Continuous forecasting of French economic growth:

Testing different models of machine learning

Paul-Armand Veillon

INSEE, Département de la conjoncture*

INSEE publishes its quarterly GDP growth forecast for the current quarter and the next quarter or two in the Conjoncture in France report each quarter. This forecast is based on those for each of the components of GDP such as household consumption or industrial production. The forecasts for these components are themselves based on short-term outlook indicators such as the business climate or the industrial production index. While only one forecast is published each quarter, new indicators are released almost daily and each new piece of information is likely to change the estimate of economic growth that appears most likely at a given date. New day-to-day or "nowcasting" forecasting models make it possible to take these frequent publications of new indicators into account for the quarterly growth forecast.

These models are developed through the use of statistical learning methods (known as "machine learning") on the one hand, and through open access in real time to hundreds of cyclical indicators ("open data") on the other hand. For example, since 2016, the Federal Reserve Bank (Fed) in Atlanta has published an updated growth forecast every week, based on a forecasting model of this type.

This brief presents a first proposal for continuous forecasting models for quarterly variations in French growth. The data used include the short-term outlook indicators published by the Banque de France, INSEE, OECD, Markit and various ministerial statistical offices. Several models are tested, including supervised statistical learning models such as random forest model and factor models.

The first results show that the forecast can vary significantly in the course of a quarter (between +0.2% and +0.4% for Q3 2019, for example), these variations following the publication of an indicator with a sharp rise or fall. The models used tend to converge at the end of the quarter and have an error, measured by the Root Mean Squared Forecast Error (RMSFE), of around 0.20 points. The forecast error ranges from 0.28 points at the beginning of the quarter to 0.20 points at the end of the quarter. The 80 % - confidence interval for Q3 2019 growth forecast thus rose from [–0.1; 0.6] in July to [0.0; 0.5] at the end of September. ■

* at the time of writing this study. The author would like to thank Clément Rousset for his help.

Continuous forecasting of French economic growth

INSEE growth forecasts are very heavily based on tendency surveys and indices such as the IPI

The first available estimate of current GDP, published in the national accounts, only becomes available one month after the end of each quarter. And yet, accurately forecasting short-term variations in GDP is a major priority for economic decision-makers. Their decisions are therefore informed by the short-term forecasts regularly published by various institutes and businesses. For example, in its Conjoncture in France published in December and June of each year, the INSEE makes forecasts for the next two quarters. These initial figures are then revised in the March and October forecasts. The forecasts published by the INSEE are based primarily on tendency surveys and short-term outlook indicators such as the industrial production index (IPI) or the turnover indices (CA). These forecasts are then integrated into an accounting framework which replicates the structure of the national quarterly accounts, ensuring consistency in terms of the accounting balances.

The rise of new statistical methods and the proliferation of data sources make real-time forecasting a possibility

Although only one forecast is published each quarter, it may be fine-tuned during the quarter in question following the publication of new indicators. The proliferation of data sources and the emergence of new forecasting methods now make it possible to continuously predict economic activity using a large number of short-term variables. These innovative methods, united under the umbrella term "nowcasting," provide a coherent statistical framework for the calculation of daily forecasts of GDP variation. By way of an example, the Federal Reserve in Atlanta is a pioneer in this field, publishing new forecasts which incorporate the most recent economic indicators on an almost daily basis. These sources range from the number of building permits issued to the level of production capacities, PMI indicators and surveys focusing on purchasing managers.

These methods are used here to create a new forecasting tool designed to continuously track the quarterly variations of French GDP. The database created for this purpose contains over a hundred temporal variables published by four different institutes. A daily forecast is produced using methods capable of summarising a very large number of variables in a single prediction.

Two results highlight the pertinence of such a tool. Firstly, forecasting error decreases continuously over the course of the forecasting quarter, falling by more than a third between the beginning and end of the quarter. As such the quality of each new prediction, measured by the degree of error in the empirical forecast, increases considerably. The best forecast is always that which is based on the most recent information. Furthermore, forecasting, as well as varying within the quarter, is highly sensitive to the publication of new indicators. Forecasts are therefore not to be considered as fixed values for a given quarter: they constantly evolve in response to the information available.

The diversity and frequency of the data available make continuous forecasting possible

Forecasters rely primarily on qualitative data for economic activity and quantitative data for production and consumption

The tendency surveys are the first data sources used by the forecasters. They are subsequently complemented by the publication of the first quantitative indicators such as the industrial production index or registration data, among others. Although these indicators provide more quantitative information than the qualitative questions contained in the business tendency surveys, their publication delay of over a month limits their usefulness for forecasting purposes. For example, the INSEE tendency survey for the manufacturing industry is published 25 days after the start of the month in question, whereas the industrial production index is published 40 days after the end of the month. As such, at the end of any given quarter, the forecasters are equipped with survey data for the whole period but quantitative data for the first month only. The advent of Big Data also raises the prospect of utilising new forms of data such as media articles, search engine traffic and even data obtained via social media. Nevertheless, the value of these new sources appears to be limited when it comes to analysing the French outlook (Bortoli & Combes 2015a, Bortoli et al. 2017).

The tendency surveys are the first indicators of economic activity available for forecasting purposes

The INSEE currently conducts around a dozen surveys covering households as well as businesses in the services, industrial and construction sectors. Their early publication makes them a useful variable for forecasters attempting to predict economic activity. By construction, these surveys are prospective: the 20,000 businesses included in the samples used for the tendency surveys are quizzed about their activity, their headcount and their expected output for the coming three months. They are also asked about the variation in these variables over the preceding three months. Their answers are summarised as "increasing," "decreasing" and "stable." The balances of opinion, which summarise these qualitative responses, are calculated as the difference between the percentages of "increasing" and "decreasing" responses. The forecasters use calibration techniques to determine the average relationship between these balances and economic activity, in order to construct forecasts. Other organisations such as the Banque de France and market analysis firm Markit also conduct tendency surveys. They provide information which is different but complementary to the INSEE surveys: they interview a different sample of businesses over a different period, with questions which are phrased differently from those used by the INSEE. The three composite indicators published by each of the organisations, although strongly correlated with one another, also demonstrate their own individual fluctuations. Furthermore, it may be pertinent to incorporate data from surveys focusing on the short-term outlook in the Eurozone or OECD countries, such as those published by the INSEE.

Quantitative indicators, published later on, by their construction provide a better quality of information on economic activity

Although the tendency surveys provide a useful signal as to trends in activity, this signal is sensitive to noise. Three-option qualitative responses cannot provide as much information as hard quantitative data. Furthermore, the questions may be open to interpretation (Bortoli et al. 2015b). Quantitative indicators, meanwhile, are based on real data such as household consumption or output figures. With the exception of vehicle registrations data, they are published more than a month after the fact, but provide quantitative information which is very close to the first estimates contained in the quarterly accounts. Three indicators published by INSEE are particularly important when constructing the first estimate of GDP: the industrial production index, published within 40 days, is an advanced indicator of industrial output compiled using data from the monthly branch surveys. The monthly series for household consumption of goods, published within 30 days, provide an initial estimate of the final consumption of households. The business turnover index, published almost 60 days after the end of the month in question and calculated using VAT declarations, gives an idea of spending on services. For these variables, the growth overhang is incorporated into the forecast. Financial variables such as loan demand from households and businesses, interest rates and market data can also be used to predict variations in GDP. The majority of these variables are published monthly by the Banque de France, and integrated into the forecasting database.

On average, a new indicator is published approximately once every three working days

The diagram below shows the date of publication of the principal indicators used for forecasting purposes within each quarter. The diagram begins on the first day of the quarter in question and ends 30 days after the end of the quarter, when the first estimate from the quarterly national accounts is published. In any given month the first available data are the tendency surveys published by Markit and the INSEE, around 18 and 24 days after the start of the month respectively, while the industrial turnover index is published 89 days after the start of the month. In total, over the four months shown here, new data are published on 34 of the 96 working days, an average of one new publication every three working days. It is thus theoretically possible to issue a new forecast every three days, incorporating a new set of information. Finally, of the 64 datasets published, 30 are tendency surveys conducted by the INSEE, the Banque de France or Markit, 13 are sets of financial data published by the Banque de France and the OECD, while 21 are quantitative indicators published by the Banque de France, INSEE and the Ministerial Statistical Services. This diverse array of indicators and data sources allows for forecasts based on a greater wealth of information than that generally used by forecasters, although the use of the resulting forecasts requires a certain degree of caution.

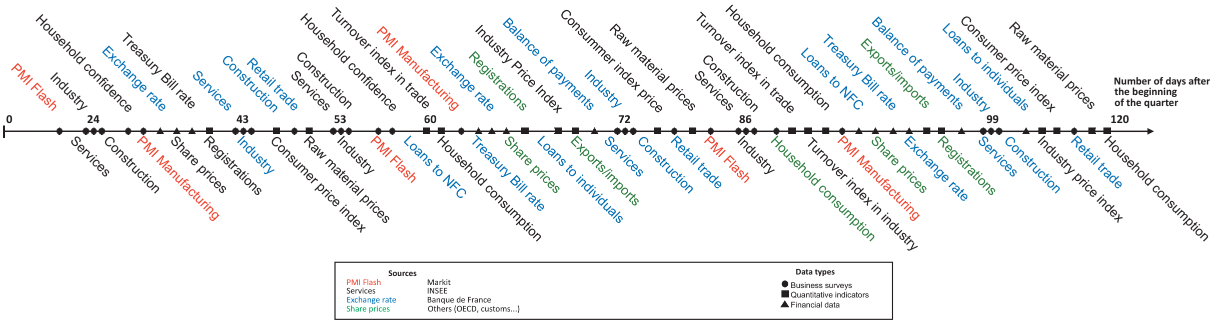
Thank to machine learning methods, it is possible to construct a forecast based on a number of indicators greater than the number of available observations

Two methodological problems arise when attempting to predict in real time the evolution of an economic aggregate: how do we aggregate data with different frequencies (monthly or quarterly) and publication dates? How do we construct forecasts with a number of variables (N) which is often greater than the number of observations (T), a problem which can be expressed as "N>T"? While the aggregation of heterogeneous and missing data is a problem specific to real-time forecasting, N>T is a classic dilemma of forecasting known as the "curse of dimensionality." Our method here is to apply the solutions proposed in the existing literature to the task of producing a first estimate of GDP growth for the quarterly national accounts.

There are various solutions to the problem of missing data

More than a decade ago, Dubois and Michaux (2006) were already examining the "problem of missing data" in relation to the quarterly forecasting of industrial output using the monthly tendency surveys. Their proposed solution was to create three quarterly series corresponding to the first, second and third months of each quarter. Depending on the availability of data, they would then integrate one, two or all three of these quarterly series. However, the drawback of this method is that it multiplies by three the number of variables, thus accentuating the problem of dimensionality. A common variant of these methods, known

1 - Calendar of publication of outlook indicators



as the bridge equation, is to predict the values for the missing months using an auto-regressive model. However, one consequence of extending the data in this manner is to add inertia to the forecast. As such, we have opted instead to calculate a quarterly average for the data available as of the forecasting date for tendency surveys, and to use the growth overhang for the other variables. The advantage of this approach is to prioritise the diversity of data sources over the addition of delays to a small number of variables, and also to make our forecasts more sensitive to the publication of new data.

Adding a large number of variables certainly improves the degree to which the model fits the data. Nevertheless, this adjustment may be detrimental to forecasting. In such situations, known as "overfitting," the estimated model is too close to the past data used and not sufficiently relevant to future developments. We thus felt it necessary to use a parsimonious model, incorporating a limited number of variables (*see the Box on Overfitting*).

One solution is thus to select a limited number of variables. Dubois & Michaux (2006), in the forecasts produced by the outlook department, were the first to employ a GETS (General to specific modelling) statistical method based on the selection of variables. This approach consists of successively eliminating non-significant variables, starting with the most general model and conducting a certain number of specification tests at each step. Where selection was previously done by hand or using less effective algorithms such as ascending and descending selection[1], using GETS instead enabled us, subject to certain conditions, to obtain the best linear forecasting model.

Factor models are capable of condensing a large number of variables into a few factors

Dynamic factor models offer a simultaneous response to the problems of missing data and high dimensionality. Pioneered by the work of Stock & Watson (2002) and Doz et al. (2011), these models have rapidly gained in popularity and are now used by many organisations, including the Fed and the ECB. Generally speaking, factor models allow us to obtain a parsimonious representation of a set of variables, summed up in a relatively small number of factors. The most well-known of these methods is principal component analysis. Meanwhile, dynamically representing these factors in the form of a space-state model allows us to take missing values into account. This method, which is conceptually very appealing, has been applied to forecasts for French GDP growth by Bessec & Doz (2012), and is also used in this article. A principal component analysis (PCA) model which does not take factor dynamics into account, an approach more frequently used in the existing literature, was also tested.

Statistical learning models offer new solutions to the "curse of dimensionality"

Different potential methods of machine learning (ML) represent a new approach to forecasting which no longer relies on the pre-specification of the relationship between an endogenous variable and the exogenous variables, but depends instead on an algorithm which finds the right model to minimise an objective function. Thanks to their predictive capacity, algorithms such as LASSO (Least Absolute Shrinkage and Selection Operator) and the random forest approach have spawned a growing body of literature focusing on the forecasting of macroeconomic aggregates with the help of Machine Learning. Biau, Biau & Rouvière (2006) notably applied the random forest method to the responses to the INSEE tendency surveys for the industrial sector, in order to forecast manufacturing output. Nevertheless, the deployment of these methods needs to abide by a certain number of elementary principles in order to avoid the pitfall of overfitting. Other automatic learning algorithms might also be used, such as neural networks. But these models often rely

[1] Stepwise ascending and descending selection algorithms allow us to test only a small number of models, which generally do not turn out to be the most effective.

upon a large number of parameters, requiring a quantity of observations so great that optimisation is not possible. The methods used in this case are LASSO[2] and random forests. The former allows us to create a linear model based on a sub-set of variables selected automatically, while the latter is based on the construction of decision-making trees (*see Box*).

The performance of these methods was then compared with that of a simple model using only the dynamics of the variable we were seeking to forecast (an autoregressive moving average model, or ARMA) and a simple calibration process based solely on the business climate in France.

The forecasting model yields results which vary considerably within the forecasting quarter, while error falls by around 40%

The quality of a forecast is measured in terms of its RMSFE (Root Mean Squared Forecast Error). As illustrated in *Figure 2*, for a given date t, the model is trained with data stretching up to date t and a forecast is then generated for the date *t+1*. The error on date *t+1* is calculated as the difference between the forecast and the value actually recorded on *t+1*. RMSFE is then calculated as the square root of mean forecasting error. The training data begin in Q4 2001 and forecasting errors are calculated for the period stretching from Q1 2011 to Q1 2019. In the rest of this section the forecasting data from Q3 2019 are given for

[2] The regulating hyperparameter λ was selected by a process of cross-validation with the training data.

2 - RMSFE calculation

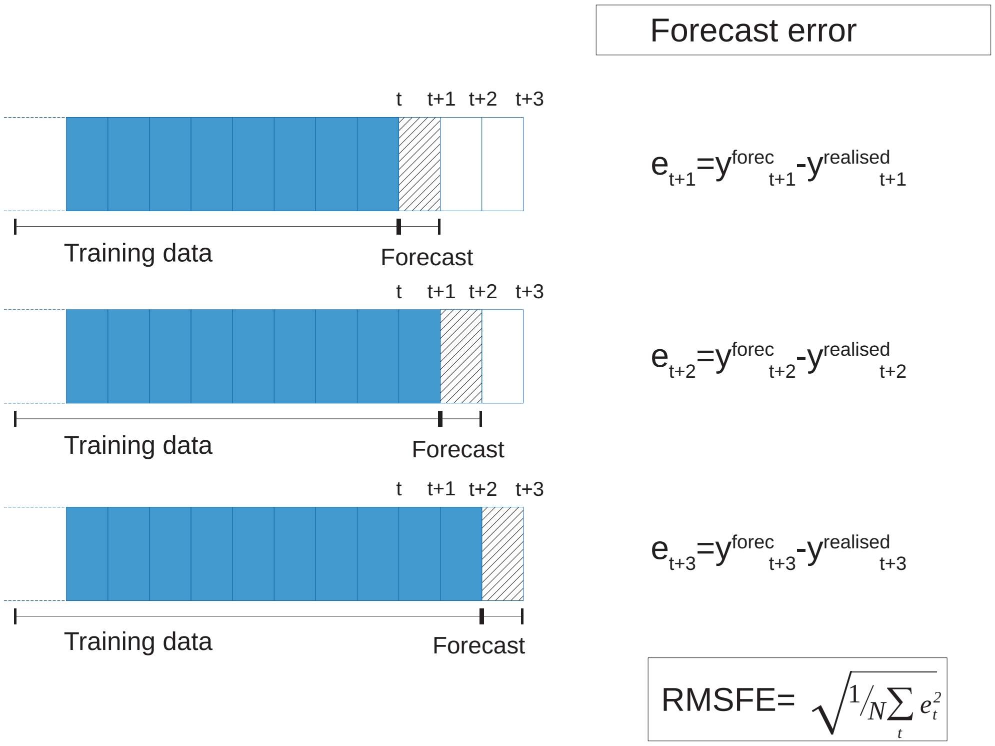

Forecast error
t t+1 t+2 t+3
Training data
Forecast
$e_{t+1}=y^{forec}_{t+1}-y^{realised}_{t+1}$
t t+1 t+2 t+3
Training data
Forecast
$e_{t+2}=y^{forec}_{t+2}-y^{realised}_{t+2}$
t t+1 t+2 t+3
Training data
Forecast
$e_{t+3}=y^{forec}_{t+3}-y^{realised}_{t+3}$
RMSFE= $\sqrt{{}^{1}/_{N}\sum_{t} e_t^2}$

illustration purposes, with the quarterly forecast for GDP growth in Q3 2019 as the objective.

Table 1 shows the RMSFE and the absolute value of maximum error with the data available 100 days after the start of the quarter, i.e. 20 days before publication of the first estimate in the quarterly accounts. All of the models perform better than those used as standard in this forecasting period. LASSO and random forest were the models with the lowest RMSFE.

While their forecasts follow a relatively similar progression, the models differ in terms of their volatility

As new information becomes available, the forecast evolves significantly and differently from model to model. Figure 3 shows the evolution of the forecasts for quarterly growth of French GDP in Q3 2019 generated by the LASSO, random forest and PCA models. The forecasts of all three models follow a broadly similar trajectory, with the main difference being their volatility or sensitivity to new publications. The PCA model is the most volatile, yielding a forecast which varies between +0.07% and +0.47%, followed by the LASSO model whose results vary between +0.17% and +0.43%. Finally, the random forest model yields forecasts varying between +0.18% and +0.38%. Higher volatility also indicates that the model's maximum absolute error is also higher. Calculated for all quarters preceding the most recent estimate, this maximum error ranges from 0.53% to 0.77% in absolute terms depending on the model. Nonetheless, the LASSO and random forest models, i.e. those with the lowest RMSFE, are almost perfectly identical throughout the quarter. In

Table 1 - Forecasting quality of the models used

Mean quadratic error and maximum error of the principal forecasting models used between 2011 and 2019

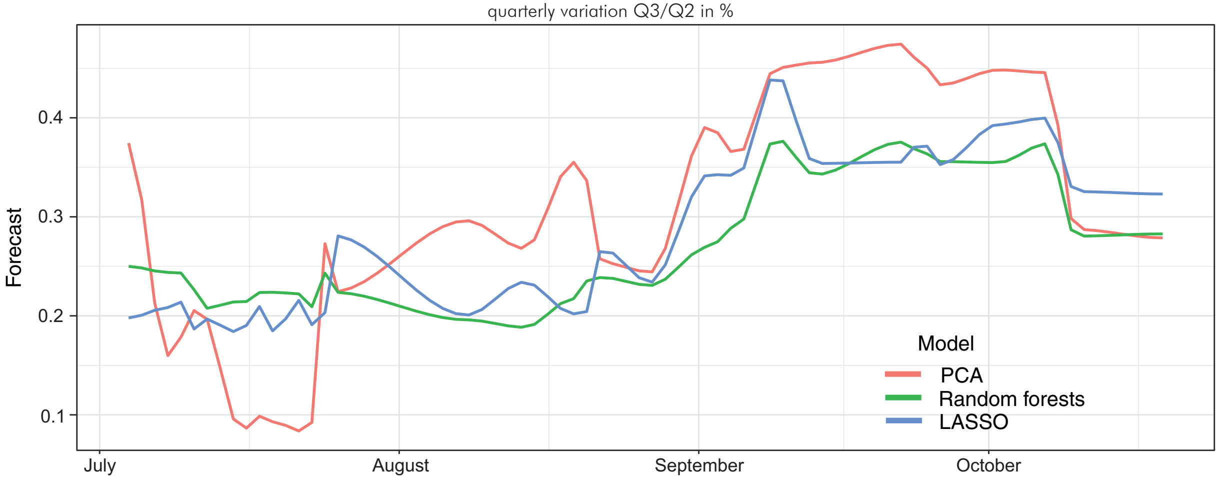
Model	Arima	Climate in France	Gets	LASSO	Forêts Al.	ACP	Dynamic-factor
RMSFE	0.33	0.28	0.23	0.20	0.19	0.23	0.22
Maximum error	0.77	0.55	0.65	0.53	0.55	0.62	0.66

Key: The maximum error of the LASSO model at T+100 days is 0.53 points (absolute value of the difference between the predicted quarterly growth rate of GDP and the rate actually recorded in current estimates) for the period 2011-2019. The corresponding RMSFE of 0.20 is calculated based on the forecasting error observed at forecasting date T+100 days, for all quarters in the same period

Source: INSEE, Banque de France, OECD, Markit, authors' calculations.

3 - Evolution of the GDP growth estimate over the course of Q3 2019

quarterly variation Q3/Q2 in %



Key: on 18 October 2019, the forecast yielded by the random forest model is 0.28.

Source: INSEE, Banque de France, OECD, Markit, authors' calculations.

the rest of this section we will look more closely at the random forest model, which offers the dual advantage of relatively low RMSFE and maximum error.

During the quarter in question forecasting error shrank by 40%

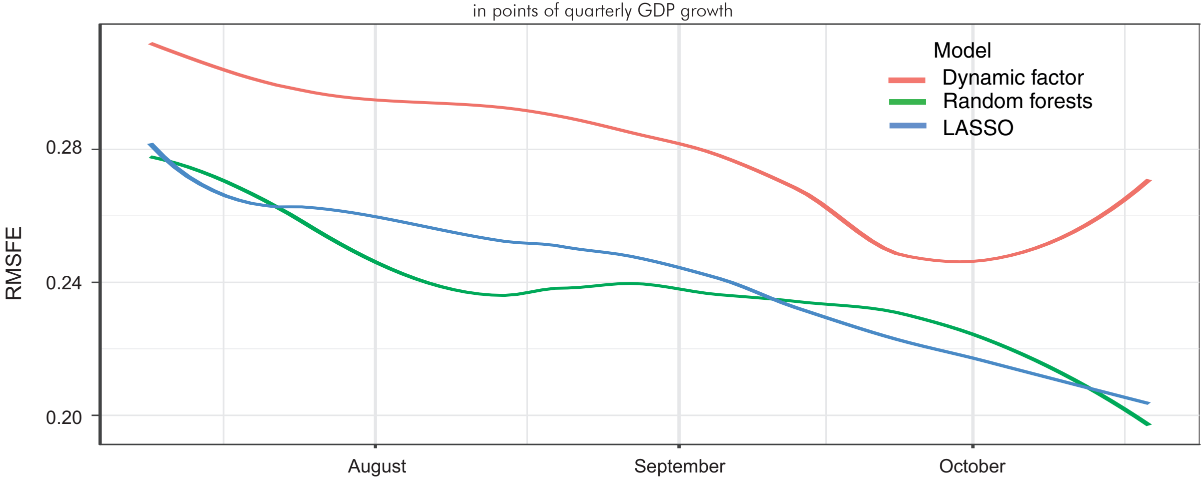
As the forecast evolves, the forecasting error shrinks as the quarter progresses, as more information becomes available regarding the current economic situation. *Figure 4* shows the evolution of the forecast generated by the random forest model and the reduction in its forecasting error over the course of Q3 2019. Forecasting error shrank by around 40% between the beginning of the quarter and the eve of the publication of the national accounts. To put it slightly differently, the 80% confidence interval of this forecast is +/– 0.38 percentage points at the start of the quarter and +/– 0.25 points by the end of the forecasting period.

Forecasting variations can be attributed to the publication of specific indicators

The growth forecast for Q3 2019 hit its lowest point in mid-August, at +0.18%. This coincides with the publication of two outlook indicators which are of particular importance for forecasting (cf. hereunder): the industrial output index for June, published on 9 August, dropped 2.2%; meanwhile the balance of output forecasts in the manufacturing

4 - Evolution of empirical forecasting error (RMSFE) during Q3 2019

in points of quarterly GDP growth

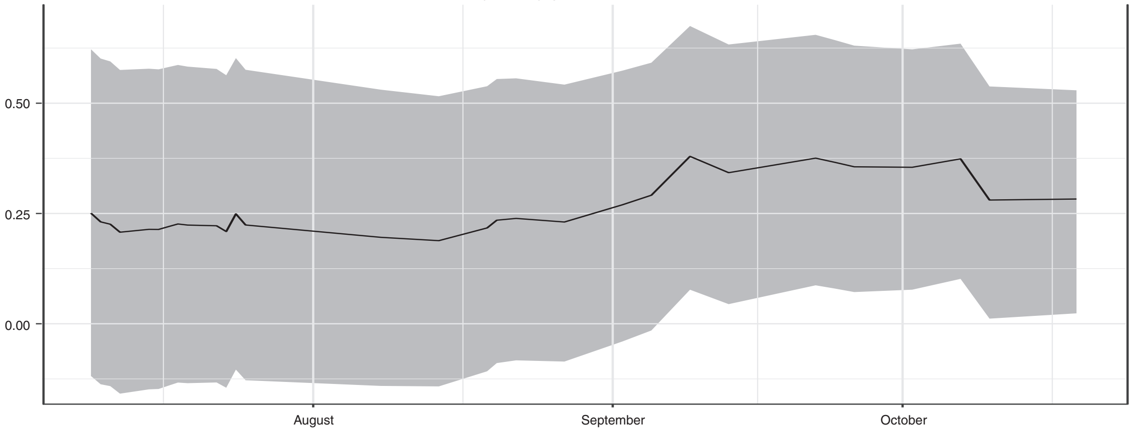


Key: on 1st September 2019, the forecasting error (RMSFE) of the random forest model is 0.24 points of GDP growth

Source: INSEE, Banque de France, OECD, Markit, authors' calculations.

5 - Growth forecast Q3 2019 with a confidence interval of 80%, using the random forest model

quarterly growth in %



Key: on 18 October 2019, the growth forecast for Q3 yielded by random forest model is +0.28%. The value with the confidence interval at 80% is between 0.03 and 0.52

Source: INSEE, Banque de France, OECD, Markit, authors' calculations.

industry, derived from the monthly outlook survey published by the Banque de France, dropped two points on the same date. One month later the forecast was back up to +0.39%, buoyed by the increase in the business climate indicator for industry, published by the Banque de France (+3.4 points), and the slight upturn in the industrial production index for the month of July, published on 10 September (+0.3%). The sharp decrease in the forecast in early October can be partly explained by the fall in the PMI indices, and also by the slight decrease seen in the IPI for August.

The random forest algorithm allows us to identify the most important variables for the forecasting of quarterly GDP growth

With the random forest method, it is possible to measure the importance of each forecasting variable (see the Box on Forecasting using the random forest method). This importance is calculated in terms of the predictive gain associated with each variable. For example, the balance of opinion for future output from the manufacturing industry is capable of reducing RMSFE by 13.5% for a forecast produced in mid-July. Tables 2 and 3 show the ten most influential variables for forecasts produced in the months of October and July respectively, i.e. one month after the end of Q3 2019 and in the first month of the quarter. The majority of the most influential indicators are connected with the manufacturing industry. Industrial output makes a very significant contribution to quarterly variations in GDP, a contribution which is disproportionate to its share of the total value added by all sectors. Moreover, the most influential variables are taken from a large number of different sources: OECD, Insee, Banque de France, Markit. This multiplicity of sources makes it possible to significantly improve the quality of the forecast. Finally, those indicators which can be considered weak signals, such as share prices, are among the most influential variables in July but have been superseded by October by quantitative indicators such as the industrial production index.

Table 2 - Importance of forecasting variables in the random forest model, as of mid-October (*T+100*)

Variables	Importance
Other industrial products (C5), variation in orders received – Banque de France, September	12.7
Manufacturing industry, past variation in output – Banque de France, September	12.4
Manufacturing industry, output forecast – Banque de France, September	10.8
Industrial production index, manufacturing industry – INSEE, August	10.0
Industrial production index, intermediate goods – INSEE, August	9.9
Industrial production index, capital goods – INSEE, August	9.1
Manufacturing PMI – Markit, September	8.5
Other industrial products (C5), output forecasts – Banque de France, September	7.0
Business climate in the construction industry – INSEE, September	6.9
Capital goods (C3), output forecasts – Banque de France, September	6.7
Manufacturing PMI, new orders – Markit, September	6.6
Monthly consumption of households, manufactured goods – INSEE, August	5.6

These new tools enable us to track in real time the evolution of economic forecasts as new indicators are published. They also allow forecasters, when used in conjunction with existing tools, to address new questions such as: "How did the forecast evolve over the course of the quarter?" "Which indicators had the biggest influence on forecasts?" and "How precise is our forecast at any given moment?". Nonetheless, this initial prototype has certain limitations and will require further research. First and foremost, machine learning is a field of research which has undergone a profound transformation over the past decade, and which continues to develop apace. The models of automatic learning used in this study may themselves need to evolve as further progress is made in the field. Moreover, real-time forecasting of quarterly growth is based on statistical analysis and is no substitute for economic analysis. It does not allow us to clearly establish a causal relationship between the fluctuations of a given indicator and the growth of GDP; it simply reflects the correlation between certain indicators and developments in GDP, based on historic data. Finally, the performance of our model over a single quarter is not sufficient proof of its robustness. It is therefore not possible to predict how it will react in times of crisis, periods in which, by definition, indicators depart significantly from their past trends. ■

Table 3 - Importance of variables in forecasting with the random forest model in mid-July (*T+15*)

Variables	Importance
Manufacturing industry, output forecast – Banque de France, June	13,5
Composite Index, business survey, OECD – OECD, June	9,5
Share prices, France – OECD, June	8,5
Other industrial products (C5), output forecast – Banque de France, August	8,0
Share prices, USA – OECD, June	7,8
Outlook turnaround indicator for services – INSEE, June	6,9
Transport equipment (C4), variation in orders received – Banque de France, June	6,7
Business climate, manufacturing industry – INSEE, June	6,6
Transport equipment (C4), variation in international orders received – Banque de France, June	6,6
Manufacturing PMI, new orders – Markit, June	6,1
Other industrial products (C5), variation in orders received – Banque de France, June	5,9

Key: in mid-July, the OECD's Composite business survey index improves RMSFE by 13.5%
Source: INSEE, Banque de France, OECD, Markit, authors' calculations.

Box 1: Overfitting

The purpose of a predictive model is to produce the most accurate forecast possible for an unobserved variable based on auxiliary observations. In this respect, the priority is not to maximise its adjustment with the data used in the estimation process: the objective is to build a model which is sufficiently general that it will yield a good forecast when used with new observations. The quality of a forecasting model is therefore assessed using a different data set from that used in its construction. To do this, all of the initial data is split into a learning sample, designed to estimate the properties of the model, and a validation sample, designed to assess the model's performance when used with an unknown sample.

The capacity of a model to be generally applicable is inversely proportional to its complexity, as per Occam's razor. The simpler a model is, the more its empirical performance will depend on the particularities of the data used in its estimation. To illustrate this principle, consider the example of a data set generated by a function (*f*) to which we then added noise (*epsilon*), so that: *y(x)=f(x)+epsilon*. We observe only *y(x)* and *x*. The forecaster's objective is to identify the function g which will best approximate *f*. We can approximate this function using a polynomial of degree p, wherein the higher the value of p, the more complex the model. *Figure 6* shows, by way of an example, the function *f* we wish to estimate (in black), an estimate calculated using a degree 2 polynomial (in red) and a degree 11 polynomial (in blue). Although the degree 11 polynomial is most closely adjusted to the data, the degree 2 polynomial offers a better estimate of *f*. The degree 11 polynomial mistakenly incorporates some of the uncertainty introduced in the data generation process. This is an example of overfitting.

6 - Simulation of overfitting

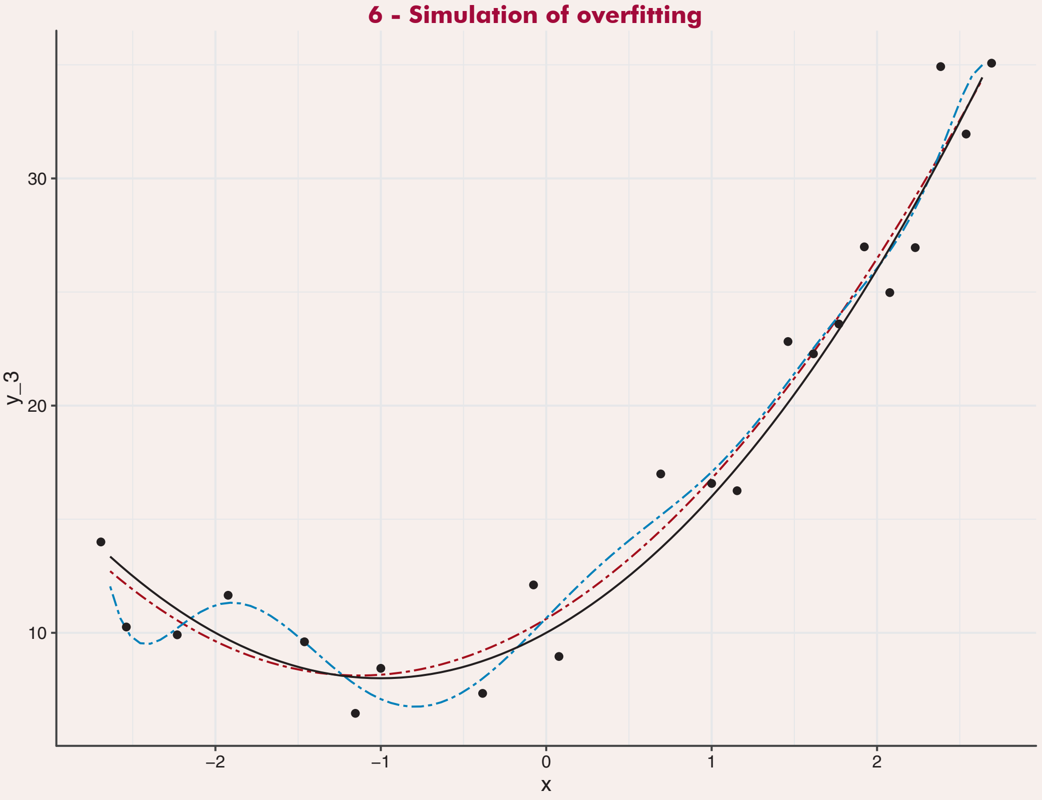
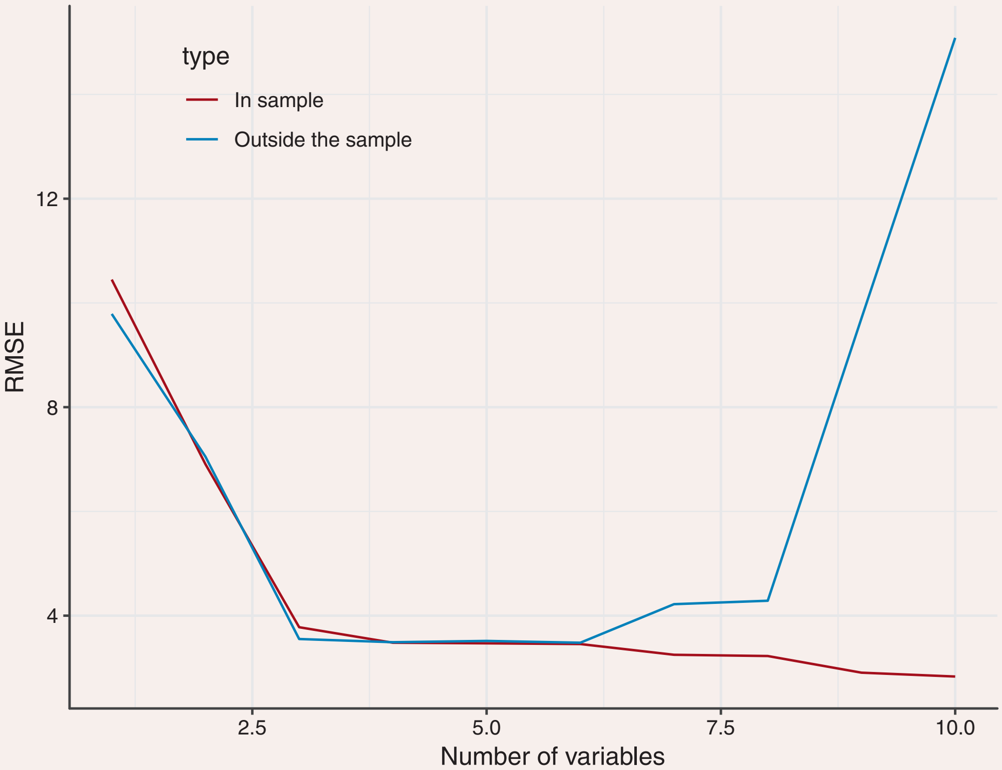


Figure 7 shows the error, in the form of RMSE (the Root Mean Square Error between the actual and predicted values), within the sample and outside the sample when we increase the number of variables. The error within the sample decreases along with the number of variables used. The more variables there are in the model, the better the model will fit the training data. However, when the number of variables exceeds 4, the error outside the sample increases. Once again, this is a case of overfitting. The model is not sufficiently general, and is too sensitive to uncertainty. ■

7 - Growth forecast for the third quarter of 2019 including confidence intervals, created using the random forest model



Box 2: Forecasting using random forests

The ‘random forest’ method is a machine learning method first developed by Leo Breiman in 2001. This algorithm is based on the construction of multiple decision-making trees, built using slightly different data samples.

Decision-making trees allow us to divide a set of observations into homogenous groups using a set of discriminant variables (predictive variables) and an output variable (predicted variable). They also have the advantage of being easy to construct and yielding a graphical representation which is simple to interpret. The trees are constructed using the CART[1] algorithm (Breiman, 1984). The general principle is to recursively divide the data set into groups. With each new division, the two sub-sets constructed are as homogenous as possible for the predicted variable[2]. The final step, known as pruning, involves constructing an optimal sub-tree using the final tree constructed in the previous step. The underlying idea is that the final tree contains a very large number of branches. This tree has very high variance and low bias; an example of overfitting. One solution is therefore to construct a family of sub-trees derived from the trimmed-down final tree, choosing from this family the tree which best minimises forecasting error.

Figure 8 shows a decision tree for forecasting the quarterly variations of GDP. This tree was created using all of the indicators available 20 days before publication of quarterly national accounts. It can be read as follows: if, in a given quarter, the growth overhang of the IPI in month 2 is greater than –1.5%, the standardised business climate in the construction sector is over 1.9 and the growth overhang of exports is greater than –1.7%, then the growth forecast is +0.97%. The percentage below the forecast indicates the share of observations from the sample which are included in this category. It is also worth noting that the algorithm has selected quantitative indicators as well as variables taken from the tendency surveys.

However, this algorithm has one major defect: instability. A slight modification to the sample may yield a very different decision tree, and thus very different predictions. The solution proposed by Breiman is to aggregate the predictions from multiple trees, generated with a degree of uncertainty. The algorithm is as follows:

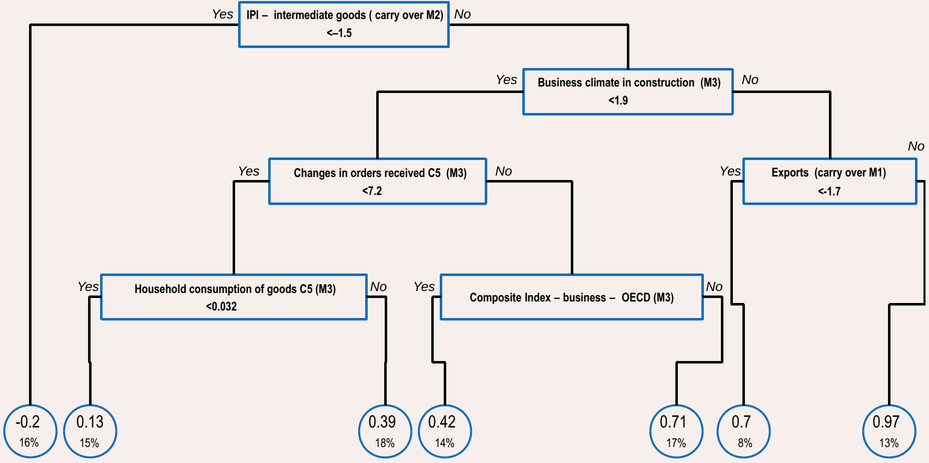
- Sampling with replacement of a set number *N* of observations in order to establish a training sample
- Random selection of *p/3* variables from the set of p predictive variables available
- Construction of a decision tree using these variables and the sample, using the CART algorithm
- Repeat this operation 1000 times to create 1000 different decision trees. The final prediction is the mean value of the predictions generated by all of the trees.

As such, each tree is generated by a different leaning process and their forecasts are weakly correlated. One of the key criteria for forecasters is the interpretability of the resulting model. With the random forest method this is made possible by quantifying the importance of the variables, calculated based on the predictive associated with each variable. *Tables 2 and 3* show the respective importance of these variables when forecasting quarterly variations at two different dates. ■

[1] Classification and Regression Trees

[2] To be more precise, with each division the two sub-sets minimise the variance within the sub-groups.

8 - Decision-making tree for forecasting the growth of GDP


Yes
IPI – intermediate goods (carry over M2)
<–1.5
No
Yes
Business climate in construction (M3)
<1.9
No
Yes
Changes in orders received C5 (M3)
<7.2
No
No
Yes
Exports (carry over M1)
<-1.7
Yes
Household consumption of goods C5 (M3)
<0.032
No
Yes
Composite Index – business – OECD (M3)
No
-0.2
16%
0.13
15%
0.39
18%
0.42
14%
0.71
17%
0.7
8%
0.97
13%

Reading: if the IPI acquisition is less than –1.5%, the model’s prediction is –0.1% of the training sample data have an IPI acquisition less than –1.5%

Bibliography

Bessec M. et Doz C. (2012), “Short-Term Forecasting of French GDP Growth Using Dynamic Factor Models”, *Économie et prévision*, 2012, n°199

Breiman L. (2001), “Random forests”. *Machine learning*, n°45, p.5-32

Breiman L., Friedman J., Olshen R. and Stone C. (1984), “Classification and regression trees”, Wadsworth & Brooks

Bortoli C. et Combes S. (2015), “Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues”, *Conjoncture in France*, Insee, March, p.43-56

Bortoli C., Combes S. et Renault T. (2017), “How to forecast employment figures by reading the newsaper”, *Conjoncture in France*, Insee, March, p.35-43

Bortoli C., Gorin Y., Olive P.-D. et Renne C. (2015), “New advances in the use of INSEE’s business tendency surveys to analyse the short-terme economic outlook”, *Conjoncture in France*, March, p.25-41

Doz C., Giannone D., et Reichlin L. (2011), “A two-step estimator for large approximate dynamic factor models based on Kalman filtering” , *Journal of Econometrics*, 2011, n°164

Dubois E. et Michaux E. (2006), “Étalonnages à l’aide d’enquêtes de conjoncture : de nouveaux résultats”, *Économie & Prévision*, n°172

Stock J. et Watson M. (2002), “Forecasting using principal components from a large number of predictors”, *Journal of the American Statistical Association*, 2002, n°460 ■