Aix Marseille School of Economics

Master 2: Econometrics, Big Data and statistics

Macroeconomic Forecasting with a large French database

Author:
Coraline BEST

Professors: Sullivan Hue, Pierre Michel

March 9, 2024

Table of contents

1 Introduction

Among the first attempts to use multiple time series for macroeconomic forecasting, one can cite James H. Stock and Mark W. Watson 2002b and James H Stock and Mark W Watson 2002a. In these seminal articles, Stock and Watson presented a method for using factor augmented autoregressive models to forecast important macroeconomic variables such as Consumer Price Index (CPI), industrial production or unemployment. In James H. Stock and Mark W. Watson 2006 I can read that back in 2006 using multivariate time series for macroeconomic forecasting was already becoming the norm both for private companies and in academia. This move can be motivated by two main reasons. First, multivariate time series allow us to model interactions between variables and better understand the economic system. Second, adding additional time series can substantially improve forecast accuracy. This move to multivariate time series analysis was further accelerated by the increase in computing power, easier access to machine learning and statistical programming packages and large macroeconomic databases.

If today this method is the norm, it still has some challenges to tackle. The first one is related to the selection of variables. As macroeconomic data, like GDP or CPI are indexes aggregating large amounts of information, one could use a vast array of possible explanatory variables. Economic theory does not necessarily help a lot too, as phenomenons like business cycles, inflation or economic growth can be explained by many different theories. Therefore I am tempted to take all the available variables in one dataset to perform theanalysis without any pre-selection. This brings us to the second problem, related to the curse of dimensionality. Bellman 1961 illustrates how, to retain the same prediction accuracy with a larger number of predictors, one needs to increase in a greater proportion the amount of training data. The challenge lies in the availability of macroeconomic data, as the best-case scenario typically allows us to access monthly data only dating back to the end of World War 2. For example, in the application to French data, I must start the dataset in 2000 as many variables were not reported before. This leaves us with only a few hundred data points to perform the analysis. This means that I need to use models that either perform information reduction or variable selection. This is problematic as I am often not only concerned with prediction accuracy but also interpretability and analysis of the underlying economic phenomenons. Therefore I need to pay close attention to the working of the models. Finally one other challenge is forecasting uncertainty. For example central banks are concerned with controlling the inflation rate, and rely on forecasts to adjust their policies. A simple point forecast is not sufficient in this case, as high uncertainty would make it too imprecise and useless. Therefore the forecasting models need to be able to return prediction intervals to have a measure of the forecast's quality.

From these few points, I can present the study of this domain. In this report, I will present how I compiled and used a large french monthly macroeconomic dataset to forecast the CPI. I try to imitate as close as possible the FRED-MD data set as presented in McCracken and Ng 2016. In the case, I try to apply some machine learning models, the Random Forest, the Elastic-Net and the Support Vector Regression in the forecasting exercise. These machine learning models provide two hedges over the traditional econometric models. They can perform variable selection or information reduction, and can model non-linearities (except for

the Elastic-Net) and more complex relationships between the variables. The machine learning models are compared to one traditional econometric model for forecasting, the Autoregressive Moving Average (ARMA) model. In this forecasting exercise, I will tune the model accordingly to the constraints of time series data, perform Principal Component Analysis (PCA) and Partial Least Squares (PLS) to study their benefits, look at the variable selection capacities of the model for interpretability and use conformal prediction to retrieve prediction intervals to the forecasts. This will be performed in the database I have collected, having data from January 2000 to July 2023. This dataset regroups 69 macroeconomic variables from different sources. The forecasting exercise will focus on the period between January 2016 to December 2019, or 36 months. I made this choice as forecasting macroeconomic variables during the Covid-crisis can prove to be a complex task, as illustrated in Goulet Coulombe, Marcellino, and Stevanović 2021. I used a rolling window setup, with one step ahead monthly forecasts.

To study the question of macroeconomic forecasting with machine learning models, interpreting these forecasts and retrieving prediction intervals, I divided this report in the following parts. First I give some background on the history of inflation in France, and the economic use of forecasting inflation. I then provide a short literature review on the use of machine learning models and large datasets for macroeconomic forecasting. In a second part I present the data, the forecasting set up, the models used, the tuning strategy and the additional methods I have used : conformal prediction, PCA, PLS, the Augmented Dickey Fuller test and the Diebold Mariano test. Finally in a last part I will look at the performance of the models, and try to find some interpretation around variable importance and variable selection.

Overall I find that the best performing model is the Elastic-Net. Its performance is further improved by using factors of the original data transformed by PCA and to a lesser extent by PLS. This result was robust both when compared to a Diebold Mariano test and by studying the uncertainty of the forecasts with prediction intervals produced by conformal prediction. This result is coherent with the rest of the literature and shows that machine learning models utilizing large macroeconomic datasets combined with PCA can substantially improve the accuracy and the quality of inflation forecasts compared to some of the traditional econometric models.

1.1 Economic Problem

Macroeconomic forecasting has been early on interested in the prevision of inflation. Going back to the influential article Jr. 2003, the main goal of macroeconomic policy is to smooth out the effects of the business cycle that cause spikes and lows of consumption for the households. This smoothing happens through two main channels, monetary and fiscal policies. When I look at the definition of inflation given by the American Federal Reserve (Fed) [1], it is a sustained and global increase in prices of the economy, measured by a price index calculated as a representative bundle of goods (the Consumer Price Index). I see that controlling inflation is essential to ensure a smooth consumption for households across time, or otherwise they would rapidly lose their purchasing capacity. Even if multiple theories of inflation exist, being a monetary phenomenon, a commodities phenomenon or even being tied to the employment

[1] Link to the definition of inflation according to the Fed

rate, governments and central banks need to deal with it in a practical manner. To do so, inflation forecasting is an essential tool to plan their economic policies, and set the anticipation of the agents. The European Central Bank (ECB) and the Fed have set an objective, that can be relaxed in certain times, of a change in global price of 2% per year. From this, inflation forecasts are a tool to set the economic policy to either support inflation with monetary and fiscal expansion, like after the Great Financial Crisis in Europe, or juggle it through monetary and fiscal tightening like in the 1980s with Paul Volcker's policies in the USA.

If I go back to the French case [2], following the common methodology adopted in the Eurozone, the French central bank uses the MAPI model to perform its inflation forecasts (more details in Ulgazi and Vertier 2022). For long term forecasts, the model is semi-structural, it simulates the structure of the economy and is not only concerned with pure forecasting. However in the short run, meaning a few months, the model is fully flexible and is based on desegregated data and tries to find trends and correlations within the series to forecast inflation. In the exercise I stick to this second case, where I constructed a large macroeconomic database with multiple series broken into their sectoral components. I try to produce competitive short term forecasts using machine learning models and an ARMA model.

1.2 Small recent history of French inflation

In this part, I briefly trace the evolution of the dynamics explaining inflation in France since World War 2. This allows us to better understand the workings of this phenomenon and better understand the workings of the models.

Following World War 2, because of large amounts of debt owned to foreign investors, the value of the Franc sharply decreased, creating a large inflationary period peaking with a 58.7% increase in prices in 1948. This was increased by successive devaluations of the currency, causing a large imported inflation in a country with destroyed industrial capacities after the war. The inflationary episode was tamed during the 1950s through successive reforms to finally stabilize its exchange rate with the dollar, which was the only gold standard currency in the Bretton-Woods system, and having a status of reserve currency following the devaluations of the British pound.

In the following year, inflation would still fluctuate rapidly across months. This was due to the stop and go economic policy of the government at the time. Following Keynesian economic thoughts, governments implemented strong contra-cyclic economic policies with large fiscal spendings and decreases in the main rates during economic slowdowns, or strong fiscal contractions and spikes in interest rates for overheating periods. Inflation would follow this business cycle and in addition follow the devaluations of the Franc and political turmoils caused by the war in Algeria, the crisis of 1958 or the events of May 1968.

The next paradigm shift followed the end of the Bretton Woods agreements both caused by the too large circulation of foreign dollars compared to the American gold reserves. This caused the end of the fixed exchange rate regime in the West, causing inflationary episodes due to

[2] Link to a Banque de France blog post about its methodology

currency fluctuations. In addition the successive oil crisis and Iranian revolutions greatly increased energy prices which in turn raised the global prices of the economy. On top of these external shocks, the successive governments maintained an interventionist and Keynesian economic policy up until 1983 which fed the price-wages cycle. This caused inflation to shift from 4% on average during the 1960s to almost 10% between 1971 and 1985.

The situation slowly stabilized in the late 1980s with the effects of the austerity fiscal policies, the change of policy of the Central Bank and the implementation of the European Monetary System. Mainly, the Central Bank changed its policy goal from simply smoothing the spikes and downs in the business cycle in terms of employment and economic growth, to ensuring the stability of prices with a target of a 2% inflation rate. This is mainly due its newly found independence from the government, its inspiration from other Central Bank's policies like in the USA or Germany and the self imposed fixed exchanged rate with the Deutsche Mark which was at the time one of the most stables currencies.

The success of these policies, and the creation of the Euro, created a time of low and stable inflation around its 2% target from 1985 up to 2008, called the Great Moderation. Note that in 1991, the methodology behind the CPI was harmonized across Europe to give birth to the current measure of inflation.

The next change happened after the Great Financial Crisis. This even opened a period of non conventional monetary policies in order to restore confidence in the financial markets and foster demand after the recession. These policies were maintained up until the Covid-crisis as the Eurozone approached multiple times periods of deflation. Nonetheless even if the money in circulation in the economy greatly increased, the change in prices remained very low under the 2% target. The aftermath of the Covid-crisis started a new era of sustained price increased, pushed by bottlenecks in global supply, increases in energy and food prices following the war in Ukraine and large increases in money supply as argued by some.

This review highlights some of the complex causes of inflation. Inflation is a complex phenomenon that can be caused by multiple factors. When from the 1950s up until the 1970s inflation was caused by excess in the aggregate demand, or devaluations, the 1970s showed how global prices were dependant on energy prices. The 1980s showed the importance of wage growth and Central Bank's policies to juggle prices, when the aftermath of the Great Financial Crisis showed the limitations of monetary policy in managing global prices. From these few points I can cite some elements that are causes of inflation : exchange rates, status of a currency, aggregate demand, energy prices, interest rates and money supply, wage growth or bottlenecks in industrial production.

1.3 Literature review

I look in this subsection at the literature around the use of large databases and machine learning for macroeconomic forecasting. First I can cite a few large datasets. I focus on datasets having at least 100 variables using monthly data. In this category, the reference is FRED-MD for American data, presented in McCracken and Ng 2016. It both provides a large number

of economic series, around 120, with most going back to 1959. Reporting the same variables than FRED-MD, but for different countries, I can find a Canadian dataset in Fortin-Gagnon et al. 2022 and a dataset for the United-Kingdom in Goulet Coulombe et al. 2021. However compared to FRED-MD, most of the series are not available before the 1980s. For a larger dataset with around 200 variables concerning Japan, one can also at Maehashi and Shintani 2020, tracking data from 1974 to today.

After presenting the data I can make a list of horse races comparing the performance of macroeconomic forecasts using machine learning models. Among them I can find Medeiros and Mendes 2016, Medeiros, Vasconcelos, et al. 2021, J. C. Chen et al. 2019,Goulet Coulombe, Marcellino, and Stevanović 2021, Goulet Coulombe, Leroux, et al. 2022, Maehashi and Shintani 2020 and Milunovich 2020. These horse races compare their forecasts on one or two macroeconomic datasets, often having the same structure than FRED-MD. Most of the articles cited above use multiple forecasting periods and forecast for one or multiple macroeconomic variables. From all these horse races a few conclusions seem to emerge. First of all, models using many predictors seem to always improve or at least be on par with traditional univariate models. Second of all, non-linear models seem to give another hedge, but it is more localised depending on the data and the variable. For instance, in Goulet Coulombe, Marcellino, and Stevanović 2021 this hedge is during the covid crisis and periods of recessions in Goulet Coulombe, Leroux, et al. 2022, while in Milunovich 2020 and Maehashi and Shintani 2020 it is for long term forecasts. Other horse races focused on more specific subjects can include Kim and Swanson 2018 and Goulet Coulombe et al. 2021, focusing on how and which data transformations can improve forecast accuracy for models with many predictors. The conclusion seems that as a whole, using PCA and variations of PCA analysis to pre-process the data seems to improve the forecast accuracy of most models. Concerning specifically shrinkage models, one can also look at Li and W. Chen 2014 and Smeekes and Wijler 2018. They show how shrinkage methods able to perform variable selection can enhance the performance of factor models by helping to select the relevant factors in the data. The second also shows in its horse race how these sparse models can even perform well on data that has been generated by a data generating process with a factor structure. Both go in the direction that even if the data generating process is not sparse, sparse models can still perform well or can be used in combination with factor models to produce good forecasts.

Overall this literature can give us a few direction to orient the study. The first is that machine learning models can perform well in macroeconomic forecasting, especially when using large datasets. The second is that data transformations, especially PCA factors, can substantially improve the forecasting performance without any additional observations or variables. Finally, there is no unique best set-up. Most studies find that many models can perform well at the same time, and these models can have very different characteristics, being based on factors, being non linear or not, sparse or not, using bayesian set ups or not. The result seem dependant on the country, the time period and the forecasting horizon. Additional to that Goulet Coulombe, Leroux, et al. 2022 gives us a few insights on the best tuning practices for the models. K-Fold or time series cross validation are both valid method performing as good as each others. In addition, when possible, tuning models with the BIC is also a valid strategy and delivers forecast as good as the ones tuned on cross validation.

2 Materials and Methods

This section provides an overview of the dataset used, the models implemented, and the empirical strategies applied in the analysis.

2.1 Presentation of the data

This subsection presents the data used detailing its collection process and formatting. The data was collected from multiple French public and free sources such as INSEE, Banque de France, DREES, DARES. The data is revised and updated continuously. After handling missing values and dropping them, I have monthly data spanning from January 2000 to July 2023. Therefore, I have a total of 70 variable and 283 observations.

The dataset encompasses a diverse set of 69 macroeconomic series, offering a comprehensive view of various facets of the economy. These series are categorized into Economic Indicators, Household Consumption Insights, Business and Employment Metrics, Consumer Price Index, Financial Market Dynamics, and Monetary Aggregates, providing a good perspective on economic performance. Additionally, it integrates variables related to the Industrial Production Index, offering a comprehensive view across industries and goods. Variables intricately linked to unemployment metrics provide insights into labor market dynamics, while those derived from monthly household surveys contribute valuable data on consumer confidence trends. The dataset also includes variables that capture Commodity Prices and Exchange Rates, along with information on Interest Rates and Bonds. Notably, I have included variables related to construction and housing. It includes variables like the number of authorized housing constructions, housing starts, authorized building site areas for non-residential premises, and the area of non-residential building sites started.

The forecasted variable is the consumer price index using the consumer price index total variable. The dataset covers a wide range of categories, each offering valuable insights into various facets of the French economy. A comprehensive table with all variables is provided in appendix.

Stationarity is a key concept in time series analysis, and it is important for several reasons. First, stationarity simplifies the modeling process and improves the accuracy of the forecasts. Non-stationary series, by contrast, may exhibit trends or seasonality, introducing complexities that challenge the accurate modeling and prediction of future values. In a stationary time series, the mean and variance are supposed to remain constant over time. Conversely, non-stationary data can lead to varying mean and variance, resulting in models that are more intricate and less interpretable. This is why I performed the Augmented Dickey-Fuller (ADF) test, to check the stationarity of the series. This test was originaly proposed by Dickey and Fuller 1979. In the context of the ADF test, the null hypothesis posits a unit root and non-stationarity, while the alternative hypothes is suggests stationarity. The test statistic from the ADF test is compared to critical values to determine whether to reject the null hypothesis. If the p-value is less than a chosen significance level (commonly 0.05), I would reject the null hypothesis and conclude that the series is stationary. Following the application of this test, I identified 11 series as stationary, leaving a total of 58 non-stationary series. In order to

achieve stationarity, I applied first-order differentiation to all non-stationary series, proving effective in making them stationary. This process was repeated until I achieved stationarity for all variables, but no variables needed to be differenced twice in the case. The forecasted variable, the CPI was also taken in differences to be stationary, this means that in the forecasting exercice I do not directly forecast the CPI index, but the change month to month of the CPI index. This does corresponds to the common use of inflation metrics, measuring year to year, quarter to quarter or month to month global changes in prices.

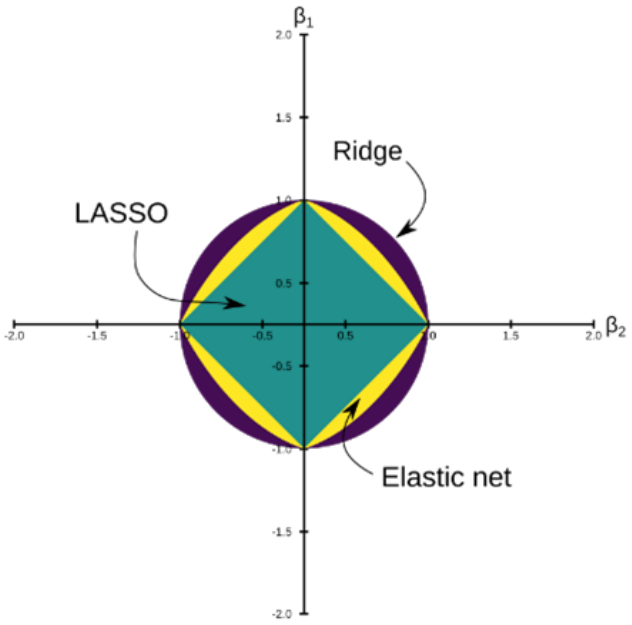
2.2 Description of the models

In this subsection I provide a brief presentation of each model used in this report.

2.2.1 Elastic Net

Elastic Net was first proposed by Zou and Hastie 2005. It is a penalization technique used in linear regression and machine learning to address the limitations of Lasso (L1) and Ridge (L2). Indeed, it combines the penalties of both L1 and L2 norms, allowing for simultaneous feature selection and handling multicollinearity. The L1 part of Elastic Net is able to perform variable selection, while L2 part helps when variable are highly correlated. Elastic Net has a parameter, named alpha, that controls for the intensity of regularization and a mixing parameter that determines the trade-off between L1 and L2 penalties, allowing to fine-tune the model's behavior based on the specific characteristics of a dataset. If this hyperparameter is set to 1, then it will correspond to pure Lasso, and to 0 will represent Ridge. The benefit of Elastic Net lies in its ability to handle datasets with highly correlated features while still performing variable selection. Therefore, it makes it a robust and interpretable model.

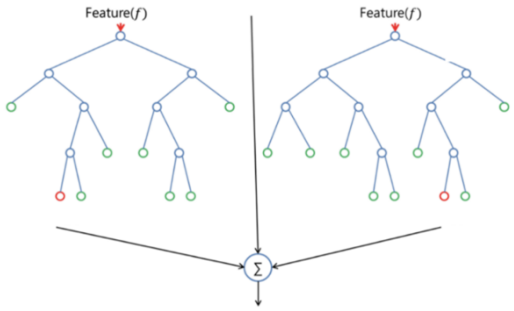
Figure 1: Elastic Net Illustration



2.2.2 Random Forest

Random Forest is a model based on Classificaiton and Regression Trees used both for classification and regression tasks proposed by Breiman 2001. The fundamental idea behind Random Forest is to build a multitude of decision trees on bootstrap samples during the training phase and combine their predictions to improve accuracy and robustness. However model averaging works best when the models are not correlated. To ensure this, at each new tree split, a random subset of the total variables is selected to perform the split on. This ensures that the trees do not perform the same splits on the same variables if only a few of them have a very high predictive power, to ensure diversity among the trees. In addition, because of its tree structure, the Random Forest deals well with mutlicolinearity and the curse of dimensionality. In addition overfitting is controlled by restricting the size and the shape of the forest by ensuring a minimum number of observations in the end leaves or by limiting the size of individual trees. During the prediction phase, the Random Forest aggregates the individual predictions of each tree, using a majority vote for classification or averaging for regression.

Figure 2: Random Forest Illustration

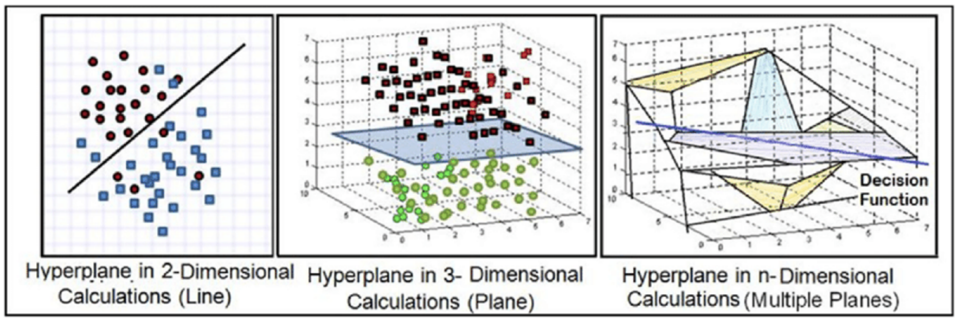


2.2.3 Support Vector Regression

Support Vector Regression (SVR) is a regression model first proposed by Drucker et al. 1996. It is based on Support Vector Machines which were developed earlier by Boser, I. M. Guyon, and V. N. Vapnik 1992, I. Guyon et al. 1991 and Cortes and V. Vapnik 1995. When the goal of SVM was to find the hyperplane that best separates data points in different classes, allowing for some margin of error, the SVR can be thought in 3 dimensions of trying to fit a tube that will contain the most data points possible. In addition, the training data points are mapped into a higher-dimensional space using a kernel function. This mapping can allow for an easier fitting of the model. The tube size, determined by the parameter ϵ, is chosen strategically to encompass as much relevant data as possible while maintaining a compact size to avoid becoming an uninformative average. For this the model relaxes the constraint for some observations falling off the tube, determined by a slackness parameter designated by C. With its structure the SVR is able to model complex non linear relationships between the variables and is also robust to high dimensionality and multicolinearity because of its slackness

parameter. Indeed the model accept to ignore some information in order to better fit to other data points.

Figure 3: SVM Illustration



2.2.4 ARMA

Autoregressive Moving Average (ARMA) is a statistical method used for time series analysis and forecasting. I use it in the study as the reference model to compare the performance of other forecasts to. The model is composed of two distinct part: the autoregressive (AR) component and the moving average (MA) component . In the autoregressive part, the model captures the linear relationship between the current observation and its past values. The moving average component, on the other hand, accounts for the influence of past error terms on the current observation. It can be modeled as a linear regression of the current value against the past error terms that are viewed here as the past economic shocks. The ARMA model is expressed as ARMA(p, q), where 'p' denotes the order of the autoregressive part, and 'q' represents the order of the moving average part. By determining appropriate values for 'p' and 'q' based on the characteristics of the time series, ARMA provides a flexible framework for capturing both short-term and long-term dependencies within the data.
Compared to the other models the ARMA model has two main characteristics. It is solely linear and it is based only on the past values, and the shocks, of the forecasted variable. This also means that I cannot use on the ARMA model PCA or PLS factors. One final note is that the ARMA model assumes in its construction that, if it is well tuned, the combination of the AR and MA terms fully capture the information and leave residuals that are normal with no autocorrelation left inside. This implies that ARMA model has its own prediction intervals that are based on normal residuals.

2.2.5 Other remarks

I use this last remark to mention the models I unsuccessfully tried to implement. I first tried to implement a Long Short Term Memory (LSTM) neural network. This was unsuccessful for three main reasons. First I did not find in the literature clear examples of LSTM model structures that performed well in macroeconomic forecasting that I could copy or take inspiration from. This left us blind and I tried different structure that gave bad performances. The second reason is the limited dataset that I use. I have a quite imbalanced data in the sense that I have only a bit more than twice observations than variables. The limited number of observations, for the large number of variables, gave poor performances with this model. Finally LSTM models can be long to train, which was very time consuming for us, so I preferred to move on and better focus on the rest of the analysis.

The second failed model was a Vector Autoregression (VAR) model. Once again I mainly droppped because of time constraints. I was unsure of which VAR model to use. Indeed this is a multivariate model were multiple series are forecasted and modeled at same time. The question of which series or even factors of the series, computed by PCA or PLS, to select was left open. I tried multiple combinations that did not give much better performances than the simple ARMA model. As this only made the forecasting problem more complex for not much information, I preferred to leave it.

2.3 Empirical strategies and model tuning

This subsection delves into the empirical strategies used, outlining the periods considered for forecasting, the methods implemented for dimension reduction, and the tuning of the models.

2.3.1 Forecasting set-up

I possess monthly data spanning from January 2000 to July 2023; however, I made a deliberate choice to exclude variables beyond December 2019. The decision was made to facilitate a forecasting exercise on the most recent data while excluding the influence of the Covid-19 crisis. The figures in the descriptive statistics subsection illustrates that the abrupt economic shutdown let to significant shifts in variables such as industrial production and unemployment, creating large and brief fluctuations that could distort the forecast performance. These fluctuations do not align with the characteristics of a conventional economic crisis scenario.

The out-of-sample period spans 4 years, from January 2016 to December 2019, amounting to 48 observations. The in-sample period covers January 2000 to December 2015, constituting 16 years of data or 192 observations. Forecast are exclusively conducted with a forecast horizon equivalent to one month (h=1) to provide monthly predictions.

use the same forecasting setup than in Medeiros, Vasconcelos, et al. 2021. The individual variables are modeled following :

$$y_{t+h} = G_h(\mathbf{x}_t) + u_{t+h} \tag{1}$$

Here $h = 1, ..., H$ and $t = 1, ..., T$. The variable of interest at month $t + h$ is given by y_{t+h}. Concerning $\mathbf{x}_t = (x_{1t}, ..., x_{nt})$, it is a vector containing n-covariates at time t. These covariates are the observed data and factors computed on the data. Note that in all of the set-ups in addition the three lagged values of the forecasted variable as additional regressors. $G_h(.)$ represents the mapping between the covariates and the variable of interest, based on the forecasting horizon h.

From this model the forecasting equation is given by :

$$\hat{y}_{t+h|t} = \hat{G}_{h,t-R_h+1:t}(\mathbf{x}_t) \tag{2}$$

$\hat{G}_{h,t-R_h+1:t}$ is the estimated target function based on data from time $t - R_h + 1$ up until t. R_h is the window size. Its size is given by $R_h = T_{full} - h - 1$. T_{full} is the combined length of the in-sample and out-of-sample periods, representing the full used data. The window size also depends on the forecasting horizon h.

A rolling window scheme is used here, meaning that for each forecast the size of the in-sample period remains the same, and the window continuously shifts such that the first in-sample observations in the window are progressively discarded. One motivation for using this scheme would be, following Pesaran, Pick, and Pranovich 2013 that this setup is more robust to structural breaks. As older data is progressively excluded, recent changes in the data generating process are more taken into account. However, this comes at the cost of not using all of the available data to fit the model.

2.3.2 Dimensional Reduction Methods

The curse of dimensionality poses significant challenges in forecasting analysis, particularly when dealing with datasets with a high number of variables. According to Bellman 1961, the curse of dimensionality arises when attempting to accurately estimate a function with an increasing number of variables, requiring an exponential number of data points. In other words, when the dataset becomes larger, I need even more observations in order to have satisfactory forecasts. As the number of features increases, the computational complexity grows as well as the risk of overfitting. High dimensional data makes it challenging to interpret relationships between variables.

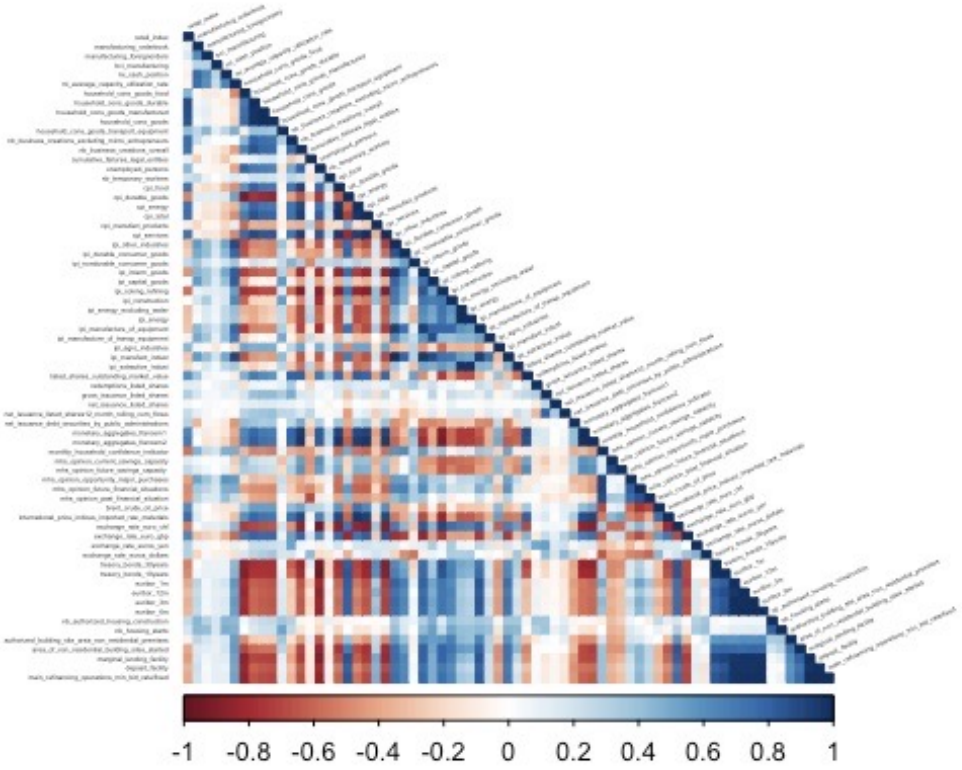
Macroeconomic datasets can have a large number of economic series, 70 in the case, but face limitations due to a relatively smaller number of observations (283 observations). This imbalance makes Ordinary Least Squares (OLS) inefficient for forecasting exercises, necessitating the use of methods capable to perform information reduction. In recent years, these models have been effectively implemented in macroeconomic forecasting demonstrating superior performance compared to traditional models, like see in the litterateur review. However, the adoption of these methods introduces additional complexities. They engage in information reduction through diverse approaches, offer multiple tuning possibilities, and exhibit varying performances contingent on the data's characteristics. These factors significantly impact the forecasting effectiveness of the models, necessitating the establishment of guidelines to ensure their successful and meaningful application.

To address this issue, I used the Principal Component Analysis (PCA). In simple terms, PCA (Principal Component Analysis) is a technique that reduces a large set of variables to a smaller one while retaining most of the original information. Before applying PCA, it's crucial to standardize the data. Standardization ensures that each variable contributes equally to the analysis. This step is necessary because if there are significant differences in the ranges of the initial variables, those with larger ranges might dominate the analysis, potentially leading to biased results. Here is the following formula that one can apply:

$$Z = \frac{X - \bar{X}}{s_X}$$

Where X is the original variable, $\bar{X}$ is the mean of X, and s_X is the standard deviation of X. After standardizing the data, I calculate the covariance matrix to analyze the relationships between input variables. High correlation between variables can indicate redundant information. To illustrate this concept, I use a correlation matrix for visualization. Along the diagonal, the correlation coefficients are precisely equal to 1. A deeper color signifies a stronger correlation, either positive (if it's a darker shade of blue) or negative (if it's a darker shade of red).
From the covariance matrix, the computation of principal components is derived. These components are original variables created through linear combinations of the original variables. The combinations are created in a manner that renders the new variables, known as principal components, uncorrelated. Additionally, the majority of the information present in the initial variables is concentrated in the first components.

Figure 4: Correlation Matrix



Similar to PCA, I also conduct Partial least squares (PLS). It is particularly useful when dealing with high dimensional data, multicollinearity, and situations where the number of features is greater than the number of observations. PLS has some relations to the principal component regression and can be viewed as an extension. As PCR, PLS can be used as a dimension reduction method, which aims to reduce the number of variables used to predict a given outcome variable. PLS first identifies a new set of features that are linear combinations of the original features and then fits a linear model via least squares using these M new features. Unlike PCR, PLS identifies these new features in a supervised way, meaning it uses as well the outcome variable Y. The idea is to find directions that help explain both the response and the predictors. Therefore, PLS aims at finding the fundamental relations between the X and Y matrices to model the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. The components that are selected give the maximal reduction in covariance of the data.

In the forecasting analysis, I used the associated factors selected by PCA and PLS respectively. Referring to James H. Stock and Mark W. Watson 2002b (and J. Stock and M. Watson 2016 for more details behind the theory), who initially used large datasets for forecasting, macroeconomic data often demonstrates a factor structure. This suggests that a linear combination of multiple variables, termed a factor, closely tracks many macroeconomic variables. The implication is that economic series are not explained by a small set of variables but rather

by aggregating information from many variables. To leverage this, forecasting models should incorporate numerous variables while still achieving information reduction without exclusion. In the PCA set up, I used the factors that would represent 90% of the variance of the original dataset. This represents 25 factors. In the PLS set up, because the factors are not computed to solely capture the variance of the dataset, but are also related to the outcome variable, I decided to select the same number of factors compared to the PCA set up. In this set up, the 25 PLS factors explain around 73% of the variance of the outcome variable.

2.3.3 Conformal Prediction

As mentioned in introduction, a single point forecast for inflation policies is not very helpful. A more useful tool would be an interval to have a better picture of the studied phenomenon. However, providing prediction intervals is a complex task. In the case, except for the ARMA model, I do not assume a functional form of the data or a data generating process. Therefore, I do not have any prior knowledge on the model, but more specifically on the residuals of the model. Without knowing the form of the residuals, I cannot say if the point forecast is an accurate picture of future inflation or if it just a point among very wide distribution of possible realisations. To tackle this uncertainty, I use conformal prediction.

Conformal Prediction (CP) was first proposed by Alex Gammerman, Volodya Vovk, and V. Vapnik 2013, and more explored in V. Vovk, A. Gammerman, and Shafer 2010. In regression set ups CP aims at producing an interval which will, based on a tolerance level, have a certain probability to contain the true value. This property can be achieved in small sample, with the only assumptions that the data is exchangeable, meaning that it comes from the same data generating process. For instance, for a 5% level using an XGBoost model, based on the training data, CP will give us a prediction interval that will contain the predicted value with a 95% probability. When dealing with time series data, CP is applied through the EnbPI algorithm. Its goal is to predict, in addition to the model producing the point forecast, the standard deviation of the residuals. With them it is now possible to construct prediction intervals. This is done by repeatedly fitting the model on the bootstrap samples of the training data. This way I construct a distribution of the predictions, and I can retrieve the quantiles associated. With knowledge about the distribution of the prediction I can now asses the uncertainty around the point forecast. In addition the EnbPI algorithm makes this process dynamic by slowly moving the window of the used data for creating the bootstrap samples. This updating ensures that measures of uncertainty is not affected by changes in the data-generating process, which would make it biased.

I implemented this framework with all the machine learning models to produce prediction intervals. I didn't include it for the ARMA model as I used an auto-arma model, for the reasons explained in the description of the ARMA model. Nonetheless this method is limited as the residuals are often not normal, producing miss-specified prediction intervals. This miss specification can be measured by the coverage rate of the prediction intervals compared to the actual value of the outcome. This coverage measures the proportion of the realized values being inside the prediction intervals. If the intervals are based on a 95% rate, having a coverage lower than this value would indicate that the residuals are miss-specified or that the conformal

prediction set-up is not adapted to the kind of data I have. Note that the small number of predicted values also affects negatively the coverage rate.

2.3.4 Diebold Mariano test

To globally compare the performance of the models, I decided to use the Diebold Mariano (DM) test to check if one forecast of reference is significantly better than a battery of other forecasts. This allows us to reject cases where I would try to make sense of marginal improvements in terms of RMSE, which probably are just a coincidence and do not reflect a real improvement in terms of forecasting accuracy. The test was proposed by Diebold and Mariano 1995 and further refined in Harvey, Leybourne, and Newbold 1997. The test has for null hypothesis that on expectation, the difference in terms of forecast error, evaluated with a given loss function, of both models is equal to 0. This means that both forecasts are of equivalent performance. Under the assumption that the loss differentials are covariance stationary, meaning that the mean and variance of the difference in forecast errors doesn't change across time, then the sample average of the loss differential converges asymptotically to a normal distribution. This is a very strong result for a moderately restrictive assumption as this assumption would be violated in case of a structural break causing the two models to suddenly change their forecasting performance. If the data generating process remains the same, this assumption seems probable. Therefore, when performing a DM test, I look at the p-value to see, in the case it is under the significance level, if the two forecasts are of significantly different accuracy. In the case, since I only have 48 forecasts I select a significance level of 20% as the low number of observations to perform the test on could negatively impact its power. With this higher threshold, the test will be less capable to finely discriminate between forecasts of similar accuracy, and will return a more conservative, but surer, result.

2.3.5 Model Tuning

In general terms a "TimeSeriesSplit" cross-validation approach is employed to evaluate the performance of the models in the regression process for being able to obtain the best tunning parameters. The dataset is sequentially split into five folds, ensuring that each testing set chronologically follows its corresponding training set. This is the main difference with K-fold cross validation where the folds can be mixed up in any order. Here the training sample grows along each fold and always maintains the data order such that I can only use validation data more recent than the training data. The model undergoes hyperparameter tuning using GridSearchCV in each iteration of a rolling window. For each window, the training set is used to train the model, and the subsequent testing set assesses its predictive capabilities. I set the testing sample in the cross validation data to 12 observations. This test size is a complex tradeoff. Because of structural breaks I don't want a period too long to test the model on data coming from different inflation regimes. However I also want a sufficient number of observations to have a more general evaluation of the performance. This is why I considered that one or two years were a good choice. I kept a single year because of the limited data. This approach, tailored for time series data, aids in capturing temporal patterns and evaluating the model's ability to generalize across different time points. Even if I have chosen this approach, note that in the forecasting exercise of Goulet Coulombe, Leroux, et al. 2022, in terms of

point forecast accuracy, K-fold cross-validation performs as well as the time series validation on macroeconomic time series.

The primary tuning parameters considered for each model were as follows:

1. Elastic Net:

    - `alpha`: Regularization parameter (tested values: 0.01, 0.025, 0.05, 0.075, 0.1, 1). This parameter controls the strength of the regularization. A higher alpha increases the penalty for model complexity, helping prevent overfitting by shrinking the coefficients.
    - `l1_ratio`: Mixing parameter (tested values: 0, 0.1, 0.25, 0.5, 0.75, 0.9, 1). Represents the balance between L1 (Lasso) and L2 (Ridge) regularization. A value of 1 corresponds to pure Lasso, and 0 corresponds to pure Ridge. The choice of l1_ratio influences the type of regularization applied, impacting feature selection and coefficient sparsity.
    - Feature Selection: Elastic Net inherently performs feature selection by encouraging some coefficients to become exactly zero. Variables with non-zero coefficients from the best Elastic Net model are selected, contributing to a more parsimonious model with a reduced set of relevant features.

2. Random Forest:

    - `n_estimators`: Number of trees in the forest (tested value: 200). Increasing this parameter typically improves model performance up to a certain point, as it allows the ensemble to capture more complex patterns in the data. However, excessively large values may lead to overfitting. Here I select a unique value as too large number of trees dramatically increase the fitting time of the model.
    - `max_depth`: Maximum depth of the trees (tested values: 5, 10, 20). Controlling tree depth helps prevent overfitting. A deeper tree can capture more intricate relationships in the training data, but it may struggle to generalize well to unseen data.
    - `min_samples_split`: Minimum number of samples required to split an internal node (tested values: 4, 6, 8, 10). Increasing this value prevents the model from creating nodes that only fit a small number of samples, promoting robustness against noise.
    - `max_features`: Number of variables selected to perform a split on (tested values: 'sqrt', 'log2'). By reducing this value, I ensure that the different trees in the forest are less correlated such that I benefit from tree averaging to reduce the variance of the predictions.

3. SVM:

    - `C`: Regularization parameter (tested values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1). Controls the trade-off between achieving a smooth decision boundary and

classifying training points correctly. A smaller C encourages a simpler decision boundary, while a larger C allows for more complex boundaries that fit the training data more closely.

- `gamma`: Kernel coefficient for 'Radial Basis Function' (tested values: 1, 0.1, 0.01, 0.001, 10). It defines the influence of a single training example, with low values indicating a broader influence and high values leading to a more localized influence. Prevents overfitting by adjusting the smoothness of the decision boundary. Note that I selected the Radial Basis Function as kernel based on the work of Exterkate et al. 2016 that recomends using this kernel for macroeconomic forecasting.

4. ARMA

- `max_p`: Maximum AR lags of the model (tested values: 8). As I use an autoarma method, I specify the maximum number of possible AR lags to construct the grid on.
- `max_q`: Maximum MA terms of the model (tested values: 8). As I use an autoarma method, I specify the maximum number of possible MA coefficients to construct the grid on.

Finally the results are compared in terms of Root Mean Squared Error (RMSE). Given the forecast $\hat{y}_{t+h}$ I can compute the forecasting error $\hat{e}_{t,m,h} = y_t - \hat{y}_{t,m,h}$. This error is specific to a model m and was made with data up until $t-h$. From this, I can compare each model forecasting at horizon m using the RMSE defined as follows, with $T - T_0$ being the length of the forecast :

$$RMSE_h = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}_{t,h}^2} \tag{3}$$

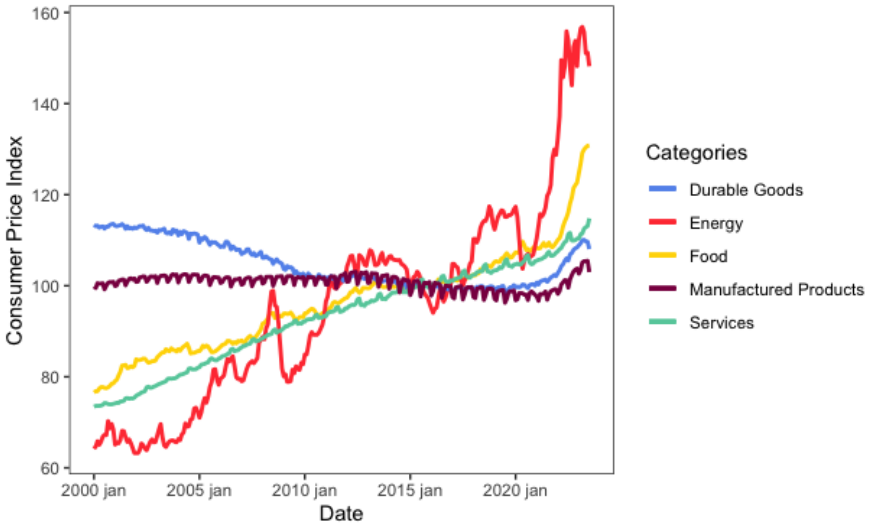
3 Results

This section begins by outlining descriptive statistics for the dataset, followed by a presentation of the results obtained.

3.1 Descriptive statistics

This subsection aims at providing some visual descriptives statitics of the dataset. The graph below represents the consumer price index (CPI) over time.
The Consumer Price Index (CPI) is a statistical measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and energy for instance. It is used to evaluate changes in the cost of living and is a key indicator of inflation or deflation in an economy. The CPI is commonly utilized by economists, policymakers, and researchers to assess trends in consumer prices and to adjust income and investment values for the impact of inflation.

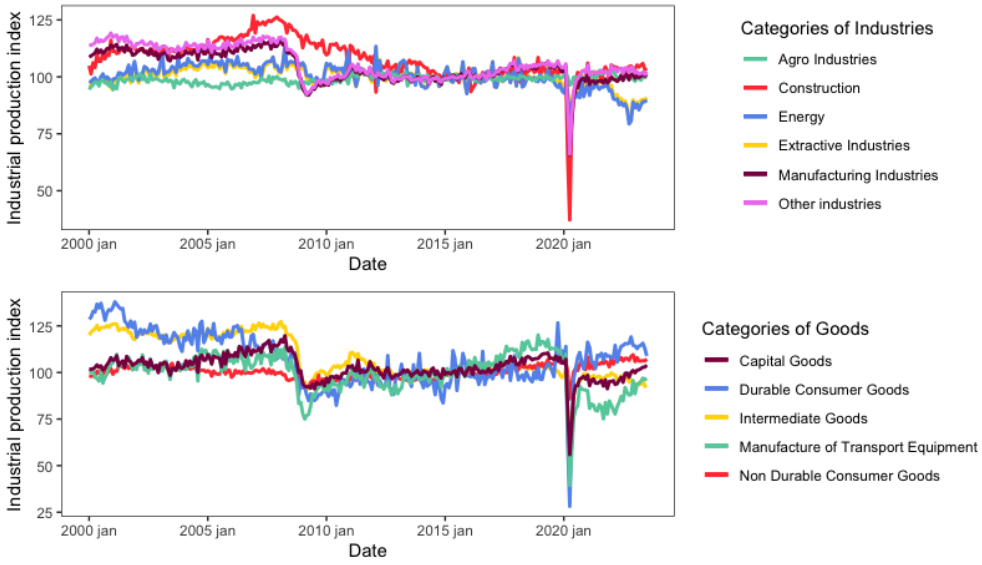
Figure 5: Consumer Price Index over time



The graphical representation reveals noteworthy trends in Consumer Price Index (CPI) components over the observed period. Specifically, the CPI for energy exhibits a consistent upward trajectory, indicating a continual increase in energy-related prices throughout the timeline. Conversely, the CPIs for food and services depict persistent upward trends, illustrating a sustained rise in prices within these categories over the years. Notably, as of 2023, the CPI for energy surpasses others, signifying a notable increase in energy-related costs compared to other CPI components. Examining the CPI for manufactured products, the graph portrays a relatively stable pattern, with marginal fluctuations over the years. This suggests that the overall prices for manufactured goods have remained relatively constant without substantial variations. Interestingly, the CPI for durable goods experienced its peak between 2000 and 2010, followed by a decline. These insights from the graph provide valuable information about the dynamics of price changes in different CPI components over time.

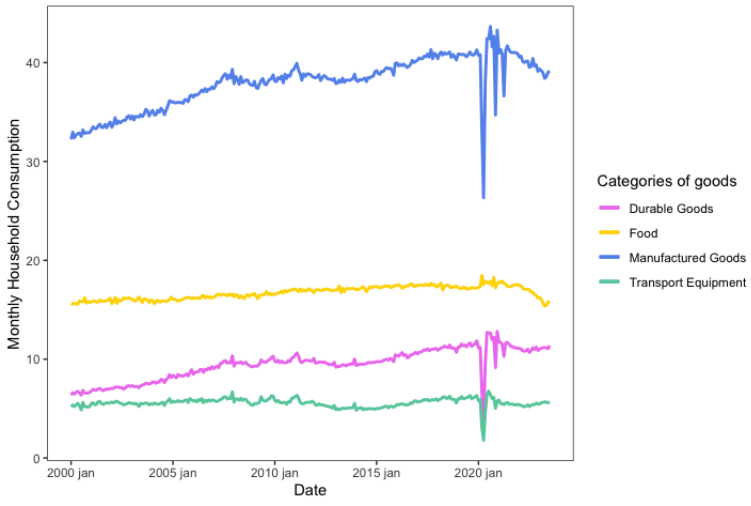
The plots below illustrate the Industrial Production Index (IPI) across various industries and goods over time. The IPI is an economic indicator that measures the output of the industrial sector of an economy. it quantifies the fluctuations in the production volume of a diverse range of goods over time, providing a comprehensive measure of industrial sector performance. Both graphs depict a relatively stable trend in the IPI over time, with a notable exception around January 2020. During this period, a significant downward spike is evident, attributed to the global impact of the Covid-19 pandemic. This phenomenon is observable across all industry categories and goods, highlighting the widespread and profound influence of the pandemic on industrial production.

Figure 6: Industrial Production Index over time



The following figure depicts the household consumption over time across different categories, including durable goods, food, manufactured goods, and transport equipment. In parallel with the observations in the industrial production index, a significant decline is observable in January 2020. This decline is particularly pronounced for manufactured goods, transport equipment, and durable goods. Interestingly, there seems to be no discernible impact on food consumption. Perhaps unsurprisingly, despite lockdown measures during the pandemic, households did not decrease their food consumption, instead there is a slight increase observed.

Figure 7: Household Consumption over time



Finally, below is a table that displays some information concerning the minimum, maximum, mean and variance for some variables that were not ploted above.

Table 1: Summary Statistics

Variables	**Min.**	**Max.**	**Mean**	**Median**
Retail index	62.15	112.22	97.74	98.22
Manufacturing order book	-52.76	41.46	7.03	7.48
Manufacturing foreign orders	-76.22	21.98	4.35	5.16
Manufactured industries cash position	-8.49	23.96	10.40	10.86
Number of business creations overall	18658	93716	46537	47481
Consumer Price Index Total	79.40	117.47	95.77	97.28
Exchange rate euro-chf	0.9561	1.6762	1.3130	1.2310
Exchange rate euro gbp	0.5794	0.9525	0.7828	0.8214
Exchange rate euros yen	91.88	169.02	128.54	129.39
Main refinancing operations	0.00	4.75	1.53	1.00

3.2 Result of predictive analysis

Table 2: Forecasting results

Models	Random Forest	Elastic-Net	SVR	ARMA
RMSE	0.141	0.13	0.16	0.169
Average prediction width	0.316	0.397	0.441	0.664
Coverage	73%	89%	72%	0.83%
Models, on PCA factors	**Random Forest**	**Elastic-Net***	**SVR**	**ARMA**
RMSE	0.132	0.062	0.09	–
Average prediction width	0.345	0.196	0.228	–
Coverage	73%	90%	83%	–
Models, on PLS factors	**Random Forest**	**Elastic-Net**	**SVR**	**ARMA**
RMSE	0.131	0.102	0.106	–
Average prediction width	0.336	0.346	0.307	–
Coverage	71%	89%	92%	–

The visualisation of the forecasts and their prediction intervals against the actual values can be seen in appendix at 8, 9, 10 and 11.

On these forecasts, I selected the Elastic-Net with PCA factors forecasts to be the reference forecast in the DM test, because it had the lowest RMSE and the smallest average prediction

interval. In bold and with a *, together with the best forecast in terms of RMSE, were marked the models that did not reject the null hypothesis of equal forecast accuracy at the 20% significance level. In the case this means that for every pair of tests the Elastic-Net computed on PCA factors had a significant difference in forecast accuracy. This means that the DM test supports the idea that this forecast is significantly better than all the others.

Following the forecasting exercise I can draw a few conclusions. Overall the best performing model is the Elastic-Net, which works best in terms of point forecasts and prediction intervals in both set-ups. The second conclusion is that the PCA set-up improves the forecasting accuracy of the models both in terms of point forecast and of prediction intervals. Finally all the machine learning models perform better than the reference ARMA model. One other very important remark is that the Elastic-Net is actually a Ridge in the PCA and PLS cases. For every forecast point in the rolling window in these two specifications, the value of the mixture parameter selected by cross validation was 0 making the model a Ridge.

Concerning the predictions intervals I can make the following remarks. The best point forecasts also have the smallest average prediction width. However one recurring problem is the coverage of these prediction intervals, that is systematically smaller than the expected 95%. For the ARMA normal this can come from the misspecification of the residuals of the model. The model assumes that the residuals are normal, which tends to not be the case with macroeconomic forecasts. For the other models using conformal prediction, this can come from the settings of the code. I used the MAPIE package on python, and I might not have given the best settings to the function according to the situation. Also the length of the forecasts can play a role too. As I forecast a small sample, 48 observations, the coverage properties of conformal prediction might not be the best with such a number. The same can play a role with the coverage of the ARMA forecast. Overall, with its small average prediction width and the Elastic-Net on PCA factors produce accurate and rather sure individual forecasts because of its relatively high coverage.

Compared to the literature, these results seem coherent. Goulet Coulombe, Marcellino, and Stevanović 2021 and Goulet Coulombe et al. 2021 both point out that forecasting on data transformed by PCA improves the point forecast accuracy of the models. Maehashi and Shintani 2020, Goulet Coulombe, Marcellino, and Stevanović 2021 and J. M. Chen 2014 also find that shrinkage models, especially in periods of stable economic growth, are competitive models that can perform well. As a whole this could support the hypothesis, better explained in J. Stock and M. Watson 2016, that macroeconomic data exhibits a factor structure.

The one surprising result is that it's a linear model, the elastic net, that delivers the best performance. One might imagine that a model like a SVR or Random Forest could benefit from its capability to model complex relationships. If I go back to Goulet Coulombe, Marcellino, and Stevanović 2021 the improvement brought by non linear models is the most seen at forecasting horizons larger than 3 months. For one month ahead forecast they find no difference between linear and non linear models.

3.3 Variable importance

In the analysis, the Random Forest and Elastic-Net have as strength to be able to perform variable selection or at least show metrics of which variables were most used for forecasting. I can study these results to identify which variables were most useful for forecasting inflation. The selection performed by the two models differs slightly. In the Random Forest, the metric measures how often a variable was selected to perform a split. In the case Ionly report the variables that were selected more than 5% of the times. For the Elastic-Net, Ireport the variables selected by the $\mathcal{L}_1$ norm of the LASSO. As the Elastic-Net is a mixture of a LASSO and a Ridge, it performs variable selection. Here Ireport the 10 variables that were most often selected by this norm.

Table 3: Variable importance following Random Forest and Elastic-Net

Model	Random Forest	Elastic-Net
Metric	10 most selected variables for tree splits	10 variables most selected by the L1 norm
	cpi energy	manufacturing foreign orders
	crude brent price	nb business creation overall
	ipi energy except water	cumulative failure legal entities
	ipi energy	unemployment overall
	ipi extractive industries	temporary employment
	ipi other industries	cpi energy
	swiss franc exchange rate	ipi coking refining
	household consumption transport equipment	ipi construction
	ipi intermediary goods	ipi extractive industries
	dollar exchange rate	redemptions listed shares

Here the two models paint two different pictures. The Random Forest seems to select variables more related to the supply side, foreign and exogenous explanation of inflation while the Elastic-Net seems more of mixture of supply and demand causes of inflation, more related to domestic factors. For instance in the Random Forest Ifind multiple variables related to energy prices, particularly oil, exchange rates and industrial production. These variables can explain inflation through imported inflation, either because of energy or exchange rates, or it can explain inflation with industrial bottlenecks using the indexes of industrial production. Concerning the Elastic-Net inflation can be explained with aggregate demand, through foreign demand and employment levels and domestic supply related to business creations and failures.

This result shows the complexity of variable selection for macroeconomic forecasting. As economic theory can provide multiple explanation for a same phenomenon, it is difficult based on prior knowledge to justify the selection of a set of variable. Here different models use quite different sets of variables to perform the same forecast. Also the set of variable could even change across time based on training and testing periods. For instance for predicting inflation during Covid, industrial bottlenecks seems the most important phenomenon to target, while during the 70s and 80s energy crises, oil and gas prices seemed the most important.

Another metric Ican study is the composition of the factors resulting from PCA analysis. The PCA factors assign weights to the variables having the most importance in their linear mixture. This differs a bit from variable selection. Here I do not look for the variables that were the most used to forecast the particular variable, but rather which variables, and particularly groups of combined variables, are the most important to explain the variability in the data set.

Table 4: 3 Most important PCA and factors and their 5 biggest weights

Factor 1	Factor 2	Factor 3
household cons durable goods	Manufacturing foreign orders	IPI energy
IPI other industries	Manufacturing capacity utilization rate	households future savings capacity
IPI manufacturing	Manufacturing cash position	EURIBOR 6 months rate
IPI intermediary goods	Manufacturing order book	IPI extracting industries
household cons manufactured goods	nb housing starts	EURIBOR 12 months rate

In this table I show the 3 factors explaining the most variance in the dataset. Together, they represent roughly 35% of the total variance. For each of these factors I show the 5 variables having the largest loading weights, meaning that they are the most important in the linear combination. I see that each factor can have some interpretability as they tend to group similar variables together. First factor is related to the production and consumption of manufactured goods, the second factor is related to economic situation of manufacturing companies and the last factor is broadly related to savings. This way, when I feed factors into the models, I can look at their composition to retrieve some interpretation over the explanatory variables.

Nonetheless I will not go over all the factors selected by the forecasting models as this would be too long and not necessarily helpful. Indeed, it is possible to give an interpretation to the first most important factors in terms of variance, but the deeper I go and the more the linear combination of the variables become complex to catch the remaining variance in the data, and I lose the clear interpretability. I can nonetheless make some observations. In the forecasting example I kept the factors that would explain 90% of the variance of the original data. This in turn gives us 30 factors out of the 69 original variables. In this forecasting exercise the 5 most selected varibles for the splits of the Random Forests on the factor data were factor 3, 1, 19, 13 and 5. Note that the lagged values of the CPI is not even in the top 10. The fact that the best performing forecast, even if capable too, hasn't been made on selected variable can be surprising. I can put it in perspective with Ulgazi and Vertier 2022 where the short term forecasts of the model are made on all the variables with no structure behind it.

I see similar results with the analysis of PLS factors. First of all, I have to note that PLS factors are computed with respects to the total CPI time series, therefore I expect that the composition of its factors are more related to explanations of inflation. I can observe this in the following table.

Table 5: 3 Most important PLS and factors and their 5 biggest weights

Factor 1	Factor 2	Factor 3
IPI extractive industries	IPI investment goods	IPI energy except water
IPI energy except water	manufacturing order book	household opinion future savings
IPI energy	Brent price	CPI energy
Brent price	Unemployment rate	Retail index
Households consumption of transport goods	CPI energy	Price index of imported raw goods

Note that these 3 factors explain 25% of the variance of the total CPI time series. Concerning their interpretation I see that variables related to energy are very important and are represented for the three factors in the largest loading weights. Next I mostly see variables related to aggregate demand (consumption of transport goods, Unemployment, retail index). These represent both exogenous and endogenous explanations of inflation.

Concerning the factors selected by the models, they were related to the order of the factors. The models selected most the first factors, which are the ones explaining the most variance in the outcome. Also once again the Elastic-Net following cross validation tented to be a Ridge or have values of alpha close to the ridge case.

4 Conclusion

In this paper, I conducted an analysis of macroeconomic forecasting, focusing on the CPI using a French database comprising 69 macroeconomic series. Leveraging advanced techniques such as PCA and PLS for dimension reduction, I explored various machine learning methodologies, including random forest, support vector machine, elastic net and the traditional ARMA time series model. The approach involved forecasting with variables selected through PLS and PCA, as well as incorporating the PCA factors directly. Rigorous model tuning was performed using grid search and cross validation for time series.

One surprising result that emerges through this analysis is the good performance of Elastic Net, a linear model, when applied to PCA Factors, especially with the coefficient set to ridge. This model demonstrated higher forecasting performance compared to all other models. Moreover, all of the machine learning models outperform the predictive performance of the ARMA model. These findings lend support to the hypothesis, described in J. Stock and M. Watson 2016, that macroeconomic data inherently exhibits a factor structure. This alignment with existing literature enhances the credibility of the results and underscores the robustness of the observed patterns. These results were supported by the result of the Diebold-Mariano test and the study of the prediction intervals.

One last area of study concerned the importance of variables and the construction of PCA and PLS factors. I saw that the set of variables selected by each methodology varied greatly, nonetheless these methods bring additional interpretability to the forecasts. The highlighted variables could be put in perspective of different theories of inflation. For instance the most relevant variables according to the Random Forest were variables related to energy and price

of imports, variables that side with the exogenous explanations of inflation.

Further expansions of the study could include nowcasting and the inclusions of novel series of macroeconomic data. Indeed, in the exercise I was able to produce a competitive short term inflation forecast using the large macroeconomic database. The forecasting exercise could be both expanded to slightly longer term forecasts, like 3 to 6 months, and more frequent forecasts. Nowcasting tries to answer the second question by forecasting economic indicators produced by statistical agencies before their publication (monthly or quarterly). For instance the Cleveland Fed produces a daily inflation nowcast [3] based on high frequency financial and retail data. Nowcasts, because they use high frequency data often use novel sources of data. For instance Macias, Stelmasiak, and Szafranek 2023 uses web scrapped prices to give early estimates of food and drinks inflation or Menzie, Baptiste, and Sebastian 2023 uses real time marine traffic or short term production of steel and semi conductors to nowcast global trade.

[3] link to the nowcasts

5 Appendix

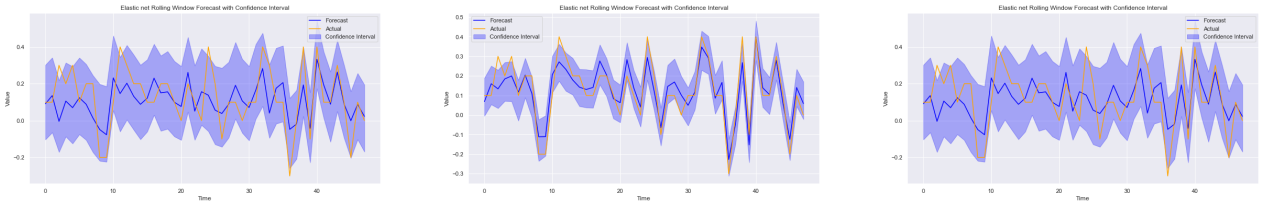
Table 6: Variable Categories and Name, part 1

Category	Variable Name
Lagged values of the predictor	**1st lagged value, 2nd lagged value, 3rd lagged value**
Time-related	date
Economic Indicators	retail_index, manufacturing_orderbook, manufacturing_foreignorders, bci_manufacturing, mi_cash_position, mi_average_capacity_utilization_rate
Household Consumption	household_cons_goods_food, household_cons_goods_durable, household_cons_goods_manufactured, household_cons_goods, household_cons_goods_transport_equipment
Business and Employment	nb_business_creations_excluding_micro_entrepreneurs, nb_business_creations_overall, cumulative_failures_legal_entities, unemployed_persons, nb_temporary_workers
Consumer Price Index (CPI)	cpi_food, cpi_durable_goods, cpi_energy, cpi_total, cpi_manufact_products, cpi_services
Industrial Production Index (IPI)	ipi_other_industries, ipi_durable_consumer_goods, ipi_nondurable_consumer_goods, ipi_interm_goods, ipi_capital_goods, ipi_coking_refining, ipi_construction, ipi_energy_excluding_water, ipi_energy, ipi_manufacture_of_equipment, ipi_manufacture_of_transp_equipment, ipi_agro_industries, ipi_manufact_indust, ipi_extractive_indust
Financial Markets	listed_shares_outstanding_market_value, redemptions_listed_shares, gross_issuance_listed_shares, net_issuance_listed_shares, net_issuance_debt_securities_by_public_administrations
Monetary Aggregates	monetary_aggregates_francem1, monetary_aggregates_francem2

Table 7: Variable Categories and Names, part 2

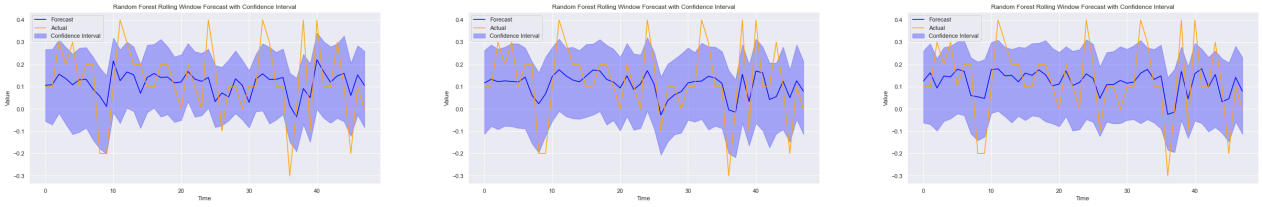
Category	Variable Name
Consumer Confidence	monthly_household_confidence_indicator, mhs_opinion_current_savings_capacity, mhs_opinion_future_savings_capacity, mhs_opinion_opportunity_major_purchases, mhs_opinion_future_financial_situations, mhs_opinion_past_financial_situation
Commodity Prices and Exchange Rates	brent_crude_oil_price, international_price_indices_imported_raw_materials, exchange_rate_euro_chf, exchange_rate_euros_yen, exchange_rate_euros_dollars
Interest Rates and Bonds	tresory_bonds_30years, tresory_bonds_10years, euribor_1m, euribor_3m, euribor_6m, main_refinancing_operations_min_bid_rate_fixed
Construction and Housing	nb_authorized_housing_construction, nb_housing_starts, authorized_building_site_area_non_residential_premises, area_of_non_residential_building_sites_started

Figure 8: Forecasting performance of Elastic Net



(a) Elastic Net (b) Elastic Net with PCA (c) Elastic Net with PLS

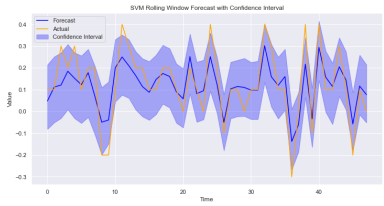
Figure 9: Forecasting performance of Random Forest



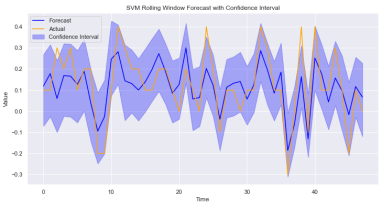
(a) Random Forest (b) Random Forest with PCA (c) Random Forest with PLS

Figure 10: Forecasting performance of Support Vector Regression

(a) Support Vector Regression

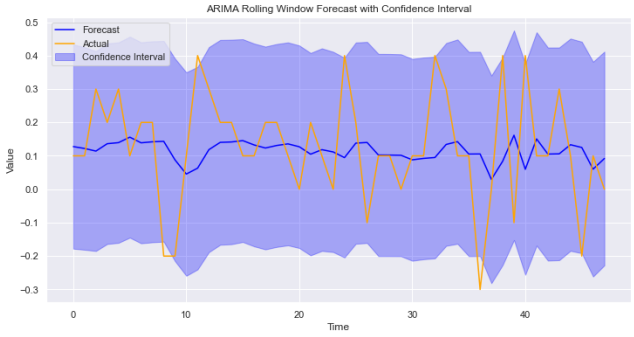


(b) Support Vector Regression with PCA



(c) Support Vector Regression with PLS

Figure 11: Forecasting performance of ARMA



References

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 144–152. ISBN: 089791497X. DOI: 10.1145/130385.130401. URL: https://doi.org/10.1145/130385.130401.

Breiman, Leo (Oct. 2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010950718922.

Chen, James Ming (2014). "Measuring market risk under the Basel accords: VaR, stressed VaR, and expected shortfall". In: *Stressed VaR, and Expected Shortfall (March 19, 2014)* 8, pp. 184–201.

Chen, Jeffrey C. et al. (Jan. 2019). "Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators". In: *Big Data for Twenty-First-Century Economic Statistics*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 373–402. URL: https://ideas.repec.org/h/nbr/nberch/14268.html.

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). "Support-Vector Networks". In: *Mach. Learn.* 20.3, pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. URL: https://doi.org/10.1023/A:1022627411411.

Dickey, D. and Wayne Fuller (June 1979). "Distribution of the Estimators for Autoregressive Time Series With a Unit Root". In: *JASA. Journal of the American Statistical Association* 74. DOI: 10.2307/2286348.

Diebold, Francis and Roberto Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business Economic Statistics* 13.3, pp. 253–63. URL: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63.

Drucker, Harris et al. (1996). "Support Vector Regression Machines". In: *Advances in Neural Information Processing Systems*. Ed. by M.C. Mozer, M. Jordan, and T. Petsche. Vol. 9. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

Exterkate, Peter et al. (2016). "Nonlinear forecasting with many predictors using kernel ridge regression". In: *International Journal of Forecasting* 32.3, pp. 736–753. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2015.11.017. URL: https://www.sciencedirect.com/science/article/pii/S0169207016000182.

Fortin-Gagnon, Olivier et al. (Nov. 2022). "A large Canadian database for macroeconomic analysis". In: *Canadian Journal of Economics/Revue canadienne d'économique* 55.4, pp. 1799–1833. DOI: 10.1111/caje.12618. URL: https://ideas.repec.org/a/wly/canjec/v55y2022i4p1799-1833.html.

Gammerman, Alex, Volodya Vovk, and Vladimir Vapnik (2013). "Learning by Transduction". In: arXiv: 1301.7375 [cs.LG].

Goulet Coulombe, Philippe, Maxime Leroux, et al. (2022). "How is machine learning useful for macroeconomic forecasting?" In: *Journal of Applied Econometrics* 37.5, pp. 920–964. DOI: https://doi.org/10.1002/jae.2910. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.2910. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2910.

Goulet Coulombe, Philippe, Massimiliano Marcellino, and Dalibor Stevanović (2021). "CAN MACHINE LEARNING CATCH THE COVID-19 RECESSION?" In: *National Institute Economic Review* 256, pp. 71–109. DOI: 10.1017/nie.2021.10.

Goulet Coulombe, Philippe et al. (2021). "Macroeconomic data transformations matter". In: *International Journal of Forecasting* 37.4, pp. 1338–1354. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2021.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0169207021000777.

Guyon, I. et al. (1991). "Structural Risk Minimization for Character Recognition". In: *Advances in Neural Information Processing Systems*. Ed. by J. Moody, S. Hanson, and R.P. Lippmann. Vol. 4. Morgan-Kaufmann. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/10a7cdd970fe135cf4f7bb55c0e3b59f-Paper.pdf.

Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the equality of prediction mean squared errors". In: *International Journal of Forecasting* 13.2, pp. 281–291. ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(96)00719-4. URL: https://www.sciencedirect.com/science/article/pii/S0169207096007194.

Jr., Robert E. Lucas (Mar. 2003). "Macroeconomic Priorities". In: *American Economic Review* 93.1, pp. 1–14. URL: https://ideas.repec.org/a/aea/aecrev/v93y2003i1p1-14.html.

Kim, Hyun Hak and Norman R. Swanson (2018). "Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods". In: *International Journal of Forecasting* 34.2, pp. 339–354. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2016.02.012. URL: https://www.sciencedirect.com/science/article/pii/S0169207016300668.

Li, Jiahan and Weiye Chen (Dec. 2014). "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models". In: *International Journal of Forecasting* 30, pp. 996–1015. DOI: 10.1016/j.ijforecast.2014.03.016.

Macias, Paweł, Damian Stelmasiak, and Karol Szafranek (2023). "Nowcasting food inflation with a massive amount of online prices". In: *International Journal of Forecasting* 39.2, pp. 809–826. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2022.02.007. URL: https://www.sciencedirect.com/science/article/pii/S016920702200036X.

Maehashi, Kohei and Mototsugu Shintani (2020). "Macroeconomic forecasting using factor models and machine learning: an application to Japan". In: *Journal of the Japanese and International Economies* 58, p. 101104. ISSN: 0889-1583. DOI: https://doi.org/10.1016/j.jjie.2020.101104. URL: https://www.sciencedirect.com/science/article/pii/S0889158320300411.

McCracken, Michael W. and Serena Ng (2016). "FRED-MD: A Monthly Database for Macroeconomic Research". In: *Journal of Business & Economic Statistics* 34.4, pp. 574–589. DOI: 10.1080/07350015.2015.1086655. eprint: https://doi.org/10.1080/07350015.2015.1086655. URL: https://doi.org/10.1080/07350015.2015.1086655.

Medeiros, Marcelo C. and Eduardo F. Mendes (2016). "1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors". In: *Journal of Econometrics* 191.1, pp. 255–271. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2015.10.011. URL: https://www.sciencedirect.com/science/article/pii/S0304407615002638.

Medeiros, Marcelo C., Gabriel F. R. Vasconcelos, et al. (2021). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods". In: *Journal of Business & Economic Statistics* 39.1, pp. 98–119. DOI: 10.1080/07350015.2019.1637745. URL: https://doi.org/10.1080/07350015.2019.1637745.

Menzie, Chinn, Meunier Baptiste, and Stumpner Sebastian (2023). "Nowcasting World Trade with Machine Learning: a Three-Step Approach". In: 917. URL: https://ideas.repec.org/p/bfr/banfra/917.html.

Milunovich, George (2020). "Forecasting Australia's real house price index: A comparison of time series and machine learning methods". In: *Journal of Forecasting* 39.7, pp. 1098–1118. DOI: https://doi.org/10.1002/for.2678. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2678. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2678.

Pesaran, M. Hashem, Andreas Pick, and Mikhail Pranovich (2013). "Optimal forecasts in the presence of structural breaks". In: *Journal of Econometrics* 177.2. Dynamic Econometric Modeling and Forecasting, pp. 134–152. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2013.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0304407613000687.

Smeekes, Stephan and Etienne Wijler (2018). "Macroeconomic forecasting using penalized regression methods". In: *International Journal of Forecasting* 34.3, pp. 408–430. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2018.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207018300074.

Stock, J.H. and M.W. Watson (2016). "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics". In: *Handbook of Macroeconomics*. Ed. by J. B. Taylor and Harald Uhlig. Vol. 2. Handbook of Macroeconomics. Elsevier. Chap. 0, pp. 415–525. DOI: 10.1016/bs.hesmac.2016.04. URL: https://ideas.repec.org/h/eee/macchp/v2-415.html.

Stock, James H and Mark W Watson (2002a). "Macroeconomic Forecasting Using Diffusion Indexes". In: *Journal of Business & Economic Statistics* 20.2, pp. 147–162. DOI: 10.1198/073500102317351921. eprint: https://doi.org/10.1198/073500102317351921. URL: https://doi.org/10.1198/073500102317351921.

— (2002b). "Forecasting Using Principal Components from a Large Number of Predictors". In: *Journal of the American Statistical Association* 97.460, pp. 1167–1179. ISSN: 01621459. URL: http://www.jstor.org/stable/3085839 (visited on 07/04/2023).

— (2006). "Chapter 10 Forecasting with Many Predictors". In: ed. by G. Elliott, C.W.J. Granger, and A. Timmermann. Vol. 1. Handbook of Economic Forecasting. Elsevier, pp. 515–554. DOI: https://doi.org/10.1016/S1574-0706(05)01010-4. URL: https://www.sciencedirect.com/science/article/pii/S1574070605010104.

Ulgazi, Youssef and Paul Vertier (2022). "Forecasting Inflation in France: an Update of MAPI". In: 869. URL: https://ideas.repec.org/p/bfr/banfra/869.html.

Vovk, V., A. Gammerman, and G. Shafer (2010). *Algorithmic Learning in a Random World*. Springer US. ISBN: 9781441934710. URL: https://books.google.fr/books?id=HPj5kQAACAAJ.

Zou, Hui and Trevor Hastie (2005). "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2, pp. 301–320. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/3647580 (visited on 03/27/2023).