(1):53–90.
13. Andreou, E., Ghysels, E., and Kourtellis, A. (2010). Regression models with mixed sampling frequencies. Journal of Econometrics, 158(2):246–261.

14. Medeiros, M. C., & Mendes, E. F. (2015). L1-Regularization of High-Dimensional Time-Series Models with Flexible Innovations. *Textos para discussão*, No 636, Department of Economics PUC-Rio (Brazil), p. 226.
15. Medeiros, M. C., Vasconcelos, G., & Freitas, E. (2016). Forecasting Brazilian Inflation with High-Dimensional Models. *Brazilian Review of Econometrics*, 36(2), 223–254. doi: 10.12660/bre.v99n992016.52273.
16. Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society. Series B (methodological)*. Wiley. 58 (1): 267–88. [JSTOR 2346178](https://doi.org/10.2307/2346178).
17. Anesti, N., Kalamara, E., & Kapetanios, G. (2021). Forecasting UK GDP growth with large survey panels (SSRN Scholarly Paper 3855557). <https://doi.org/10.2139/ssrn.3855557>
18. Babii, A., Ghysels, E., & Striaukas, J. (2020). Machine Learning Time Series Regressions with an Application to Nowcasting (arXiv:2005.14057). arXiv. <https://doi.org/10.48550/arXiv.2005.14057>
19. Banking services. (2023). Retrieved 17 December 2023, from <https://www.bankofengland.co.uk/banking-services>
20. Checherita-Westphal, C. D., & Rother, P. (2010). The Impact of High and Growing Government Debt on Economic Growth: An Empirical Investigation for the Euro Area. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1659559>
21. Von Mises, L., & Ebelin, R. M. (1996). *The Austrian Theory of the Trade Cycle and Other Essays*. Center for libertarian studies. <https://books.google.ee/books?id=0JigDXuzm9YC>
22. Doreswamy, & Vastrad, C. M. (2013, December 10). Performance Analysis Of Regularized Linear Regression Models For Oxazolines And Oxazoles Derivative Descriptor Dataset. arXiv.Org. <https://doi.org/10.5121/ijcsity.2013.1408>
23. Elliott, G., & Timmermann, A. (2016). Forecasting in Economics and Finance. *Annual Review of Economics*, 8(1), 81–110. <https://doi.org/10.1146/annurev-economics-080315-015346>

24. Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Machine Learning and Knowledge Extraction*, 1(1), Article 1. <https://doi.org/10.3390/make1010021>
25. Giglio, C., & Brown, S. D. (2018). Using elastic net regression to perform spectrally relevant variable selection. *Journal of Chemometrics*, 32(8), e3034. <https://doi.org/10.1002/cem.3034>
26. Hills, S. (2010). The UK recession in context—What do three centuries of data tell us?
27. Kuznets, S. (1934). National Income, 1929-1932. In *National Income, 1929-1932* (pp. 1–12). NBER. <https://www.nber.org/books-and-chapters/national-income-1929-1932/national-income-1929-1932>
28. Marsilli, C. (2014). Variable Selection in Predictive MIDAS Models (SSRN Scholarly Paper 2531339). <https://doi.org/10.2139/ssrn.2531339>
29. NIPA Handbook: Concepts and Methods of the U.S. National Income and Product Accounts | U.S. Bureau of Economic Analysis (BEA). (n.d.). Retrieved 13 January 2024, from <https://www.bea.gov/resources/methodologies/nipa-handbook/>
30. Oppers, S. E. (2002). The Austrian Theory of Business Cycles: Old Lessons for Modern Economic Policy? (SSRN Scholarly Paper 879306). <https://papers.ssrn.com/abstract=879306>
31. Petropoulos and Kourentzes—Improving forecasting via multiple temporal aggreg.pdf. (n.d.). Retrieved 2 January 2024, from [https://orca.cardiff.ac.uk/id/eprint/84329/1/Foresight%202014%20MAPA%20\(pre-print\).pdf](https://orca.cardiff.ac.uk/id/eprint/84329/1/Foresight%202014%20MAPA%20(pre-print).pdf)
32. Siliverstovs, B. (2017). Short-term forecasting with mixed-frequency data: A MIDASSO approach. *Applied Economics*, 49(13), 1326–1343. <https://doi.org/10.1080/00036846.2016.1217310>
33. Stakeholder Capitalism: A brief history of GDP - and what could come next. (2021, December 13). *World Economic Forum*.

<https://www.weforum.org/agenda/2021/12/stakeholder-capitalism-episode-1-a-brief-history-of-gdp/>

34. Wang, H., Li, G., & Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 63–78. <https://doi.org/10.1111/j.1467-9868.2007.00577.x>
35. West, K. D., & McCracken, M. W. (1998). Regression-Based Tests of Predictive Ability. *International Economic Review*, 39(4), 817–840. <https://doi.org/10.2307/2527340>
36. What Is Keynesian Economics? - Back to Basics—Finance & Development, September 2014. (n.d.-a). Retrieved 13 January 2024, from <https://www.imf.org/external/pubs/ft/fandd/2014/09/basics.htm>
37. Zhang, Y., Ma, F., & Wang, Y. (2019). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? *Journal of Empirical Finance*, 54, 97–117. <https://doi.org/10.1016/j.jempfin.2019.08.007>
38. Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statist. Sin.*, 7, 221-264.
39. Schmelzer, C. H., Martin Arnold, Alexander Gerber, and Martin. (n.d.). 4.4 The Least Squares Assumptions | Introduction to Econometrics with R. Retrieved 13 January 2024, from <https://www.econometrics-with-r.org/4.4-tlsa.html>
40. Stock, J. H., & Watson, M. W. (2020). *Introduction to econometrics*. Pearson. <https://thuvienso.hoasen.edu.vn/handle/123456789/11496>
41. Altelbany, S. (2021). Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study. *Journal of Applied Economics and Business Studies*, 5(1), 131-142. <https://doi.org/10.34260/jaebbs.517>
42. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1),1-22. DOI: 10.1163/ej.9789004178922.i-328.7
43. Hoerl, Arthur E.; Kennard, Robert W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics*. 12 (1): 55–67. doi:10.2307/1267351

44. Garrison, R. W. (2004). Overconsumption and Forced Saving in Mises-Hayek Theory of the Business Cycle. *History of Political Economy*, 36(2), 323–349.
45. Kaufman, R.L. *Heteroskedasticity in Regression: Detection and Correction*; Sage Publications: Thousand Oaks, CA, USA, 2013; Volume 172
46. Hastie, T., Tibshirani, R., Friedman, J., & Hastie, T. (2016). *The elements of statistical learning* (2nd ed.)
47. Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253-263.
48. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 2005, 67, 301–320.
49. Tutz, G.; Ulbricht, J. Penalized regression with correlation-based penalty. *Stat. Comput.* 2009, 19, 239–253.
50. Sefton, J and Weale, M (1995), *Reconciliation of National Income and Expenditure: balanced estimates of national income for the United Kingdom, 1920-1990*, Cambridge University Press.
51. Andersson, F and Lennard, J (2017), ‘Irish GDP between the Famine and the First World War: Estimates based on a Dynamic Factor Model’, Lund University Department of Economics Working Paper.
52. Broadberry, S, Campbell, B, Klein, A, Overton, M and van Leeuwen, B, (2015), *British Economic Growth, 1270-1870*, Cambridge University Press.
53. Mitchell, B R (1988), *British Historical Statistics*, Cambridge University Press, Cambridge.
54. OpenAI. (2022). ChatGPT (Dec 20 version), large language model, <https://chat.openai.com/>

**APPENDIX:**

**A) Root mean square errors of Regularization Models**

**Horizon = 1**

	<b>Expanding Window</b>	<b>Fixed Window</b>
	<b>RMSE</b>	<b>RMSE</b>
<b>Lasso</b>	849.66	963.66
<b>Ridge</b>	829.01	826
<b>Elastic Net</b>	751.1	748.56

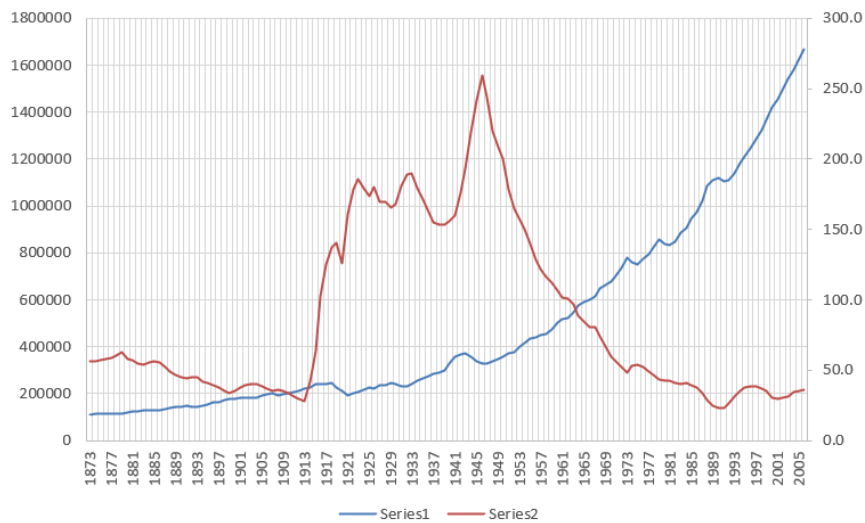
**Horizon = 2**

	<b>Expanding Window</b>	<b>Fixed Window</b>
	<b>RMSE</b>	<b>RMSE</b>
<b>Lasso</b>	956.32	1151.44
<b>Ridge</b>	833.69	831.78
<b>Elastic Net</b>	757.48	853.24

**Horizon = 3**

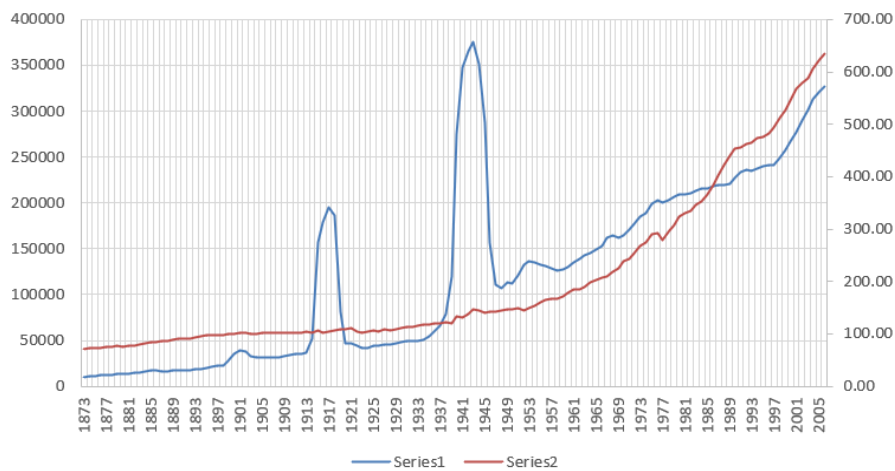
	<b>Expanding Window</b>	<b>Fixed Window</b>
	<b>RMSE</b>	<b>RMSE</b>
<b>Lasso</b>	1015.84	1195.27
<b>Ridge</b>	834.82	832.62
<b>Elastic Net</b>	756.87	882.03

## B) Combined Chart of UK GDP and Debt-GDP ratio



- Axis on the left, and Series 1 (blue line) is the UK GDP in millions of pounds (£)
- Axis on the right, and Series 2 (red line) is the Debt-GDP ratio of the UK in percentual (%)

## C) Real Consumption of UK Government



- The axis on the left, and Series 1 (blue line) is the Real Government Consumption of Goods and Services in millions of pounds (£)
- The axis on the right, and Series 2 (red line) is the Real Consumption Wages in the UK, where the 100 mark is the real consumption level in 1900, used as benchmark for the period analyzed

## Resümee

### **Aastatuhande Andmed - Kvantitatiivne uurimuslik analüüs, milles kasutatakse Ühendkuningriigi SKP kasvu regulariseerimismudeleid mitme sageduse andmete abil**

Töös uuritakse Ühendkuningriigi (UK) majanduse paljude majandusmuutujate vahelisi seoseid ning nende kaudu selgitatakse nende seoseid Ühendkuningriigi SKP aastase kasvuga. Uuringus kasutatakse Bank of Englandi andmekogumit, mis hõlmab üle tuhande aasta kestnud, iganädalasi, igakuiseid ja aastaseid andmeid. Kokku valiti analüüsimiseks 56 majandusmuutujat, sealhulgas Ühendkuningriigi aastane SKP, aastatel 1870-2006.

Teoreetiline alus tugineb selliste majandusteadlaste nagu Mises, Hayek ja Huerta Soto seisukohtadele, uurides ettevõtlust, tootlikkust ja valitsuse mõju kui SKP kasvu mõjutavaid tegureid. Doktoritöös võrreldakse prognoosimisel regulatsioonimudeli (Ridge, Lasso, Elastic-Net) ja võrdlusalusena kasutatud autoregressiivse mudeli AR (5) tulemuslikkust. Elastic-Net mudel, mis on Ridge'i ja Lasso kombinatsioon, on tuvastatud kui kõige täpsem kõigi olemasolevate akende ja horisontide puhul. Uuringus jõutakse järeldusele, et valitsusega seotud muutujad, eelkõige need, mis on seotud valitsemissektori tarbimise ja võlaga, mängivad olulist rolli Ühendkuningriigi SKP aastase kasvu selgitamisel, mis on kooskõlas Huerta Soto perspektiiviga, samas kui ettevõtlusega seotud muutujad, mis on liigitatud tööjõu, kapitali ja tootlikkusega, ei ole üheski uuritud mudelis olulisel määral esindatud.

Järeldus rõhutab valitsemiskulude olulisust SKP kasvu mõjutamisel, seades kahtluse alla eelduse, et ettevõtlusega seotud muutujatel on võrreldav tähtsus. Uuringus pakutakse välja edasiste uuringute võimalusi, sealhulgas täiendavate ennustajate uurimine, Midas-sgLasso regressiooni kasutamine ja selliste muutujate uurimine, mis võivad omada suuremat mõju SKP kasvule kui valitsusega seotud muutujad, mis tähendaks, et nad saavad nendest teguritest võimendavat mõju.

**Non-exclusive licence to reproduce the thesis and make the thesis public**

I, Pedro Marco Melo Silva, grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis Millenium Data – A Quantitative Exploratory Analysis Using Regularization Models on the UK GDP Growth Through Multi Frequency Data supervised by Luca Alfieri.

I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

I am aware of the fact that the author retains the rights specified in points 1 and 2.

I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

***Pedro Marco Melo Silva***

***16/01/2024***