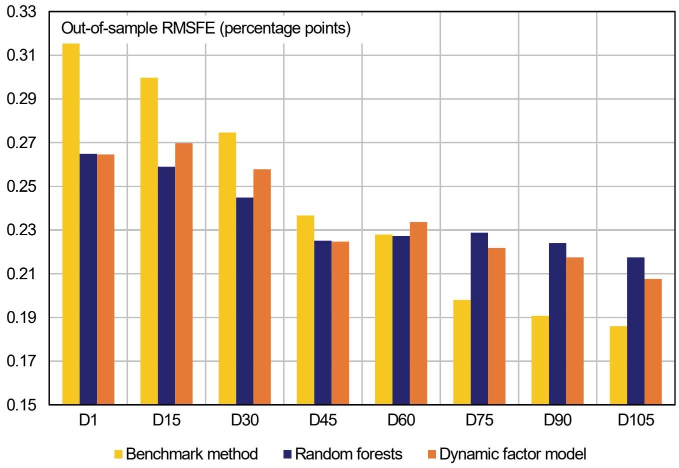
Trésor-economics

No. 254 • January 2020

Improved GDP nowcasting using large datasets

- Business outlook surveys, economic data, and financial time series provide a considerable amount of information available to forecasters for GDP nowcasting, namely to predict economic growth in the very short term . Business surveys generate not only synthetic indicators, but also disaggregated subsector survey data, containing much supplementary information in over a thousand time series that can be harnessed in an attempt to improve short-term forecasts.
- Conventional forecasting methods are ill-suited to processing such large datasets, and most GDP nowcasting remains based on linear regression on a relatively small number of variables. The past two decades, however, have seen the development of statistical methods capable of manipulating far larger datasets. One example is found in dynamic factor models, which can be used to summarise information efficiently while requiring only limited computing resources.
- More recently, as computing power has increased, machine learning methods have been developed and have gained popularity. These methods apply new techniques for filtering, sorting and processing information, such as random forests and neural networks.
- Some of these methods can improve short-term GDP growth forecasting performance by using large databases that include, among other things, subsector-level survey data chosen through a variable preselection process. Random forests appear to offer an appropriate method for selecting, at different dates, the variables most likely to provide information on current GDP.
- The greatest gains in GDP nowcasting performance arising from models based on large datasets over conventional models occurs in the earlier part of the quarter, before the first "hard" (quantitative) data becomes available.

Forecast error as a function of the forecast horizon



Source: Insee, Banque de France, PMI and financial data; DG Trésor calculations.

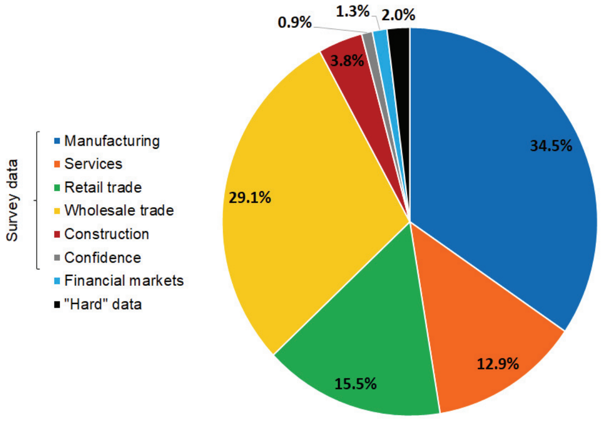
How to read this chart: The root mean square forecast error (RMSFE) is the square root of the average of the squared values of the forecast errors. D1 corresponds to the forecasts made using all data available on the 1st day of the quarter in question, D15 corresponds to the forecasts using all the data available on the 15th day, and so on; D105 corresponds to the data available just before the release of the first official estimate of GDP.

1. Hundreds of datasets are available for GDP growth nowcasting

Before publication of the first estimate of GDP – roughly thirty days after the end of the quarter – forecasters try to predict quarter-on-quarter GDP growth in real time, in order to provide as accurate a diagnosis as possible to inform economic policy making. Vast amounts of data are available for forecasting. In addition to "hard" data[1] (consumption expenditure, industrial production, etc.) and financial data (stock market prices, interest rates, VIX volatility index, etc.), considerable amounts of survey data[2] are available, both at an aggregate level (e.g., manufacturing) and at a more-detailed subsector level (e.g., automotive). The database available in France each quarter, and which is used here, thus contains over a thousand variables, the vast majority of which are survey data (chart 1).

The outlook survey data is published throughout the quarter, at different dates depending on the indicator. Insee publishes its monthly survey data at the end of the month in question, at the same time the PMI (Purchasing Managers Index)[3] is released, whereas it takes an additional ten days or so for the Banque de France survey data. The first "hard" data is available later, with monthly household consumption expenditure on goods available with a lag of 30 days, and the industrial production index (IPI) with a lag of 40 days. Taking the staggered publication dates into account, it is possible to build databases for each fifteen-day period, to include the information available to the forecaster at the time (table 1).

Chart 1: Database components



Sources: Insee, Banque de France, PMI and financial data; DG Trésor calculations.

Table 1: Date of release of main indicators

	Date of release	Survey	Main hard data
Month 1	D0	Insee and PMI (month 0)	
	D15	Banque de France (month 0)	
	D30	Insee et PMI (month 1)	Household consumption expenditure on goods (month 0)
Month 2	D45	Banque de France (month 1)	IPI (month 0)
	D60	Insee anf PMI (month 2)	Household consumption expenditure on goods (month 1)
Month 3	D75	Banque de France (month 2)	IPI (month 1)
	D90	Insee and PMI (month 3)	Household consumption expenditure on goods (month 2)
Month 4	D105	Banque de France (month 3)	IPI (month 2)
	D120	**Publication of GDP**	

How to read this table: In addition to the data already present in the database on day 30 (D30 database), the D45 database contains the Banque de France surveys for the first month in the quarter, and the IPI for the last month preceding the current quarter.
Note: This table refers to the more conventional short-term economic indicators and does not include financial variables that are available each day in near-real time.

[1] The term "hard data" refers to quantitative information, e.g., the industrial production index (IPI), as opposed to more qualitative data, e.g., data from surveys.

[2] Data for short-term economic indicators comes from (generally monthly) panel surveys of businesses or households.

[3] Another short-term economic indicator is the Purchasing Managers' Index (PMI), which is produced by IHS Markit for a large number of countries.

2. Innovative statistical methods are used to process large datasets

Conventional economic forecasting is based on econometric models that postulate what is most often a linear relationship between the variable to be forecasted (here, quarterly GDP) and explanatory variables that are highly correlated with the target variable.

These econometric models have recently been supplemented by machine learning techniques from the realm of artificial intelligence (diagram 1). To apply econometric techniques, the forecaster must make a priori assumptions regarding the form of the relationship between GDP and its explanatory variables; machine learning involves a far more open heuristic approach, with algorithms that will learn from their mistakes.

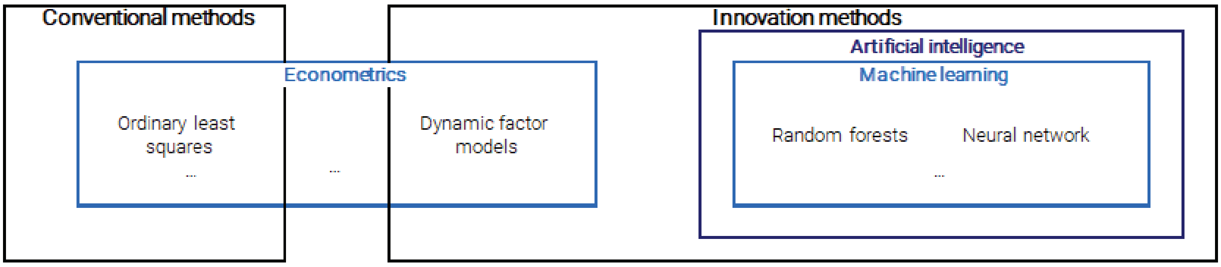
Conventional linear models on their own are incapable of processing such large datasets, for two main reasons: first, many variables are correlated, and second, variable preselection processes[4] are ill-suited to cases where the number of variables exceeds the number of observations available for each. More recent techniques, e.g., dynamic factor models (DFMs), random forests, and neural networks, are required to deal with such large datasets.

Dynamic factor models use factor analysis,[5] the principle of which is to summarise the information in a large set of variables into a relatively small number of components, called factors.[6] In forecasting quarter-on-quarter GDP growth, the estimated factors are then used as explanatory variables in a conventional linear econometric model.[7]

The random forests method[8] is a generalisation of the decision tree technique that seeks to predict the value of GDP based on a set of criteria for the explanatory variables (box 1). The random forests method starts by drawing random samples from the entire set of available data; then, for each sample, a decision tree is optimised to predict GDP. The predictions of each tree are aggregated to obtain the final prediction of the forest. This method also allows ranking of the explanatory variables by order of importance for predicting the variable of interest.[9]

The neural networks method is based on models originally inspired by the operation of biological neural networks. (Box 2 briefly describes the architectural principle of artificial neural networks, and how they are used).

Diagram 1: Econometrics and machine learning



Note: This diagram presents only the methods discussed in this paper.

[4] See Krolzig H.-M. and D. Hendry (2000), "Computer Automation of General-to-Specific Model Selection Procedures, *Economics Series Working Papers* 3, University of Oxford Department of Economics.

[5] For a detailed presentation, see Bessec M. and C. Doz (2011), "Prévision de court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques", *DG Trésor working document* no. 2011/01.

[6] These models provide a concrete response to the need for parsimony, in that they make it possible to avoid overfitting (a situation where the set of exogenous variables offers a very good explanation of the GDP observed in the sample, but where adding further observations results in a severe degradation of the model's predictive performance).

[7] The forecast is made using ordinary least squares.

[8] See Breiman (2001), "Statistical Modeling: The Two Cultures".

[9] The importance of each variable can be measured by computing the increase in the forecast error when the values of the variables are disturbed. A small increase in the forecast error would indicate that the variable in question is relatively unimportant for the forecast, whereas a significant increase in the forecast error would indicate that the variable is important for the forecast.

Box 1: Decision trees and random forests

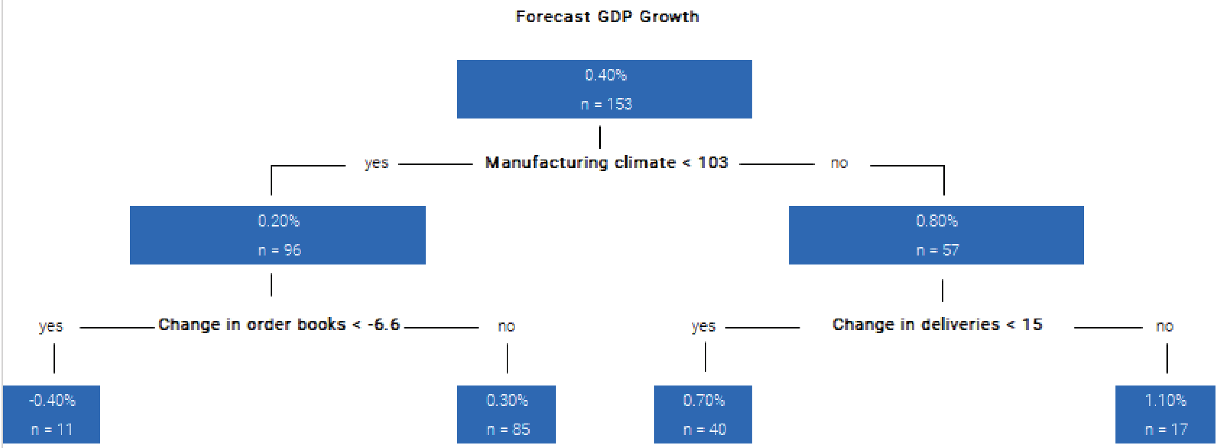
A random forest is a set of decision trees built using subsets of randomly selected data.

A decision tree aims to predict a variable of interest by applying a set of predetermined criteria. Decision trees are built using an iterative process that recursively splits the starting sample into two groups according to a criterion for a variable, such that it minimises intragroup variance and maximises intergroup variance: this defines a node, which corresponds to a decision criterion for one of the variables in the sample. The process ends when the number of observations in the resulting groups falls below a predefined threshold.

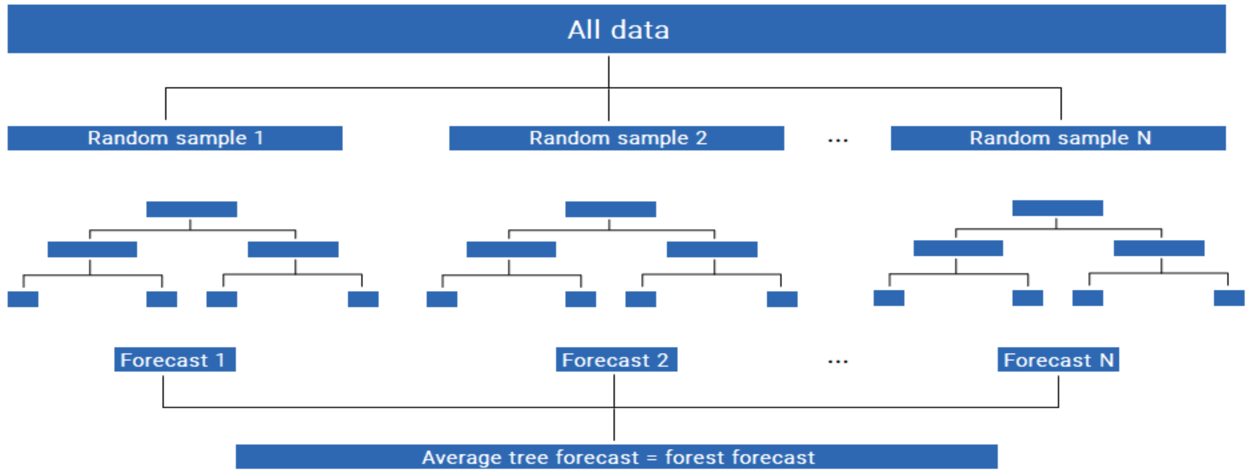
The random forests method involves generating a multitude of decision trees. Each tree in the forest is built from a subset of data established by bootstrapping, that is, drawing a random sample, with replacement, from the initial database. Bootstrapping the sample in this way makes the model less subject to overfitting. Each split into nodes introduces a further random element, because only some – randomly chosen – variables in the database are tested. This means that the same discriminant variables are not used in all trees.

The figure below represents a decision tree, once its parameters are known. In each node, the first line corresponds to the mean of the variable of interest (here, GDP growth) for the observations in the sample for that node, and the second line corresponds to the number of observations in that sample. Below, one can read the condition that splits observations into two groups. The condition is constructed to obtain (using a variance criterion) the most homogeneous data possible within the node.

In this example, for an observation in which the manufacturing climate is 102, the change in orders is 5, and the change in deliveries is 15, the forecast for GDP is 0.3%. The first node (climate = 102 < 103), where the splitting condition is "strictly less than 103", leads to the sub-tree on the left-hand side; and the node for orders, because the splitting condition "strictly less than –6.6" is not satisfied (5 > –6.6), leads to the right-hand side of the sub-tree. The variable for deliveries is not used for this observation.



When all the trees have been built, the GDP prediction corresponds to the average of the predictions of each tree. More specifically, when a new observation arrives, each of the decision trees is applied to that observation, yielding as many GDP predictions as there are trees, and the average of those predictions is the prediction of the forest.

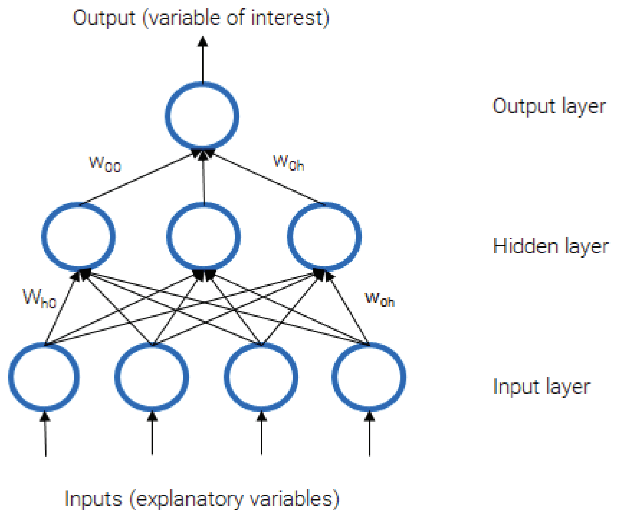


Box 2: Neural networks

Neural networks are a set of functions that aim to predict a variable of interest from observed variables. The form of the neural functions is determined on the basis of a database, and the prediction is then generated from a new observation.

A neural network is made up of multiple "neurons" that aggregate information. The neurons are organised in layers: an input layer, an output layer, and possibly one or more intermediate layers called hidden layers. The number of input neurons corresponds to the number of explanatory variables entering into the model, and the output neuron corresponds to the predicted variable of interest – in this case, the predicted change in GDP. In each layer, each neuron calculates the weighted sum of its inputs (w_{00}, ..., w_{hh}) and applies an activation (or transfer) function (e.g., logistic function or sigmoid function) to generate the output.

The specification of the neural network accordingly depends on estimation of its parameters (weights and biases), once the architecture has been defined (number of layers, number of neurons per layer, the activation functions, and so on). The weights and biases are estimated using a gradient descent optimisation algorithm to minimise the forecast error; the numbers of neurons and neuron layers, and the activation functions, are most often determined empirically.[a]



[a] In practice, the literature indicates that neural networks built with a single hidden layer provide good forecasting performance and represent the architecture most often used for purposes similar to those described in this paper (Kaastra and Boyd, 1996).

3. Subsector survey data can improve GDP growth forecasts

The three types of models - DFM, random forests, and neural networks - are applied to three different databases:

- the "narrow" database comprising aggregated survey data, "hard" data, and financial data
- the "broad" (or exhaustive) database containing the detailed (subsector) survey data,[10] "hard" data, and financial data
- the "filtered" database comprising the 100 most important variables selected in the broad database using a random forest method

Throughout each quarter, the databases are rebuilt and the forecasts are made every 15 days, as the various variables are published. All told, for each fortnight, there are eight methods for estimating GDP.[11]

[10] The "detailed" data corresponds to survey data for a subsector, e.g., the automotive industry, as opposed to survey data for a broad sector, like manufacturing, which is referred to as aggregated data.

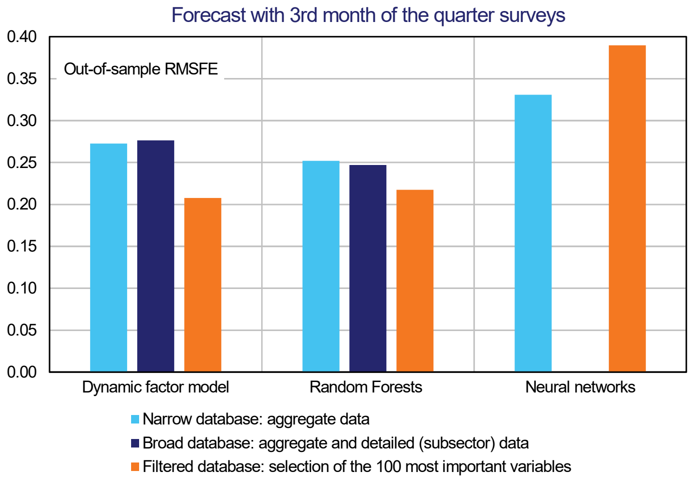
[11] Neural networks are not used for the "broad" (exhaustive) database, because computing time would be too long.

The comparison of the estimates obtained with these eight methods shows that simply increasing the size of the databases does not improve the forecast, even using the new techniques, as the models all generate similar results whether they are applied to the narrow database or the broad database (chart 2).[12]

The results improve significantly when the models are applied to a "filtered" database, that is, to the set of data filtered from the exhaustive database by the random forests.

Neural networks, on the other hand, appear to be less precise when applied to all the databases – even the filtered database – probably because the available time series are too short and the data too numerous to allow accurate estimation of the necessary parameters. This problem arises less when forecasting financial variables, which are available virtually daily on the markets, providing enough observations for estimation. This explains the popularity of neural networks in the financial sector.

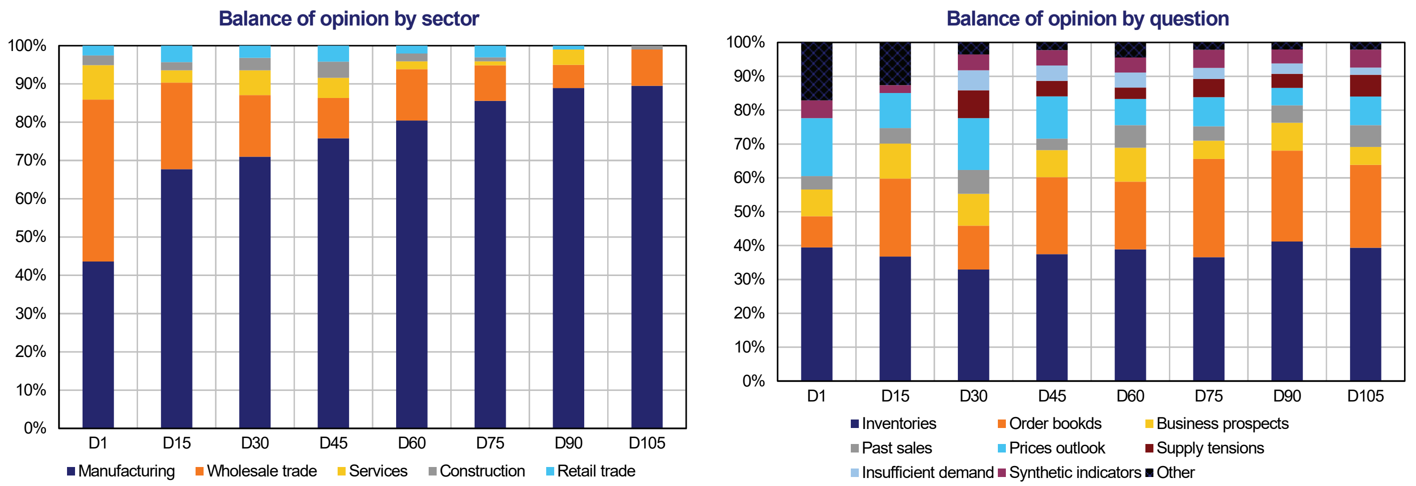
Chart 2: Forecast error by type of model and type of database



Sources: Insee, Banque de France, PMI and financial data; DG Trésor calculations.

How to read the chart: This chart displays the out-of-sample RMSFE depending on the model and the database used. Neural networks using aggregated and disaggregated data could not be estimated because they would require too much computing time.

Chart 3: Balance-of-opinion survey contributions by sector and by survey question in the "filtered database" with 100 variables



Sources: Insee, Banque de France, PMI and financial data; DG Trésor calculations.

How to read these charts: These charts set out the relative percentages of the balances-of-opinion in the surveys selected by random forests, by sector of activity and by survey question, depending on the forecast horizon. D15, for instance, corresponds to the database selected for a forecast generated on the 15th day of the quarter in question.

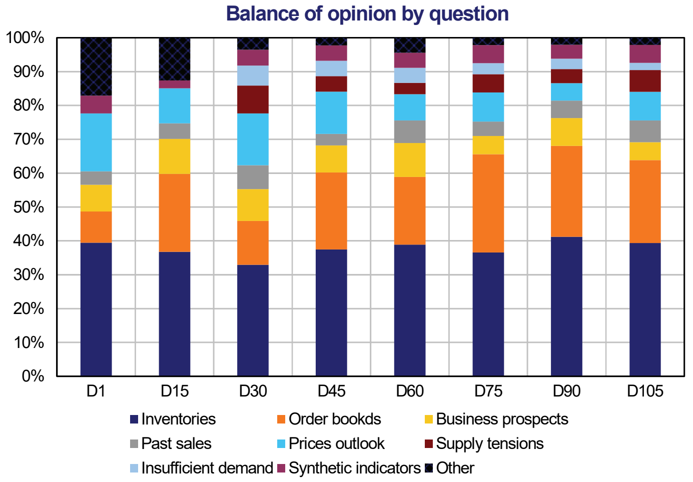
[12] The models are estimated for the 2000-2011 period and the forecasts for 2011-2018, with the parameters reestimated at each time point.

Most of the data in the database filtered by the random forest comes from the detailed (subsector) surveys – from 70% to 85% of the data, depending on the forecast horizon; this shows the extent to which the subsector surveys provide supplementary information on the economy.

More specifically, most of the survey data selected by random forests is linked to the manufacturing and wholesale trade sectors, most often relating to businesses' opinions on their order books and on their inventories of finished goods or inputs. What is more, the percentage of data from the manufacturing sector in the filtered database increases over time in the course of the quarter.

In building the database, the random forests algorithm generally selects the most recent data, particularly early in the quarter (chart 4), thus assigning greater value to recently published data, irrespective of its source.

Chart 4: Balance-of-opinion survey contributions by source, over time



Sources: Insee, Banque de France, PMI and financial data; DG Trésor calculations.

How to read this chart: On the 45th day of the quarter, when the Banque de France survey for the first month of the quarter is released, the random forests algorithm selects the most important data; 43% of the data comes from the Banque de France surveys (of which 28% is "new" data, i.e., from the survey of the first month of the quarter), 44% comes from the Insee surveys, and 8% comes from the PMI surveys.

4. Random forests and dynamic factor models improve forecasting performance mainly in the earlier part of the quarter

Random forests and DFMs improve GDP growth forecasting performance, compared to conventional methods, in the earlier part of the quarter. Neural networks, on the other hand, provide the least satisfactory results, at all times in the quarter.

More specifically, compared to a linear model with more conventional preselection of variables,[13] DFMs and random forests yield better results during the first two months of the quarter, i.e., the first half of the period during which forecasts are made. This superiority is particularly strong during the first two fortnights of the quarter (chart, page 1).

Remembering that the first data available is survey data – whereas the first "hard" data is available only at the end of the second month of the quarter – the new methods can be more efficient than conventional variable selection algorithms in selecting which surveys will be used. By contrast, during the last month of the quarter and in the first month of the following quarter, before the official growth estimate is released, the conventional method makes greater use of the industrial production index for the first month of the quarter.[14]

Because the computing demands of the new methods are very reasonable, the increased accuracy does not come at the cost of longer estimation times. On the other hand, unlike more conventional models, these methods do not allow direct analysis of the contributions of each explanatory variable to the GDP forecast;[15] they provide a rapid estimate of GDP growth, but cannot replace a detailed sector-by-sector analysis of the economy.

Maël Blanchet, Mélanie Coueffe

[13] The models are compared here to a linear model applied to a set of variables selected using a general-to-specific (GETS) procedure consisting only of synthetic survey indicators and the main hard data (see Krolzig and Hendry (2000), "Computer Automation of General-to-Specific Model Selection Procedures").

[14] The benchmark model gives significant weight to the industrial production index. In the random forests method, only one part of the explanatory variables is tested at each split, to avoid overfitting to the model. It follows that the industrial production index will not be used in each tree, and its weight will be far less in the forest than in the benchmark model.

[15] The relative contributions of the variables to the dynamic factor model forecast are not generated directly but can be obtained by ad hoc calculation; but the relative contributions to the random-forest forecast cannot be calculated.

Publisher:
Ministère de l'Économie
et des Finances
Direction générale du Trésor
139, rue de Bercy
75575 Paris CEDEX 12

Publication manager:
Bertrand Dumont

Editor in chief:
Jean-Luc Schneider
(01 44 87 18 51)
tresor-eco@dgtresor.gouv.fr

English translation:
Centre de traduction
des ministères économique
et financier

Layout:
Maryse Dos Santos
ISSN 1962-400X
eISSN 2417-9698

Recent Issues in English

December 2019

No. 253 French exports of goods to the EU
Orhan Chiali

No. 252 Use of macroeconomic modelling in public policy evaluation
Cyril De Williencourt, Florian Jacquetin

November 2019

No. 251 Gender budgeting issues in France
Axel Brunetto, Colette Debever, Mounira Nakaa, Louise Rabier

No. 250 Digital platforms and competition
Marion Panfili

https://www.tresor.economie.gouv.fr/Articles/tags/Tresor-Economics

in **Direction générale du Trésor (French Treasury)**

@DGTrésor

To receive *Trésor-Economics:* tresor-eco@dgtresor.gouv.fr

This study was prepared under the authority of the Directorate General of the Treasury (DG Trésor) and does not necessarily reflect the position of the Ministry of Economy and Finance.