Bulletin
de la Banque de France
248/5 - SEPTEMBER-OCTOBER 2023

BANQUE DE FRANCE
EUROSYSTÈME

Economic research

Nowcasting world trade in real time with machine learning

A key problem in economic assessment is that many time series arrive with long lags, posing a policy challenge. We address it for international trade in volumes by building a monthly "nowcast" (contemporaneous forecast). Using a dataset of 600 variables, our paper uses an innovative machine learning algorithm, the macroeconomic random forest – found to perform better than other linear and non-linear techniques. We employ a three-step approach composed of (i) variable pre-selection, (ii) factor extraction and (iii) machine learning regression. This approach delivers a substantially more accurate prediction compared to a Stock and Watson (2002) method based on factor extraction and OLS, with accuracy gains in between 15-30%. Compared to an autoregressive model, accuracy gains are around 30-40%. We illustrate the performance of the model during the Covid-19 pandemic.

Menzie Chinn
University of Wisconsin
Baptiste Meunier
European Central Bank
Sebastian Stumpner
Banque de France

JEL codes
C53, C55,
E37

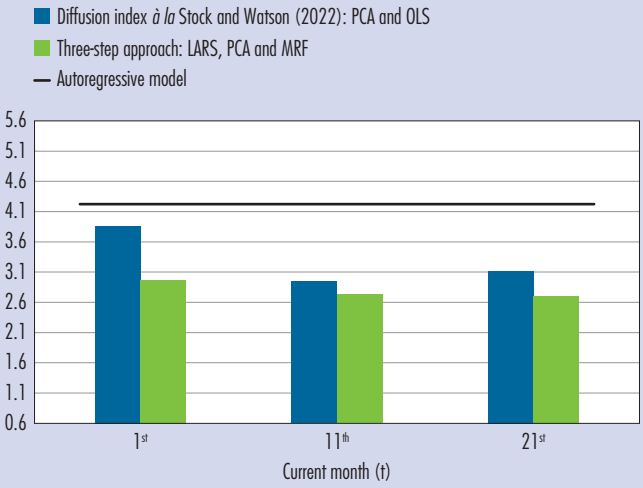
600
number of predictive variables in the dataset

26%
average accuracy gains of our method relative to a linear approach

40%
average accuracy gains of our method relative to a naïve autoregressive model

Evolution of predictive accuracy (out-of-sample RMSE)

(x-axis: forecasting day, either 1st, 11th or 21st day of the month; y-axis: percentage change, year-on-year)



Source: Authors.
Interpretation: A lower root mean squared error (RMSE) indicates a higher accuracy.
Note: The out-of-sample period is Jan. 2012 – April 2022.
LARS: least-angle regression; MRF: macroeconomic random forest; OLS: ordinary least squares; PCA: principal components analysis.

1 Official trade data are published with delay

Real-time economic analysis is often complicated by the fact that economic time series are published with significant lags. This is also the case for international trade: even though some countries publish data on trade in **values** quickly, trade in **volumes** is less timely. The Dutch *Centraal Plan Bureau* (CPB) issues estimates widely used among economists, but which are published roughly eight weeks after month end – meaning March data is available around 25 May.[1] This poses a challenge from a policy perspective, as decisions should rely on timely information.

The delay in data availability is a particularly important problem in a fast-changing economic environment. Recent years have witnessed several rapidly evolving crises, such as the 2020 Covid pandemic or Russia's invasion of Ukraine since 2022. The goal of this project is to develop a tool that allows to accurately predict the evolution of world trade in volumes with no or very short delay, even during large crises episodes.

While official data are published with delay, numerous indicators are available in the meantime. The purpose of our recent paper (Chinn et al., 2023) is to exploit such information to provide advance estimates of trade in **volumes**. Given publication delays, the purpose is not only to predict trade for the current month t ("nowcasting") but also in previous months ("back-casting" at months $t-2$ and $t-1$ for which CPB data have not yet been released). We also "forecast" at $t+1$ to assess the informative content of our method about future developments.

We identify 600 variables that provide timely information. To build our dataset, we screen through the literature on nowcasting trade, notably Keck et al. (2010), Guichard and Rusticelli (2011), Jakaitiene and Dees (2012), Bahroumi et al. (2016), Martinez-Martin and Rusticelli (2021), Charles and Darné (2022). This provides variables covering different aspects of the trade outlook (e.g. customs data, shipping costs, freight traffic) and more broadly the macroeconomic outlook, both industrial activity (e.g. steel production) and households' consumption (e.g. retail sales). Finally, commodity prices and financial indicators are included.

2 Our methodology builds around machine learning

A key novelty of our approach consists in using machine learning algorithms. Testing across different classes of algorithms, the best performing technique is found to be the "macroeconomic random forest" (MRF) of Goulet Coulombe (2020) described in Annex 1.

A second contribution is to propose a three-step framework using first variable pre-selection, then factor extraction, and finally a machine learning regression, that we describe in Annex 1. We compare different methods for each of the steps: the best-performing combination includes "least-angle regression" (LARS) for variable pre-selection (step 1), principal component analysis (PCA) for factor extraction (step 2), and the macroeconomic random forest (MRF) for prediction (step 3).[2] LARS is similar to stepwise regression when dealing with a large set of potential regressors to include variables step-by-step, but the method ensures that regression coefficients are similar in absolute value when the variables have the same correlation with the residuals (see Annex 1). The overall approach works sequentially: (**step 1**) LARS selects the 60 most informative predictors out of our dataset of 600 variables;[3] (**step 2**) selected variables are summarized in a few factors using PCA;[4] and finally (**step 3**) factors are used as explanatory variables in the regression of world trade, using the MRF.[5] While pre-selection and factor extraction have been already used in the literature (e.g. Jardet and Meunier, 2022), our contribution is their combined use in an integrated framework for machine learning. The three-step approach

1 Data is published on a timelier manner for a few advanced economies (for example, around one month for the US and around one month and half for France) but which represent a limited fraction of international trade.

2 The alternative pre-selection methods we explore are "sure independence screening", pre-selection based on t-stat, and Iterated Bayesian Moving Averaging. Alternative factor extraction methods are the 2-step estimator, the quasi-maximum likelihood estimator, and dynamic PCA. Alternative regression techniques are in Annex 3.

3 The forecaster needs to choose the number of variables to be included in the model. In our case we set this number at 60 based on empirical accuracy tests with different numbers of variables.

4 The number of factors to enter the regression is determined through the Bai and Ng (2002) information criteria as is common in the literature.

5 In the setup of the MRF, both the linear part and the random forest part are composed of the factors (see details on the macroeconomic random forest in Annex 1). In that sense, the MRF can be viewed as a generalization of our baseline OLS specification, but where coefficients of the regression are time-varying and follow a random forest algorithm. In addition, the hyper-parameters of the MRF (e.g. number of trees to grow) are set based on cross-validation.

can be viewed as an extension of the widely used "diffusion index" of Stock and Watson (2002) who combine PCA with an OLS regression. We extend it with pre-selection and machine learning.

Using these techniques, we produce in Chart 1 out-of-sample predictions of world trade, from January 2012 to April 2022. We use a real-time set-up by, at each point in time, using only the data that would have been available to the forecaster at that time. It also implies that we re-run variable pre-selection, factor extraction, and regression at each point in time. We manage the real-time data flow (meaning the asynchronous availability of the data at different dates) by using the vertical re-alignment of Altissimo et al. (2006) (see Annex 2).

3 Our three-step approach achieves significant accuracy gains

We measure performance (accuracy) by how well out-of-sample predictions fit actual data. We rely on the root mean squared error (RMSE) measuring the deviation between actual data for the year-on-year growth rate of CPB world trade (y_t) and the model prediction ($\hat{y}_t$). For predictions on an interval going from 1 to T:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}$$

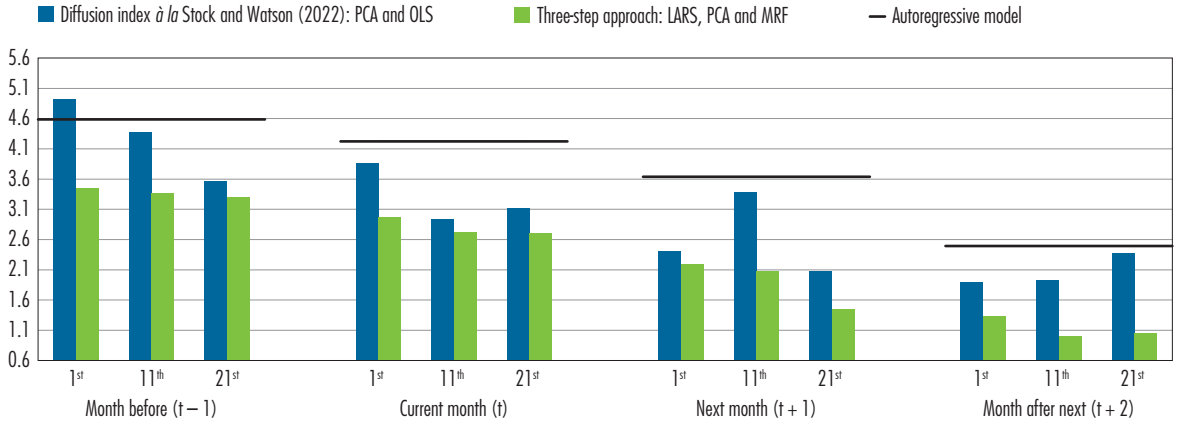
To evaluate the accuracy, we first determine the set of variables that would have been available to a forecaster in real-time on the 1st, 11th, and 21st of each month. We then run the three-step approach on these distinct datasets.

The resulting RMSEs are displayed in Chart 1 for the three-step approach (green), the Stock and Watson (2002) approach (blue) and the autoregressive model (black line). Bars get smaller from left to right, meaning that we (naturally) make less mistakes as more information become available. Indeed, to make predictions for a month t, the forecaster will have access to more information during the next month ($t + 1$, on the right of the chart) than in the month before ($t - 1$, on the left of the chart). We even reach a RMSE below 1% when back-casting. Focusing on the green bars, the RMSE declines by about 40% as we move from predicting on the 11th of month $t - 1$ to predicting on the 11th of month $t + 1$. It then falls by another 50% if we predict on the 11th of month $t + 2$.

The three-step approach consistently outperforms the two benchmarks. While the accuracy gain varies by prediction date and horizon, on average the three-step approach delivers a 26% lower RMSE than a model *à la* Stock and Watson (2002) and a 40% lower RMSE than an autoregressive model. We also show that the three-step approach using the macroeconomic random forest outperforms other three-step approaches based on other linear and non-linear regression techniques (see Annex 3).

C1 Evolution of predictive accuracy (out-of-sample RMSE)

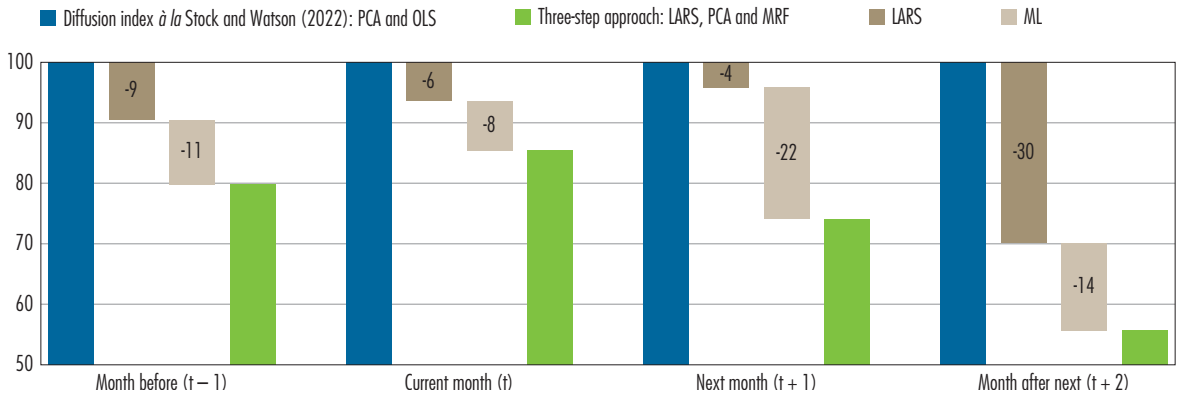
(x-axis: forecasting day, either 1st, 11th or 21st day of the month; y-axis: percentage change, year-on-year)



Source: Authors.
Interpretation: A lower root mean squared error (RMSE) indicates a higher accuracy.
Note: The Out-of-sample period is Jan. 2012 – April 2022. Stock and Watson (2002) is based on PCA and OLS. The three-step approach uses LARS, PCA and macroeconomic random forest (MRF). The performance of the autoregressive (AR) model is flat across forecasting days because the AR term does not depend on the day of the forecast. See glossary for a definition of abbreviations.

C2 Decomposition of accuracy gains (100 = PCA and OLS)

(percentage change, year-on-year)

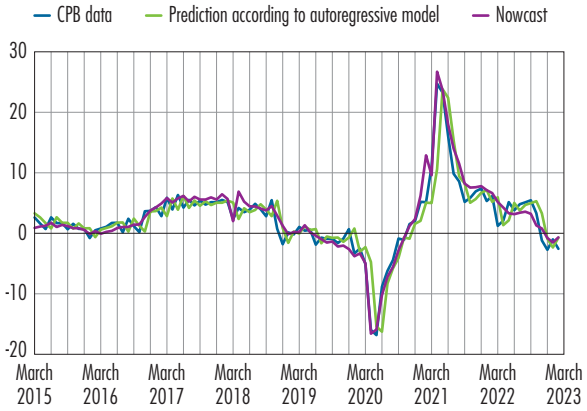


Source: Authors.
Note: Stock and Watson (2002) based on "PCA and OLS"; "3-step": final accuracy with three-step approach using LARS, PCA, and macroeconomic random forest (MRF); "LARS": pre-selection with least-angle regression; "ML": machine learning with macroeconomic random forest. Results are relative to PCA-OLS normalized to 100 for each month. Results are average gains over datasets at 1st, 11th, and 21st days of the month.

Accuracy gains are driven both by variable pre-selection and machine learning. Chart 2 decomposes the accuracy gains compared with the two-step PCA-OLS approach of Stock and Watson (2002) into gains coming from preselection and gains coming from MRF. Generally, both contribute substantially to accuracy gains. While contributions also depend on the horizon, gains from machine learning are more stable and stand between 10 and 20%.

C3 Real-time out-of-sample predictions of world trade growth during Covid-19

(percentage change, year-on-year)



Source: Authors.
Note: "CPB": *Centraal Plan Bureau*. The three-step nowcast uses LARS, PCA, and macroeconomic random forest. See glossary for a definition of abbreviations.

Finally, Chart 3 compares the evolution of the year-on-year percentage change of world trade with the prediction from our three-step nowcast. The real time prediction is based on the data extracted on the 21st of the month, for a back-casting by two months (prediction of month *t* at month *t* + 2). The chart shows that our preferred nowcast is consistently very close to the data and predicts trade growth well in times of high fluctuations when nowcasting exercises are particularly valuable, as shown by the very close predictions during Covid-19. Predictions have also remained highly accurate in the face of more recent shocks such as Russia's aggression of Ukraine in early 2022.

⁂

In the end, we obtain a forecasting model of world trade in **volumes** based on an innovative machine learning method, the macroeconomic random forest. Doing so, we have used a three-step approach featuring pre-selection and factor extraction, before machine learning. More broadly, this approach can be viewed as a practical guide for forecasters willing to use machine learning. The approach is indeed highly flexible and can be applied seamlessly to other target variables.

References

Altissimo (F.), Cristadoro (R.), Forni (M.), Lippi (M.) and Veronese (G.) (2006)
"New Eurocoin: tracking economic growth in real time", *CEPR Discussion Papers*, No. 5633.

Bai (J.) and Ng (S.) (2002)
"Determining the number of factors in approximate factor models", *Econometrica*, Vol. 70, No. 1, pp. 191-221.

Barhoumi (K.), Darné (O.) and Ferrara (L.) (2016)
"A world trade leading index (WTLI)", *Economic Letters*, Vol. 146, pp. 111-115.

Charles (A.) and Darné (O.) (2022)
"Backcasting world trade growth using data reduction methods", *The World Economy*, Vol. 45, No. 10, pp. 3169-3191.

Chinn (M.), Meunier (B.) and Stumpner (S.) (2023)
"Nowcasting world trade with machine learning – a three-step approach", *Working Paper Series*, No. 917, Banque de France, juillet.
Download the document

Efron (B.), Hastie (T.), Johnstone (I.) and Tibshirani (R.) (2004)
"Least angle regression", *Annals of Statistics*, Vol. 32, No. 2, pp. 407-499.

Goulet Coulombe (P.) (2020)
"The macroeconomy as a random forest", *arXiv pre-print*.

Guichard (S.) and Rusticelli (E.) (2011)
"A dynamic factor model for world trade growth", *OECD Economics Department Working Papers*, No. 874.

Jakaitiene (A.) and Dees (S.) (2012)
"Forecasting the world economy in the short term", *The World Economy*, Vol. 35, No. 3, pp. 331-350.

Jardet (C.) and Meunier (B.) (2022)
"Nowcasting world GDP growth with high-frequency data", *Journal of Forecasting*, Vol. 41, No. 6, pp. 1181-1200.

Kaiser (H.) (1960)
"The Application of electronic computers to factor analysis", *Educational and Psychological Measurement*, Vol. 20, pp. 141-151.

Keck (A.), Raubold (A.) and Truppia (A.) (2010)
"Forecasting international trade: a time series approach", *OECD Journal: Journal of Business Cycle Measurement and Analysis*, No. 2009/2.

Martínez-Martín (J.) and Rusticelli (E.) (2021)
"Keeping track of global trade in real time", *International Journal of Forecasting*, Vol. 37, No. 1, pp. 224-236.

Stock (J.) and Watson (M.) (2002)
"Forecasting using principal components from a large number of predictors", *Journal of the American Statistical Association*, Vol. 97, No. 460, pp. 1167-1179.

Glossary

CPB: *Centraal Plan Bureau*.

LARS: Least-angle regression.

ML: Machine learning.

MRF: Macroeconomic random forest.

OLS: Ordinary least squares.

PCA: Principal component analysis.

RMSE: Root mean square error.

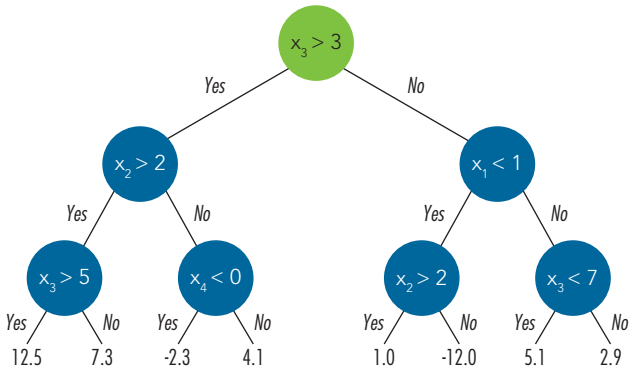
Appendices
Methodological elements

1 Definitions of some technical terms

As we test over a range of different techniques for regression (see Annex 3), an important ingredient in our work is the distinction between machine learning techniques based on **decision trees** and those based on **regressions**. The first category (random forest, gradient boosting) is the most widely used in the literature and works by aggregating several "decision trees" together. The second category of **regression**-based techniques (macroeconomic random forest, linear gradient boosting) is much less used in the literature. It is an adaptation of the first category but using linear regressions instead of, or in complement to, decision trees.

A decision tree is an algorithm used for classification or regression. A tree is composed of different nodes connected between them. Each node is a split point, corresponding to a test (a statement, which is true or false, like the examples shown in the bubbles on Chart A1) based on the value of a variable x (which can differ at each node, see for example x_1, x_2, or x_3 on Chart A1). If x meets the test (in bubbles on Chart A1), then the algorithm takes one path (a leaf) or otherwise takes the other path (a second leaf). These leaves lead to other nodes, and so on. At the extremities of all possible paths, the final leaves give the model predictions for the target variable (y, which in our case is CPB world trade) which are illustrated by the values at the bottom of Chart A1. Chart A1 provides a stylized example of a decision tree.

CA1 Example of decision tree



Source: Authors.

A random forest is an "ensemble method" using a large number of decision trees. The underlying idea is to build a large number of un-correlated trees. Then, by averaging predictions over multiple decision trees, the variance of the aggregate prediction is reduced. Key is to obtain independence between the different trees. In a random forest, this is ensured by (i) taking a different bootstrapped sample (method consisting in drawing randomly a random sample) for each tree, and (ii) considering only a subset of variables for each tree. With several independent trees, pooling the independent predictions lowers the variance and therefore increases the performance of the model.

Macroeconomic random forests (MRF; Goulet Coulombe, 2020) extend the canonical random forests which are constructed by pooling together several decision trees – hence a forest – to obtain a prediction. However, these canonical random forests tend to be too flexible for macroeconomic time series with few observations. To address this, the MRF builds on a linear part $y_t = X_t\beta_t$ as in a plain OLS – where y_t is the target variable, CPB world trade, X_t a vector of explanatory variables, and β_t the associated coefficients. But, unlike in OLS, coefficients β_t can vary through time according to a random forest. Formally, $\beta_t = F(S_t)$ where F refers to a random forest and is based on S_t, a set of variables potentially different from X_t.

LARS (Efron et al., 2004) is an iterative forward selection algorithm. Starting with no predictors, it adds the predictor x_i most correlated with the target variable y and then increases the (absolute) value of coefficient β_i so that the correlation of x_i with the residual $(y - \beta_i x_i)$ decreases. It does so until another predictor x_j has similar correlation with $y - \beta_i x_i$. There, x_j is added to the active set and the procedure continues with moving both coefficients β_i and β_j by the same amount, until another predictor x_k has as much correlation with the residual (now $y - \beta_i x_i - \beta_j x_j$). This algorithm provides a ranking of all variables by the order at which they are added. As it accounts for variables already selected, the algorithm ensures the complementarity between selected variables.

2 How do we manage the data flow in real time?

In real-time, asynchronous publication dates across the different variables lead to a "ragged-edge" pattern. Each variable has a different number of missing observations at the end of the dataset, depending on the publication delays. To address this issue, we apply the "vertical realignment" of Altissimo et al. (2006). For each variable, the last available point is taken as the contemporaneous value and the entire series is realigned accordingly. For example, for a forecaster in March 2023 who wishes to nowcast with a variable whose last observation is December 2022, the series is "re-aligned" by taking the December 2022 value as the value for March 2023. Formally, if the last observation of x_t at time T is at $T-k$, the re-aligned series is $\tilde{x}_t = x_{t-k}$ for all t between 0 and T.

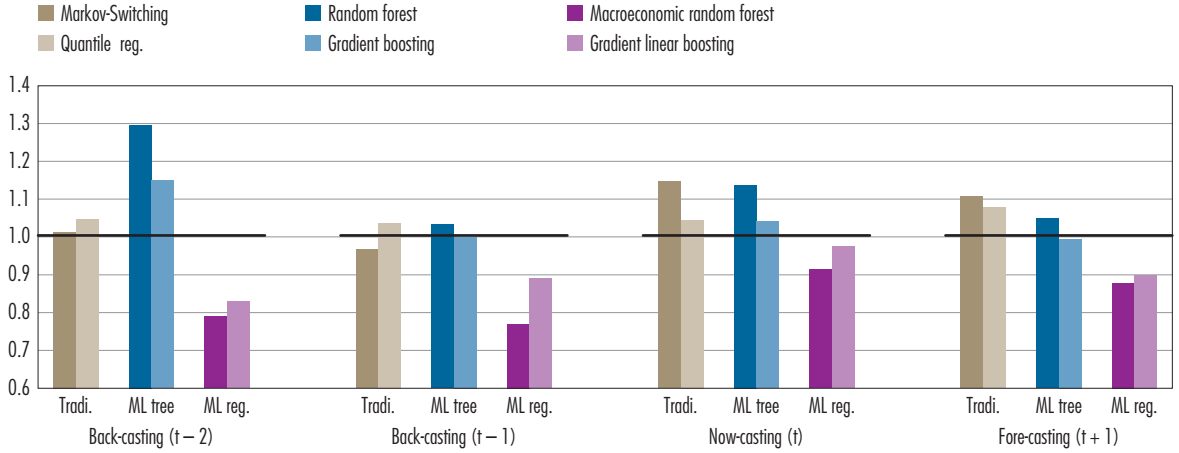
In addition, some series can have value after the date to forecast. For example, a forecaster in March 2023 might be willing to back-cast world trade in January 2023 – given long publication lag. It can happen that the timeliest series have already observations for February 2023. For example, this could be the case if the oil price (known every day) is an explanatory variable: from March 2023, the forecaster will have at his disposal the price of oil in February and March 2023. To account for this, we extend vertical realignment to avoid losing these "excess" observations. The series is re-aligned in the opposite direction as in Altissimo et al. (2006) by taking $\tilde{x}_t = x_{t+k}$ instead of $\tilde{x}_t = x_{t-k}$. Unlike in Altissimo et al. (2006) where the re-aligned series $\tilde{x}_t$ replaces the original x_t, the re-aligned series comes as a new variable. In the example for oil prices, this procedure would end up with the creation of two additional series: one re-aligned on $t-1$, meaning it would use oil prices in $t+1$ (February 2023) to predict trade in t (January 2023); and another re-aligned on $t-2$, meaning it would use oil prices in $t+2$ (March 2023) to predict trade in t (January 2023).

3 What about other non-linear regression techniques?

One interest of the three-step approach is the flexibility to incorporate different regression techniques. Before turning to the macroeconomic random forest, we test other non-linear approaches. A first set consists in "traditional" non-linear regressions: Markov-switching and quantile regressions. A second set relates to machine learning techniques. Within this second set, we distinguish between **tree**-based and **regression**-based techniques. **Tree**-based techniques include notably the random forest which pool

CA2 Accuracy (out-of-sample RMSE) relative to the linear (OLS) benchmark

(percentage change, year-on-year)



Source: Authors.
Interpretation: A lower root mean squared error (RMSE) indicates a higher accuracy.
Note: Accuracy is measured by the out-of-sample RMSE over Jan. 2012 – April 2022. Performances are presented relative to the OLS benchmark (black straight line at 1.0). Results are obtained for the average of datasets mirroring data available to a forecaster at the 1st, 11th, and 21st days of the month, using a LARS for pre-selecting the 60 most informative regressors, with factors extracted through PCA on the pre-selected set. "ML tree" = machine learning techniques based on **decision trees**; "ML reg." = machine learning techniques based on **linear regressions**.

predictions from numerous trees. It also includes gradient boosting, which is also based on trees but instead of pooling several independent trees, builds a model by adding trees iteratively – where each new tree depends on the results of past ones. The macroeconomic random forest belongs to the **regression**-based class. As explained in Annex 1, this is indeed an extension of the random forest but using a (non-linear) regression to discipline coefficients. This category also includes linear gradient boosting, a technique which works similarly as the canonical gradient boosting explained above but adds **linear regressions** instead of **trees**.

Chart A2, which displays the accuracy relative to the OLS (black line), shows that the macroeconomic random forest outperforms all other techniques employed. All results use the three-step approach with LARS and PCA – therefore any difference comes only from regression. Even in this case, the macroeconomic random forest (dark pink) beats significantly the OLS. It also outperforms other non-linear methods, traditional, and **tree**-based machine learning (ML). The outperformance is significant and consistent over the different horizons. The only method whose accuracy is close to the MRF is the linear gradient boosting (light pink) which is another machine learning technique based on linear regressions – suggesting the superiority of this class of technique for non-linear forecasting.

Published by
Banque de France

Managing Editor
Claude Piot

Editor-in-Chief
Claude Cornélis

Editor
Nelly Noulin

English Editor
Authors

Technical production
Studio Creation
Press and Communication

ISSN 1952-4382

To subscribe to the Banque de France's publications
https://www.banque-france.fr/en/alertes/abonnements