Euro area GDP forecast using large survey dataset

A random forest approach

Olivier Biau, Angela D'Elia

Directorate General for Financial and Economic Affairs - European Commission*

Abstract

Recent works in the econometric literature consider the problem of summarising efficiently a large set of variables and using this summary for a variety of purposes, including forecast (Stock and Watson, 2002; Forni *et al*., 2005; Giannone *et al*., 2008; for a wide review, see Eklund and Kapetanios, 2008). Factor analysis combined with linear modelling has been, usually, the main tool used for this objective.

This paper presents a new statistical approach to forecast macro-economic aggregates, based on the Random Forests technique, originally developed as a learning classification tool (Breiman, 2001), which is known to enjoy good prediction properties and to be robust to noise.

While the Random Forests algorithm is usually applied in medical research and biological studies, this paper investigates the potential of applying the same principles to the modelling and the forecast of macro-economic aggregates using large data sets of survey variables, in the same vein of Biau *et al*. (2007).

A specific application to euro area GDP short-term forecast, using the harmonised European Union Business and Consumer Survey data set, is shown. The Random Forests technique is exploited with a twofold objective: the first one is to obtain (trough a Monte Carlo exercise) a preliminary non-parametric forecast of GDP growth and the second one is to distinguish, among lots of candidate explanatory variables, those which significantly contribute to explain and predict the analysed phenomenon and those which mostly add random noise. Indeed, the variable importance index based on Random Forests has the advantage of selecting relevant variables independently from any functional and distributional assumptions, which makes it a robust candidate tool for variables selection. A predictive model is, then, built using as input the selected variables.

The forecast performance of this survey-based model is assessed through an out-of-sample exercise (using vintage data): at this scope, the results are compared with both the outputs from an auto-regressive (AR) model (taken as benchmark), and with the quarterly projections of the *euro zone economic outlook* (jointly released by three main European economic institutes: the German IFO, the French INSEE and the Italian ISAE), which are deemed to be among the most reliable forecasts.

* Olivier Biau, seconded national expert from the INSEE - France, and Angela D'Elia are economic analysts in the Directorate for Economic Studies and Research in the Directorate-General of Economic and Financial Affairs, Brussels. Views expressed represent exclusively the positions of the authors and do not necessarily correspond to those of the European Commission.

Olivier.Biau@ec.europa.eu; Angela.D'Elia@ec.europa.eu

Evidence is found that a well-performing and parsimonious survey-based model can be specified to forecast euro area GDP quarter on quarter growth, and that Random Forests are, therefore, an effective tool in selecting the most relevant predictive variables.

Key Words: Business and Consumer Survey data, GDP short-term forecast, Random Forests, Variables selection

JEL Classification: C8, C51, C53, C63, E3

1. Introduction

Assessing and forecasting the state of the economy is an important task for policy-makers and analysts. Since hard data (e.g. GDP) are published with a considerable delay, policy decisions have to rely on more timely information: this is the case of business tendency survey data, which – due to their early release – are widely used as potential indicators to track the economic activity.

Typically, survey information is scattered through large number of "soft" time series. As a consequence, recent works in the econometric literature consider the problem of summarising efficiently a large set of (both soft and hard) variables and using this summary for a variety of purposes, including forecast (Stock and Watson, 2002; Forni *et al*., 2005; Giannone *et al*., 2008; for a wide review, see Eklund and Kapetanios, 2008). Factor analysis combined with linear modelling has been the main tool used for this objective.

This paper presents a new statistical approach to forecast macro-economic aggregates, based on the Random Forests technique, originally developed as a learning classification tool (Breiman, 2001), which is known to enjoy good prediction properties and to be robust to noise.

While the Random Forests algorithm is usually applied in medical research, biological studies and bioinformatics (Arun and Langmead, 2006; Diaz-Uriarte and Alvarez de Andrés, 2006; Ward *et al*., 2006), this paper investigates the potential of applying the same principles to economic data, in order to model and forecast macro-economic aggregates using large data sets of survey variables. This approach has been followed successfully by Biau *et al*. (2007) in order to forecast the French manufacturing output growth from firm-level survey data.

A specific application to euro area GDP short-term forecast, using the harmonised European Union Business and Consumer survey data set, is shown. The Random Forests technique is exploited with a twofold objective: the first one is to obtain a preliminary non-parametric forecast of GDP growth and the second one is to distinguish, among lots of candidate explanatory variables, those which significantly contribute to explain and predict the analysed phenomenon and those which mostly add random noise. Indeed, the variable importance index (Breiman, 2002) based on Random Forests has the advantage of selecting relevant variables independently from any functional and distributional assumptions, which makes it a robust candidate tool for variables selection. A predictive model is, then, built using as input the selected variables.

The forecast performance of this survey-based model is assessed through an out-of-sample exercise: at this scope, the results are compared with both the outputs from an auto-regressive (AR) model (taken as benchmark), and with the quarterly projections of the *euro zone economic outlook* (jointly released by three main European economic institutes: the German IFO, the French INSEE and the Italian ISAE), which are deemed to be among the most reliable forecasts.

The paper is organised as follows. The Random Forest approach and the related variable importance measure are presented in Section 2, while the dataset used throughout the study is described in Section 3. Section 4 is devoted to obtain first, trough a Monte Carlo exercise, a preliminary non-parametric forecast of GDP growth (§ 4.1); then, to select the most relevant variables to be used in the predictive model (§4.2) and finally (§4.3) to provide an assessment of the empirical performance of the model, using vintage data. Further developments and conclusions end the paper.

2. The random forests approach

Random Forests (RF hereafter) is an efficient algorithm for both high-dimensional classification and regression problems, introduced by Breiman (2001). RF are, indeed, one of the most successful

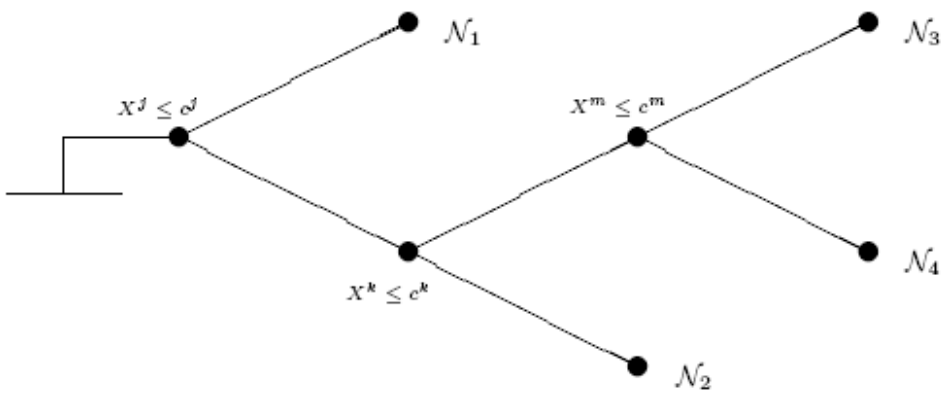
ensemble methods appearing in machine learning (Dietterich, 2000) and are known to enjoy good prediction properties. RF have been shown to give excellent performance on a number of practical problems, even if the mechanism of RF algorithms is difficult to analyze, remains largely unknown and is not clearly elucidated from a mathematical point of view (Breiman, 2002; Lin and Jeon, 2006; Biau *et al.*, 2008, Biau and Lugosi, 2008).

Let us consider a learning set $\boldsymbol{L} = \{(X_1, Y_1),\ldots,(X_n, Y_n)\}$ made of n i.i.d. observations of a random vector $(\boldsymbol{X}, Y)$. Vector $\boldsymbol{X}_i = [X^1,\ldots,X^p]$ contains p predictors or explanatory variables, say $\boldsymbol{X}_i \in \mathbb{R}^p$, and $Y_i \in \mathbb{R}$ is a numerical response. Thus, given a new realisation of $\boldsymbol{X}$, the statistical problem is to predict Y using the learning set $\boldsymbol{L}$. In regression problems, we suppose that $Y = s(\boldsymbol{X}) + \varepsilon$, where s is the so-called regression function. The principle of random forests (§ 2.1) is to combine many binary regression trees, built using several bootstrap samples on $\boldsymbol{L}$, and choosing randomly at each node the subset of explanatory variables $\boldsymbol{X}$.

2.1 From binary trees to random forest

Hastie *et al*. (2001) describe in details how to grow a binary regression tree using a dataset $\boldsymbol{L}$. Briefly, the algorithm CART (Classification and Regression Trees) automatically decides at each node both splitting variable and threshold value. Having found the best split, the dataset is then partitioned into two resulting sub-sets, and the process continues until each node reaches a user-specified minimum *nodesize* and becomes a terminal node (Figure 1).

Figure 1 An example of regression tree



Given a new $\boldsymbol{X}$, the tree regressor h is then defined on each terminal node by the empirical mean:

$$h(\boldsymbol{X}) = \frac{1}{Card\{i/\boldsymbol{X}_i \in N(\boldsymbol{X})\}} \sum_{i/\boldsymbol{X}_i \in N(\boldsymbol{X})} Y_i \tag{1}$$

where $N(\boldsymbol{X})$ stands for the terminal node containing $\boldsymbol{X}$.

The principle of random forest is to grow a large number (K) of regression trees (often many hundred) from different independent subsets of variables. For each tree and each node, RF employ

randomness when selecting a variable to split on: each decision tree is built from a bootstrapped sample of the full dataset (Efron and Tibshirani, 1993) and then, at each node, only a random sample of the available variables is used as candidate variables for split point selection. Thus, instead of determining the optimal split on a given node by evaluating all possible splits on all variables, a subset *mtry* of the input variables are randomly chosen, and the best split is calculated only within this subset (with the value *mtry* being held constant during the growth of the forest).

Having built an ensemble of K trees, the predicted outcome (final decision) is obtained as the average value over the K trees. Thus, denoting by $h_1,\ldots,h_K$ the individual tree predictors, the predicted outcome is:

$$h(\boldsymbol{X}) = \frac{1}{K}\sum_{k=1}^{K} h_k(\boldsymbol{X}). \tag{2}$$

Averaging over trees, in combination with the randomization used in growing a tree, enables random forests to approximate a rich class of functions while maintaining low generalization error. This enables random forests to adapt to the data, automatically fitting higher order interactions and non-linear effects, while at the same time keeping overfitting in check (Ishwaran, 2007). In particular in the regression setting, RF are known to give an accurate approximation of the conditional mean of the response variable (Meinshausen, 2006).This has led to a great interest in the method and applications in many fields.

Over the last years, an associated package[1] *randomForest* (Liaw and Wiener, 2002) has been developed in the free available R software, in order to implement Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code). The choice of the parameters K, *nodesize*, and *mtry*, allows a fine tuning of the algorithm itself (Genuer *et al*., 2008), with default values for regression purposes being set equal to 500, 5 and $p/3$, respectively.

2.3 The variable importance measure

An interesting and useful feature of RF is that, while they are often used for exploratory data analysis (classification and regression), they can also be used to select variables and reduce data dimensionality. This is done by ranking the variables by means of a variable importance measure and removing those variables with low rank. In the regression setting, for example, the variable importance measure for a variable X^j is the normalised difference between the prediction error when X^j is noised up by permuting its value randomly compared to prediction error under the original predictor.

More in detail (Ishwaran, 2007), a given X^j is randomly permuted in the out-of-bag (OOB) data (the data not selected by bootstrapping) and the noised up OOB data is dropped down the tree grown from the in-bag data (bootstrap data). This is done for each tree in the forest and an out-of-bag estimate of prediction error is computed from the resulting predictor. The difference between this value and the out-of-bag error without random permutation is averaged over all trees and normalised by the standard error. This is called the variable importance of X^j. Large positive values for X^j indicate that X^j is predictive (since noising up it increases prediction error), whereas zero or negative importance values identify variables not predictive.

The *randomForest* package in R allows the extraction of variable importance measure, as well as the scree plot of the variables, which are ranked accordingly to it.

[1] http://cran.r-project.org/web/packages/randomForest/index.html

3. The data set

The data set used throughout the paper is built on the Joint Harmonised European Union Business and Consumer Surveys[2]. It is made mainly of the Euro Area balances of opinion (e.g. the percentage difference between positive and negative respondents), which are interesting indicators in many respects: they are easy to implement and to read, they are subject to limited revisions across time and they are highly correlated with the corresponding aggregates of interest (e.g. economic hard variables), even though they are generally smoother. All these interesting properties, together with their timely release, explain why the balances of opinion are among the main indicators used by short-term analysts as explanatory variables in linear models.

The time series (both monthly and quarterly) that have been used in the analysis are those available at the end of the third month (S_t) of each quarter (S_q) for all the surveyed sectors (e.g. Industry, Services, Retail trade, Construction, and Consumers). Besides the level series, also the difference series (S_t - S_{t-1}, S_t - S_{t-2}, S_t - S_{t-3} for monthly questions, S_q - S_{q-1} for quarterly questions) have been taken into account, so that the data set is finally made of 172 soft series, as detailed in Table 1.

Therefore, at the end of each quarter, the dataset includes the most recent "soft" data available. This mimics precisely the operational conditions under which quarterly projections for the euro zone are made by practitioners. This is, indeed, the kind of soft data that are part of the information set used for those projections.

The only hard variable is the euro area GDP qoq growth series (first estimate, released by Eurostat), which is used as dependent variable to be predicted on the basis exclusively of the available soft survey data. The sample covers the period September 1995 – September 2009.

[2] Joint Harmonised EU program of Business and Consumer Surveys (BCS):
http://ec.europa.eu/economy_finance/db_indicators/surveys/method_guides/index_en.htm
http://ec.europa.eu/economy_finance/db_indicators/surveys/index_en.htm

Table 1 Data Set

Survey sector	Questions[a]	Level	Difference		
Industry	Monthly Questions (1 to 7)	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
	Quarterly Questions (9 to 16)	S_q	$S_q - S_{q-1}$		
	Confidence Indicator[b]	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
Services	Monthly Questions (1 to 4)	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
	Confidence Indicator[b]	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
Retail trade	Monthly Questions (1 to 5)	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
	Confidence Indicator[b]	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
Construction	Monthly Questions (1 to 5)	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
	Quarterly Questions (6)	S_q	$S_q - S_{q-1}$		
	Confidence Indicator[b]	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
Consumers	Monthly Questions (1 to 12)	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$
	Quarterly Questions (13 to 15)	S_q	$S_q - S_{q-1}$		
	Confidence Indicator[b]	S_t	$S_t - S_{t-1}$	$S_t - S_{t-2}$	$S_t - S_{t-3}$

a) The detailed list of the questions can be found in the Appendix.

b) Confidence indicators are computed as the arithmetic average of the balances of the answers to the questions: 2, 4 (with inverted sign) and 5 (Industry); 1, 2 and 3 (Services); 1, 2 (with inverted sign) and 4 (Retail trade); 3 and 4 (Construction); 2, 4, 7 (with inverted sign) and 11 (Consumers).

4. Nowcasting GDP growth through random forests

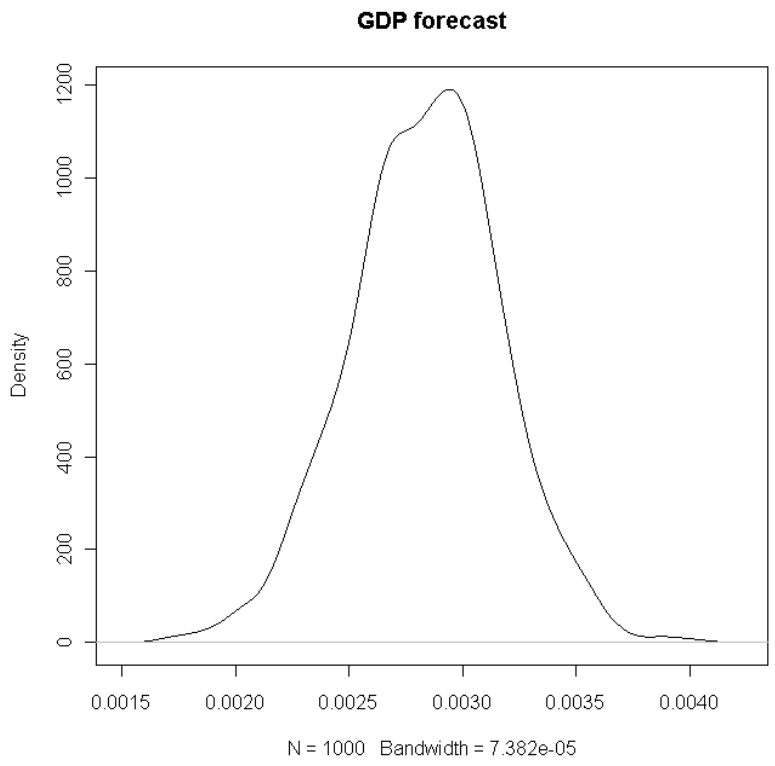
The RF approach, as described in Section 2, can be usefully exploited in order to predict GDP qoq growth in the euro area by means of the survey data. This is, indeed, a typical high-dimensional regression problem ($n << p$), as the data set is made of p = 172 possible candidate explanatory variables (time series), each consisting of n = 57 observations.

More in detail, two different avenues can be followed. First, the RF algorithm can be used in itself to obtain a non-parametric estimate of the GDP growth (§ 4.1). Secondly, one can use the variable importance measure to obtain a ranking of the explanatory variables, and then to select those variables on which to build a linear model to forecast GDP growth (§ 4.2 and § 4.3).

4.1 Non-parametric estimation of GDP growth

A Monte Carlo exercise has been set up, running 1000 replicates of random forests, each grown on K=500 trees. This yield, for example, an estimated value for euro area GDP qoq growth in 2009Q3 equal to +0.3% (Figure 2), which compares well to the value effectively observed (+0.38%, first estimate released by Eurostat on 3rd December, 2009).

Figure 2 Monte Carlo kernel density of GDP forecast for 2009Q3



Source: Our computation on European Commission and Euro Area Business Cycle Network data

For the out-of-sample analysis the sub-sample 2004Q1 – 2009Q3 has been selected. The values predicted by means of the RF have then been compared to GDP vintage data (e.g. historical released data), which are available through the Euro Area Business Cycle Network[3]. As benchmark values the output from a univariate auto-regressive (AR) model has been taken.

As the AR appears to be a poor competitor, the forecasts obtained through the RF have been then compared to the quarterly projections of the *euro zone economic outlook*[4] (jointly released by three main European economic institutes: the German IFO, the French INSEE and the Italian ISAE), which are deemed to be among the most reliable forecasts for the euro area. According to the methodological box of this publication, "the forecasts are built up with the help of different forecasting tools shared by the three institutes, using time series models based on business surveys by national institutes, Eurostat and the European Commission." In fact this publication provides 3 steps ahead projections for GDP, Industrial Production, Consumption and Inflation and describes the economic links between these main aggregates: here, however, our interest is just to assess how a data driven model like the RF model performs relatively to a fair competitor (the *euro zone economic outlook*) for GDP 1 step ahead forecasting.

The main results are shown in Table 2, where the forecast accuracy (in terms of Mean Square Errors - MSE) of the 3 models (AR, RF, *euro zone economic outlook*) is compared. Table 2 shows

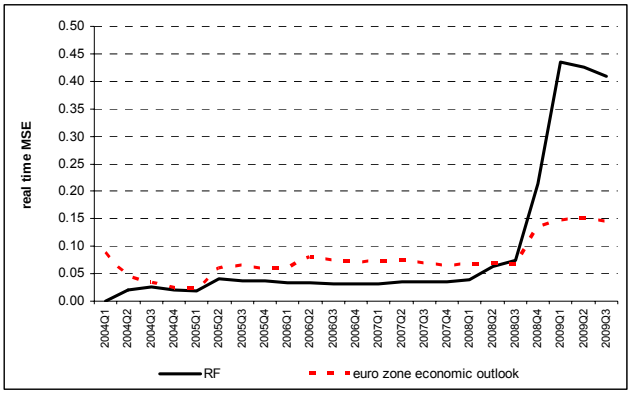
[3] http://www.eabcn.org

[4] http://www.cesifo-group.de/portal/page/portal/ifoHome/a-winfo/d2kprog/30kprogeeo/_KPROGEEOlist

that the non-parametric forecasting approach, based on random forests (RF), outperforms the univariate AR model in predicting GDP growth for the euro area, but not the *economic outlook*. However, in terms of real time MSE, it seems to perform better than the *euro zone economic outlook* until the second/third quarter of 2008. This is a noteworthy result, as the RF projections emanate from a full non-parametric tool that has as inputs exclusively soft variables.

Table 2 - Forecast accuracy

MSE	
AR	0.64
RF	0.43
euro zone economic outlook	0.15



Note: MSEs are computed on the whole out-of-sample period in the table, while the chart shows how these values develop over time (real time MSEs)

Source: Our computation on European Commission, Euro Area Business Cycle Network and IFO-INSEE-ISAE data

The performance of RF in predicting GDP growth has been far less satisfactory in 2008Q4 and in the first quarter of 2009 (see Table 1a in the Appendix). In fact, these quarters have recorded extremely negative values of GDP growth (-1.8% in 2008Q4, -2.5% in 2009Q1), which had never been observed previously in the learning set (sample) ***L***: as the RF predictor (2) is, by construction, a weighted average over *K* trees built on the learning set, it can not take negative values that are not present in ***L***. However, it is worth noticing that once negative values are observed and enter the learning set, the RF algorithm is set to learn from them and, then, to progressively adapt the predicted outcome (as for 2009Q2).

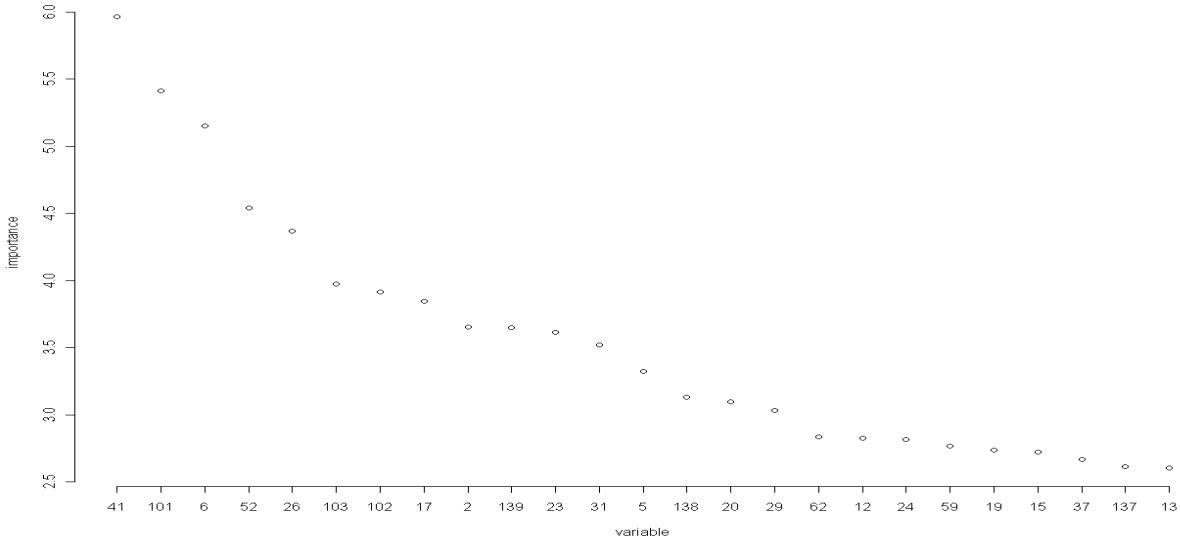
4.2 Selecting relevant variables through RF

To square the issue of the never observed recession in the learning set, a different approach in forecasting GDP growth through RF is to exploit the variable importance measure (described in section 2.1) in order to select among a large number of possible candidate explanatory variables those that are expected to be good predictors, and then to plunge them in a linear model (e.g. a bridge model). It is worth noticing that the variable selection step is also necessary as the use of standard linear regression models is not possible in situations where the number of possible predictors is large compared to the number of observations.

This two-steps strategy combines the effectiveness of RF as a powerful tool for identifying relevant variables in large survey data sets, with the known advantages of bridge models for short-term forecast of GDP (see e.g. Baffigi *et al.*, 2004).

Following this strategy, first a ranking of the 25 most predictive survey variables has been obtained (Figure 3; the variables codification is given in Table 2a in the Appendix). The ranking has been obtained by averaging the variable importance measures over 1000 Monte Carlo replicates.

Figure 3 Variable importance measure plot



Source: Our computation on European Commission and Euro Area Business Cycle Network data

Then, the 25 selected variables have been inserted as candidate explanatory variables in the specification of a linear bridge model to forecast euro area GDP growth qoq. Starting from this general specification, the Gets (GEneral-To-Specific) procedure has been used to reduce its complexity by eliminating statistically insignificant variables and to ensure the congruency of the model (Krolzig and Hendry, 2001). The Gets procedure has been implemented through the econometric freely available software Grocer (Dubois and Michaux, 2008).

4.3 The model and its performance

Using the above described strategy, finally the retained model[5] to forecast euro area GDP growth qoq includes five explanatory variables besides the constant term (Table 3).

[5] The retained model has successfully passed the standard battery of misspecification tests.

Table 3 Estimated model

Dependent Variable: GDP growth
Method: Ordinary Least Squares
Estimation period: 1995Q3 2009Q2
Included observations: 56

Code	Variable	Coefficient	T-statistics	P- value
	Constant	0.0061507	12.03	0.000000
12	INDU Q11 *Orders development over past 3 months*	0.0001111	2.68	0.0098621
24	CONS Q2 *Expectation about household financial positions over next 12 months*	0.0003239	3.25	0.0020789
41	RETA Q3 *Orders development expected over next 3 months*	0.0002008	3.15	0.0027377
62	INDU Q12 (t – (t-1)) *Export orders development expected over next 3 months*	0.0004439	4.15	0.0001278
101	INDU Q2 (t – (t-2)) *Assessment of current order book*	0.0002538	1.93	0.0588848

R-squared = 0.8583952	Adjusted R-squared =0.844234
S.E. of regression = 0.0024363	Sum of squared residuals = 0.0002968
Overall F test: F(5,50) = 60.6119091	Durbin-Watson stat =2.1843682

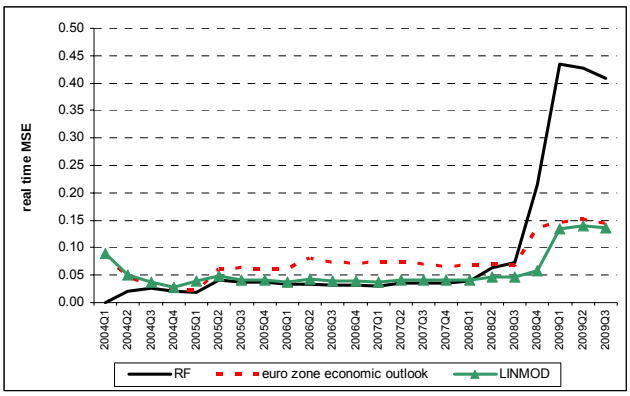
Source: Our computation on European Commission and Euro Area Business Cycle Network data

Remarkably, four out of five explanatory variables are related to the orders level and dynamics (past and expected), both in Industry and in Retail trade: this is unsurprising, as the assessments about order books are supposed to be among the most factual soft variables, reflecting the true stance of economic activity in which survey respondents are involved.

The predictive performance of the linear model (LINMOD) has, then, been assessed through an out-of-sample analysis (sub-sample 2004Q1 – 2009Q3). This means that for each point in time the parameters of the model and the forecasts have been estimated using the data that replicate the pattern of data availability at that time. The outcomes are compared to both the RF projections and to those of the *euro zone economic outlook* (Table 4 and Figure 4; more details are given in Table 1a in the Appendix).

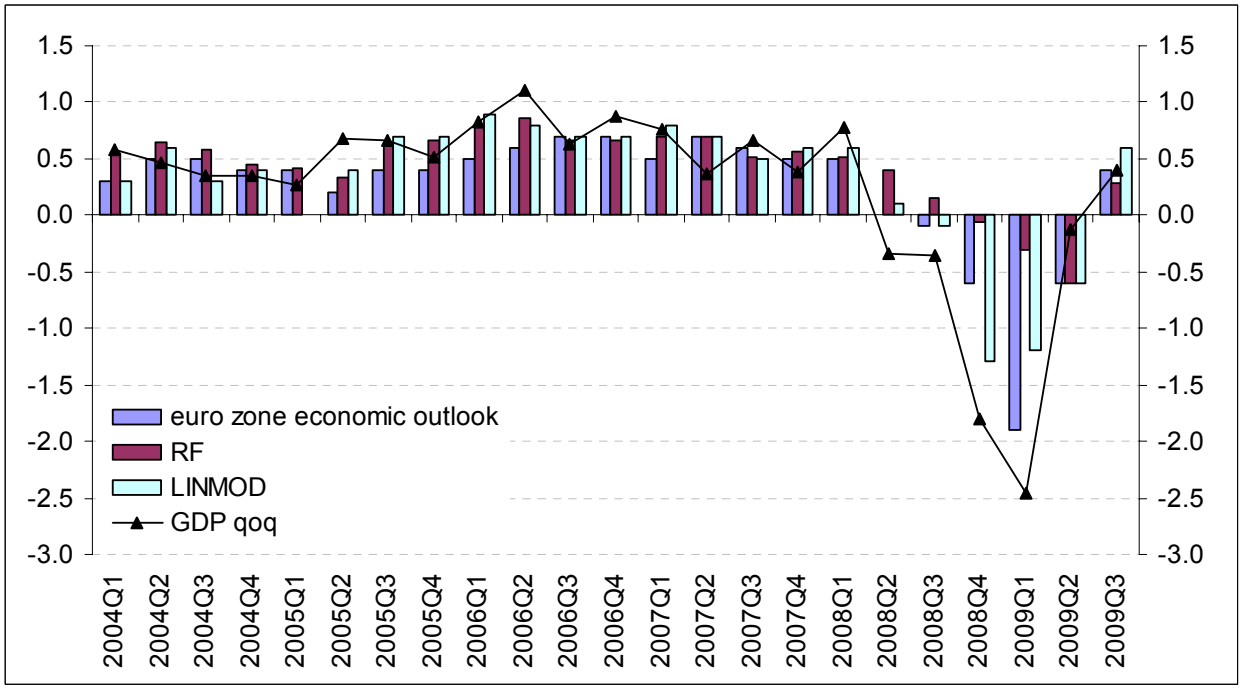
Table 4 – Forecast accuracy

MSE	
RF	0.43
euro zone economic outlook	0.15
LINMOD	0.14



Note: MSEs are computed on the whole out-of-sample period in the table, while the chart shows how these values develop over time (real time MSEs)

Figure 4 GDP growth qoq forecasts



Source: Our computation on European Commission, Euro Area Business Cycle Network and IFO-INSEE-ISAE data

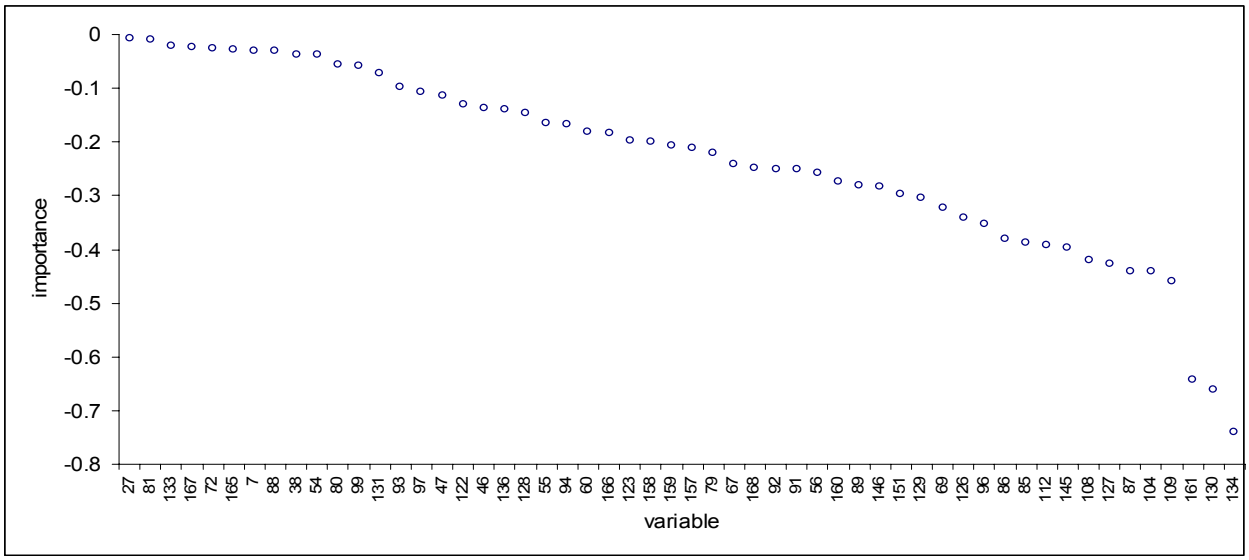
Table 4 and Figure 4 emphasize the good results achieved by using the RF algorithms to select the variables to be inserted in the preliminary specification of a linear model. We note, in particular, that:

- on the whole out-of-sample period, LINMOD performs equally well as the *euro zone economic outlook*[6],
- during the "crisis" quarters (e.g. 2008Q3 – 2009Q2), LINMOD always outperforms the *euro zone economic outlook* (and the pure RF itself).

The observed difference between the predictive performance of the pure RF approach and the combined one (LINMOD) is mainly due (as discussed in §4.1) to the structure of the learning set ***L*** where no negative values were present, at least until 2008Q2, and it is likely to gradually diminish with the lengthening of ***L***, which now encompasses also negative values. On the other hand, this difference enlightens the importance of the variable selection step and the advantages in using this RF's feature to properly specify linear predictive models.

Moreover, an interesting by-product of the RF algorithm is the possibility to identify those variables that have a negative impact (e.g. a negative value of the variable importance measure) on the predictive performance of a model. Those variables just add noise and should, therefore, be discarded *a priori* from any following model specification. An example for the dataset used throughout the paper is given in Figure 5 (for the codification of the variables, see Table 2a in the Appendix).

Figure 5 Variable importance measure plot (negative values)



Source: Our computation on European Commission and Euro Area Business Cycle Network data

5. Conclusions and further developments

A new approach (based on Random Forests technique) has been presented in order to short term forecast GDP growth in the euro area. It can be followed through two different avenues: pure

[6] The null hypothesis of the Harvey *et al*. (1997) tests of equal accuracy in forecast performance can not be rejected.

non-parametric RF, or RF combined with linear model. Using GDP vintage data, the comparative predictive performance of both strategies has been discussed and compared to the AR model output and to the *euro zone economic outlook* projections. In particular, the combined approach outperforms the AR benchmark and behaves equally well as the *euro zone economic outlook*: it is, therefore, a good candidate tool for short-term analysts, especially in situations where the large number of predictors makes the use of standard linear regression models not possible.

Besides that, it is also worthy to notice that the RF algorithm works very fast (using the R-package "RandomForest", prediction and variables selection take a few seconds). This makes possible to get the forecast of the aggregate of interest (e.g. the GDP) as soon as real time survey data are available.

For further developments, two different avenues can be followed. First, the soft dataset –used as input for the RF analysis- could be widened by adding those hard variable that are available at the end of each quarter (e.g. carry-over of industrial production and first registration of private and commercial cars). Secondly, one could investigate alternative state-of-the art forecasting methodologies for large datasets (see Eklund and Kapetanios, 2008, for a non-technical overview). Among the different variable selection methods that reduce the dimensionality of the original dataset, a potential candidate could be the LASSO technique (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996).

References

Arun K. and Langmead C.J. (2006), Structure based chemical shift prediction using random forests nonlinear regression, *4th Asia-Pacific Bioinformatics Conference*, Taiwan, 317–326.

Baffigi A., Golinelli R. and Parigi G. (2004), Bridge models to forecast the euro area GDP, *International Journal of Forecasting*, 20, 447-460.

Biau G., Biau O. and Rouvière L. (2007), Nonparametric forecasting of the manufacturing output growth with firm-level survey data, *Journal of Business Cycle Measurement and Analysis*, 3, 317-332.

Biau G., Devroye, L. and Lugosi, G. (2008), Consistency of random forests and other averaging classifiers, *Journal of Machine Learning Research*, 9, 2015-2033.

Biau G. and Devroye, L. (2008). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, University of Paris 6 Technical Report.

Breiman L. (2001), Random forests, *Machine Learning*, Kluwer Academic Publishers, 45, 5-32.

Breiman L. (2002), *Manual on setting up, using, and understanding Random Forests v3.1*, Technical Report, http://oz.berkeley.edu/users/breiman.

Díaz-Uriarte R. and Alvarez de Andrés S. (2006), Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 1-13

Dietterich, T.G. (2000), Ensemble methods in machine learning, *Lecture Notes in Computer Science,* Springer-Verlag, 1-15.

Dubois E and Michaux E. (2008) Grocer: an econometric toolbox for Scilab, http://dubois.ensae.net/grocer.html

Eklund J., Kapetanios G. (2008) A review of forecasting techniques for large datasets, *National Institute Economic Review*, 203, 109-115.

Forni M., Hallin M., Lippi M. and Reichlin L. (2005), The Generalized Dynamic Factor Model: one-sided estimation and forecasting, *Journal of the American Statistical Association*, 100, 830-840.

Genuer R, Poggi J.M. and Tuleau C. (2008), Random Forests: some methodological insights, *Rapport de recherche*, 6729, Institut national de recherche en informatique et en automatique, France.

Giannone D., Reichlin L. and Small D. (2008), Nowcasting: The Real Time Informational Content of Macroeconomic Data Releases, *Journal of Monetary Economics*, 55, 665-676.

Harvey D.I., Leybourne S.J and Newbold P. (1997), Testing the equality of prediction mean square errors, *International Journal of Forecasting*, 13, 273-281.

Hastie T., Tibshirani R.J. and Friedman J. (2001), *The Elements of Statistical Learning*, Springer-Verlag.

Ishwaran H. (2007), Variable importance in binary regression trees and forests, *Electronic Journal of Statistics*, 1, 519–537.

Krolzig H.M. and Hendry D. (2001) Computer automation of general-to-specific model selection procedures, *Journal of Economic Dynamics and Control*, 25, 831-836.

Liaw A. and Wiener M. (2002), Classification and regression by randomForest. *R News*, 2, 18–22.

Lin Y. and Jeon Y. (2006), Random forests and adaptive nearest neighbours, *Journal of the American Statistical Association*, 101, 578–590.

Meinshausen N. (2006), Quantile Regression Forests, *Journal of Machine Learning Research*, 7, 983–999.

Stock J.H., and Watson M.W. (2002), Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, 97, 1167–1179.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso, J*ournal of the Royal Statistical Society B*, 58, 267-288.

Ward M.M., Pajevic S, Dreyfuss J. and Malley J.D. (2006), Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests, *Arthritis and Rheumatism*, 55, 74–80.

Appendix

Table 1a Out-of-sample GDP forecast

	FORECAST			GDP qoq	ERROR			SUM ERROR²		
	ifo insee isae	RF	LINMOD	observed	ifo insee isae	RF	LINMOD	ifo insee isae	RF	LINMOD
2004Q1	0.3	0.6	0.3	0.6	-0.3	0.0	-0.3	0.1	0.0	0.1
2004Q2	0.5	0.7	0.6	0.5	0.0	0.2	0.1	0.1	0.0	0.1
2004Q3	0.5	0.6	0.3	0.4	0.1	0.2	-0.1	0.1	0.1	0.1
2004Q4	0.4	0.4	0.4	0.4	0.0	0.0	0.0	0.1	0.1	0.1
2005Q1	0.4	0.4	0.0	0.3	0.1	0.1	-0.3	0.1	0.1	0.2
2005Q2	0.2	0.3	0.4	0.7	-0.5	-0.4	-0.3	0.4	0.3	0.3
2005Q3	0.4	0.6	0.7	0.7	-0.3	-0.1	0.0	0.5	0.3	0.3
2005Q4	0.4	0.7	0.7	0.5	-0.1	0.2	0.2	0.5	0.3	0.3
2006Q1	0.5	0.8	0.9	0.8	-0.3	0.0	0.1	0.6	0.3	0.3
2006Q2	0.6	0.9	0.8	1.1	-0.5	-0.2	-0.3	0.8	0.3	0.4
2006Q3	0.7	0.7	0.7	0.6	0.1	0.1	0.1	0.8	0.4	0.4
2006Q4	0.7	0.7	0.7	0.9	-0.2	-0.2	-0.2	0.9	0.4	0.5
2007Q1	0.5	0.7	0.8	0.8	-0.3	-0.1	0.0	0.9	0.4	0.5
2007Q2	0.7	0.7	0.7	0.4	0.3	0.3	0.3	1.0	0.5	0.6
2007Q3	0.6	0.5	0.5	0.7	-0.1	-0.2	-0.2	1.0	0.5	0.6
2007Q4	0.5	0.6	0.6	0.4	0.1	0.2	0.2	1.1	0.6	0.7
2008Q1	0.5	0.5	0.6	0.8	-0.3	-0.3	-0.2	1.1	0.7	0.7
2008Q2	0.0	0.4	0.1	-0.3	0.3	0.7	0.4	1.2	1.2	0.9
2008Q3	-0.1	0.2	-0.1	-0.3	0.2	0.5	0.2	1.3	1.4	0.9
2008Q4	-0.6	-0.1	-1.3	-1.8	1.2	1.7	0.5	2.7	4.3	1.1
2009Q1	-1.9	-0.3	-1.2	-2.5	0.6	2.2	1.3	3.1	9.1	2.8
2009Q2	-0.6	-0.6	-0.6	-0.1	-0.5	-0.5	-0.5	3.3	9.4	3.1
2009Q3	0.4	0.3	0.6	0.4	0.0	-0.1	0.2	3.3	9.4	3.1

Source: Our computation on European Commission, Euro Area Business Cycle Network and IFO-INSEE-ISAE data

Table 2a Codification of BCS variables used as input for random forests

2	INDU	Q1	*LEVEL: monthly St; quarterly Sq*		**51**	INDU	Q1	*DIFF monthly: St - S(t-1)* *DIFF quarterly: Sq - S(q-1)*
3	INDU	Q2			**52**	INDU	Q2	
4	INDU	Q3			**53**	INDU	Q3	
5	INDU	Q4			**54**	INDU	Q4	
6	INDU	Q5			**55**	INDU	Q5	
7	INDU	Q6			**56**	INDU	Q6	
8	INDU	Q7			**57**	INDU	Q7	
9	INDU	COF[a]			**58**	INDU	COF	
10	INDU	Q9			**59**	INDU	Q9	
11	INDU	Q10			**60**	INDU	Q10	
12	INDU	Q11			**61**	INDU	Q11	
13	INDU	Q12			**62**	INDU	Q12	
14	INDU	Q13			**63**	INDU	Q13	
15	INDU	Q14			**64**	INDU	Q14	
16	INDU	Q15			**65**	INDU	Q15	
17	INDU	Q16			**66**	INDU	Q16	
18	SERV	Q1			**67**	SERV	Q1	
19	SERV	Q2			**68**	SERV	Q2	
20	SERV	Q3			**69**	SERV	Q3	
21	SERV	Q4			**70**	SERV	Q4	
22	SERV	COF			**71**	SERV	COF	
23	CONS	Q1			**72**	CONS	Q1	
24	CONS	Q2			**73**	CONS	Q2	
25	CONS	Q3			**74**	CONS	Q3	
26	CONS	Q4			**75**	CONS	Q4	
27	CONS	Q5			**76**	CONS	Q5	
28	CONS	Q6			**77**	CONS	Q6	
29	CONS	Q7			**78**	CONS	Q7	
30	CONS	Q8			**79**	CONS	Q8	
31	CONS	Q9			**80**	CONS	Q9	
32	CONS	Q10			**81**	CONS	Q10	
33	CONS	Q11			**82**	CONS	Q11	
34	CONS	Q12			**83**	CONS	Q12	
35	CONS	COF			**84**	CONS	COF	
36	CONS	Q13			**85**	CONS	Q13	
37	CONS	Q14			**86**	CONS	Q14	
38	CONS	Q15			**87**	CONS	Q15	
39	RETA	Q1			**88**	RETA	Q1	
40	RETA	Q2			**89**	RETA	Q2	
41	RETA	Q3			**90**	RETA	Q3	
42	RETA	Q4			**91**	RETA	Q4	
43	RETA	Q5			**92**	RETA	Q5	
44	RETA	COF			**93**	RETA	COF	
45	BUIL	Q1			**94**	BUIL	Q1	
46	BUIL	Q3			**95**	BUIL	Q3	
47	BUIL	Q4			**96**	BUIL	Q4	
48	BUIL	Q5			**97**	BUIL	Q5	
49	BUIL	COF			**98**	BUIL	COF	
50	BUIL	Q6			**99**	BUIL	Q6	

a) COF stands for Confidence Indicator

Table 2a (ctd) Codification of BCS variables used as input for random forests

100	INDU	Q1	*DIFF monthly: St - S(t-2)*		**137**	INDU	Q1	*DIFF monthly: St - S(t-3)*
101	INDU	Q2			**138**	INDU	Q2	
102	INDU	Q3			**139**	INDU	Q3	
103	INDU	Q4			**140**	INDU	Q4	
104	INDU	Q5			**141**	INDU	Q5	
105	INDU	Q6			**142**	INDU	Q6	
106	INDU	Q7			**143**	INDU	Q7	
107	INDU	COF[a]			**144**	INDU	COF	
108	SERV	Q1			**145**	SERV	Q1	
109	SERV	Q2			**146**	SERV	Q2	
110	SERV	Q3			**147**	SERV	Q3	
111	SERV	Q4			**148**	SERV	Q4	
112	SERV	COF			**149**	SERV	COF	
113	CONS	Q1			**150**	CONS	Q1	
114	CONS	Q2			**151**	CONS	Q2	
115	CONS	Q3			**152**	CONS	Q3	
116	CONS	Q4			**153**	CONS	Q4	
117	CONS	Q5			**154**	CONS	Q5	
118	CONS	Q6			**155**	CONS	Q6	
119	CONS	Q7			**156**	CONS	Q7	
120	CONS	Q8			**157**	CONS	Q8	
121	CONS	Q9			**158**	CONS	Q9	
122	CONS	Q10			**159**	CONS	Q10	
123	CONS	Q11			**160**	CONS	Q11	
124	CONS	Q12			**161**	CONS	Q12	
125	CONS	COF			**162**	CONS	COF	
126	RETA	Q1			**163**	RETA	Q1	
127	RETA	Q2			**164**	RETA	Q2	
128	RETA	Q3			**165**	RETA	Q3	
129	RETA	Q4			**166**	RETA	Q4	
130	RETA	Q5			**167**	RETA	Q5	
131	RETA	COF			**168**	RETA	COF	
132	BUIL	Q1			**169**	BUIL	Q1	
133	BUIL	Q3			**170**	BUIL	Q3	
134	BUIL	Q4			**171**	BUIL	Q4	
135	BUIL	Q5			**172**	BUIL	Q5	
136	BUIL	COF			**173**	BUIL	COF	

a) COF stands for Confidence Indicator

Industry survey - Questionnaire

Monthly questions

Q1 How has your production developed over the past 3 months? It has...
+ increased
= remained unchanged
− decreased

Q2 Do you consider your current overall order books to be...?
+ more than sufficient (above normal)
= sufficient (normal for the season)
− not sufficient (below normal)

Q3 Do you consider your current export order books to be...?
+ more than sufficient (above normal)
= sufficient (normal for the season)
− not sufficient (below normal)

Q4 Do you consider your current stock of finished products to be...?
+ too large (above normal)
= adequate (normal for the season)
− too small (below normal)

Q5 How do you expect your production to develop over the next 3 months? It will...
+ increase
= remain unchanged
− decrease

Q6 How do you expect your selling prices to change over the next 3 months? They will...
+ increase
= remain unchanged
− decrease

Q7 How do you expect your firm's total employment to change over the next 3 months? It will...
+ increase
= remain unchanged
− decrease

Quarterly questions

Q9 Considering your current order books and the expected change in demand over the coming months, how do you assess your current production capacity? The current production capacity is….
+ more than sufficient
= sufficient
− not sufficient

Q10 How many months of production are assured by your current overall order books?
Our production is assured for … months

Q11 How have your orders developed over the past 3 months? They have...
+ increased
= remained unchanged
− decreased

Q12 How do you expect your export orders to develop over the next 3 months? They will...
+ increase
= remain unchanged
− decrease

Q13 At what capacity is your company currently operating (as a percentage of full capacity)?
The company is currently operating at … % of full capacity.

Q14 How has your competitive position on the domestic market developed over the past 3 months? It has...
+ improved
= remained unchanged
− deteriorated

Q15 How has your competitive position on foreign markets inside the EU developed over the past 3 months? It has...
+ improved
= remained unchanged
− deteriorated

Q16 How has your competitive position on foreign markets outside the EU developed over the past 3 months? It has...
+ improved
= remained unchanged
− deteriorated

Services survey - Questionnaire

Monthly questions

Q1 How has your business situation developed over the past 3 months? It has…
+ improved
= remained unchanged
− deteriorated

Q2 How has demand (turnover) for your company's services changed over the past 3 months? It has...
+ increased
= remained unchanged
− decreased

Q3 How do you expect the demand (turnover) for your company's services to change over the next 3 months? It will…
+ increase
= remain unchanged
− decrease

Q4 How has your firm's total employment changed over the past 3 months? It has...
+ increased
= remained unchanged
− decreased

Retail trade survey - Questionnaire

Monthly questions

Q1 How has (have) your business activity (sales) developed over the past 3 months? It has… (They have...)

+ improved (increased)
= remained unchanged
− deteriorated (decreased)

Q2 Do you consider the volume of stock you currently hold to be...?

+ too large (above normal)
= adequate (normal for the season)
− too small (below normal)

Q3 How do you expect your orders placed with suppliers to change over the next 3 months? They will...

+ increase
= remain unchanged
− decrease

Q4 How do you expect your business activity (sales) to change over the next 3 months? It (They) will...

+ improve (increase)
= remain unchanged
− deteriorate (decrease)

Q5 How do you expect your firm's total employment to change over the next 3 months? It will...

+ increase
= remain unchanged
− decrease

Construction survey - Questionnaire

Monthly questions

Q1 How has your building activity developed over the past 3 months? It has...

+ increased
= remain unchanged
− decreased

Q3 Do you consider your current overall order books to be...?

+ more than sufficient (above normal)
= sufficient (normal for the season)
− not sufficient (below normal)

Q4 How do you expect your firm's total employment to change over the next 3 months? It will...

+ increase
= remain unchanged
− decrease

Q5 How do you expect the prices you charge to change over the next 3 months? They will...

+ increase
= remain unchanged
− decrease

Quarterly question

Q6 Assuming normal working hours, about how many months' work is accounted for by the work in hand and the work already contracted for?

Number of months: …

Consumer survey - Questionnaire

Monthly questions

Q1 How has the financial situation of your household changed over the last 12 months? It has...

+ + got a lot better
+ got a little better
= stayed the same
− got a little worse
− − got a lot worse
N don't know.

Q2 How do you expect the financial position of your household to change over the next 12 months? It will...

+ + get a lot better
+ get a little better
= stay the same
− get a little worse
− − get a lot worse
N don't know.

Q3 How do you think the general economic situation in the country has changed over the past 12 months? It has...

+ + got a lot better
+ got a little better
= stayed the same
− got a little worse
− − got a lot worse
N don't know.

Q4 How do you expect the general economic situation in this country to develop over the next 12 months? It will...

+ + get a lot better
+ get a little better
= stay the same
− get a little worse
− − get a lot worse
N don't know.

Q5 How do you think that consumer prices have developed over the last 12 months? They have…

+ + risen a lot
+ risen moderately
= risen slightly
− stayed about the same
− − fallen
N don't know.

Q6 By comparison with the past 12 months, how do you expect that consumer prices will develop in the next 12 months? They will…

+ + increase more rapidly
+ increase at the same rate
= increase at a slower rate
− stay about the same
− − fall

N don't know.

Q7 How do you expect the number of people unemployed in this country to change over the next 12 months? The number will...
+ + increase sharply
+ increase slightly
= remain the same
− fall slightly
− − fall sharply
N don't know.

Q8 In view of the general economic situation, do you think that now it is the right moment for people to make major purchases such as furniture, electrical/electronic devices, etc.?
+ + yes, it is the right moment now
= it is neither the right moment nor the wrong moment
− − no, it is not the right moment now
N don't know.

Q9 Compared to the past 12 months, do you expect to spend more or less money on major purchases (furniture, electrical/electronic devices, etc.) over the next 12 months? I will spend…
+ + much more
+ a little more
= about the same
− a little less
− − much less
N don't know.

Q10 In view of the general economic situation, do you think that now is...?
+ + a very good moment to save
+ a fairly good moment to save
− not a good moment to save
− − a very bad moment to save
N don't know.

Q11 Over the next 12 months, how likely is it that you save any money?
+ + very likely
+ fairly likely
− not likely
− − not at all likely
N don't know.

Q12 Which of these statements best describes the current financial situation of your household?
+ + we are saving a lot
+ we are saving a little
= we are just managing to make ends meet on our income
− we are having to draw on our savings
− − we are running into debt
N don't know.

Quarterly questions

Q13 How likely are you to buy a car over the next 12 months?
+ + very likely
+ fairly likely
− not likely
− − not at all likely
N don't know.

Q14 Are you planning to buy or build a home over the next 12 months (to live in yourself, for a member of your family, as a holiday home, to let etc.)?
+ + yes, definitely
+ possibly
− probably not
− − no
N don't know.

Q15 How likely are you to spend any large sums of money on home improvements or renovations over the next 12 months?
+ + very likely
+ fairly likely
− not likely
− − not at all likely
N don't know.