Direction des Études et Synthèses Économiques

G 2006 / 06

Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

Gérard BIAU, Olivier BIAU
et Laurent ROUVIERE

Document de travail

INSEE

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail
de la Direction des Etudes et Synthèses Économiques

G 2006 / 06

Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

Gérard BIAU*, Olivier BIAU**
et Laurent ROUVIERE***

AOUT 2006

Texte préparé pour la 28ème conférence du CIRET du 20 au 23 septembre 2006, à Rome.

Les auteurs remercient Matthieu Cornec pour sa discussion d'une version antérieure de cette étude lors du séminaire du Département des Études Économiques d'Ensemble du 26 juin 2006.

---

* Institut de Mathématiques et de Modélisation de Montpellier UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II CC 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5,

** Département de la Conjoncture - Division « Enquêtes de Conjoncture » - Timbre G120 - 15, bd Gabriel Péri - BP 100 - 92244 MALAKOFF Cedex

*** Laboratoire de Statistique, Université Rennes 2-Haute Bretagne- Campus Villejean, Place du recteur H. Le Moal, CS 24307, 35043 Rennes Cedex

Département des Etudes Economiques d'Ensemble - Timbre G201 - 15, bd Gabriel Péri - BP 100 - 92244 MALAKOFF CEDEX - France - Tél. : 33 (1) 41 17 60 68 - Fax : 33 (1) 41 17 60 45 - CEDEX - E-mail : d3e-dg@insee.fr - Site Web INSEE : http://www.insee.fr

Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

Abstract

A large majority of summary indicators derived from the individual responses to qualitative Business Tendency Survey questions (which are mostly three-modality questions) result from standard aggregation and quantification methods. This is typically the case for the indicators called balances of opinion, which are the most currently used in short term analysis and considered by forecasters as explanatory variables in linear models. In the present paper, we discuss a new statistical approach to forecast the manufacturing growth from firm-survey responses. We base our predictions on nonparametric forecasting algorithms inspired by statistical pattern recognition, such as the k- nearest neighbors and random forest regression methods, which are known to enjoy good generalization properties. Our algorithms exploit the heterogeneity of the survey responses, work fast, and allow the treatment of missing values. Starting from a real application on a French data set related to the manufacturing sector, we argue that these procedures lead to significantly better results than more traditional competing methods.

Keywords: Business Tendency Surveys, balance of opinion, short-term forecasting, manufactured production, k-nearest neighbour regression, random forests

---

Prévisions non paramétriques de la production manufacturière à partir des réponses individuelles aux enquêtes de conjoncture

Résumé

La majorité des indicateurs élaborés à partir des réponses individuelles aux questions qualitatives des enquêtes de conjoncture résultent de méthodes de quantification et d'agrégation standards. C'est le cas des soldes d'opinion, qui sont les indicateurs les plus couramment utilisés par les conjoncturistes, notamment comme variables explicatives dans des modèles linéaires de prévisions. Dans cette étude, nous présentons une nouvelle approche statistique permettant de prévoir le taux de croissance de la production manufacturière à partir des réponses individuelles des chefs d'entreprise à l'enquête de conjoncture dans l'Industrie. Notre approche est basée sur des techniques non paramétriques d'apprentissage statistique, de type k-plus proches voisins et forêts d'arbres de Breiman. Nos algorithmes sont faciles à mettre en œuvre, rapides et permettent en outre de traiter la non-réponse. Une application sur un jeu de données réelles françaises montre la supériorité des performances de ces algorithmes par rapport aux méthodes plus traditionnelles, basées sur les étalonnages du taux de croissance de la production sur les soldes d'opinion.

Mots-clés : enquêtes de conjoncture, solde d'opinion, prévision conjoncturelle, k-plus proches voisins, forêts aléatoires

Classification JEL : C8, C42, E23, E37, C14

1 Introduction

Due to their early release (by the end of the month in which they are conducted), Business Tendency Surveys (BTS) are widely used as potential indicators of the economic activity, ahead of the publication of data from quarterly national accounts. In particular, BTS results allow the elaboration of short-term forecasting models of the main aggregates of the national accounts on the basis of summary indicators derived from the surveyed responses.

Most BTS questions are qualitative and require either a positive response ("up" or "superior to average"), an intermediate one ("stable" or "close to average") or a negative one ("down" or " "inferior to average"). A large majority of summary indicators derived from the individual responses to these questions result from standard quantification methods, mostly based on a combination of the percentage of positive, stable and negative answers. This is typically the case with the so-called balance of opinion, which is the most currently used indicator for short-term analysis, and which is defined as the difference between the (generally weighted) proportion of positive responses with respect to the negative ones.

As such, these kinds of indicators encounter some criticism, essentially because they do not exploit the heterogeneity of the surveyed individual responses. In this respect, Mitchell, Smith and Weale (2004) discuss alternative indicators of the economic activity, by relating firm categorical responses to official data via ordered discrete-choice models. Their applications to British and German survey data suggest that their indicators provide more accurate early estimates of manufacturing output growth than a set of classical aggregate indicators. However, on French data, Biau, Erkel-Rousse and Ferrari (2006) find that the balances of opinion lead to better or, at least, as accurate short-term forecasts of the manufacturing production growth rate as the Mitchell, Smith and Weale indicators.

In the present paper, we discuss a new statistical approach to forecast the manufacturing growth, with two important novelties. Firstly, we propose to exploit the heterogeneity of the firm-level survey responses by working out untreated data instead of balances of opinion. Secondly, we base our predictions on nonparametric forecasting algorithms inspired by the nearest neighbor and random forest regression methods, which are known to enjoy good generalization properties (Breiman, 2001a,b, Devroye, Györfi and Lugosi, 1996). Our algorithms exploit the heterogeneity of the survey responses, work fast, and allow the treatment of missing values. We argue that these procedures lead to significantly better results than more traditional competing methods.

The paper is organized as follows. In Section 2, we describe the data set used in this study. Section 3 is devoted to the presentation of our forecasting algorithms. Finally, in Section 4, we briefly describe the INSEE (National Institute for Statistics and Economic Studies) traditional methodology and compare its performance with our model.

2 The data

Our application will be based on a French data set related to the manufacturing sector. The quarterly manufacturing production growth rate is a quantitative data derived from the Quarterly National Accounts[1]. The entrepreneur individual qualitative responses are collected by the Business Survey Unit of the French Statistical Institute. Even if the French Industry survey is carried out on a monthly basis, we decided to use quarterly observations instead of monthly observations. This was motivated by the fact that the regular short-term forecasts of the economic activity performed by INSEE are precisely made on a quarterly basis. Our analysis covers the period ranging from the first quarter 1995 to the fourth quarter 2005. Moreover, we decided to test the forecasting performance of the methods on the type of data which are used in the operational conditions of the INSEE forecasting exercises[2]. Therefore, we focused on the survey responses carried out in February, May, September and November[3].

The INSEE surveys deal with questions relating to production at the product level (not at the firm level). More precisely, each firm can declare up to four products[4] and answers questions regarding each of these products. In our analysis, we chose to retain only the biggest products (in terms of amount of sales). The total number of firms entering the survey during the considered period is 6,336[5]. On average, the number of responses during the period is equal to 17. In order to apply our methods, we selected firms whose number of responses was larger than the 3rd Quartile (Q3). Hence, we retained 1,587

[1]The empirical analysis was carried out in early spring 2006. At that period, the last published release of the French quarterly accounts was the one presenting the first results relating to the fourth quarter of 2005 (expressed in 2000 constant prices).

[2]With the same results described below, we also test the accuracy in forecast performance of the different approaches with a second set of data using the survey carried out in January, April, July and October.

[3]The "Notes de conjoncture" are issued three times a year in March, June, and December. A more concise "Point de Conjoncture" updates the June Note in October. These publications present INSEE short term forecasts.

[4]1.4 product per firm is declared on average.

[5]Note that about 4,000 industrial entrepreneurs are interviewed during each survey. However, owing to economic developments (closure or restructuring of enterprises), the sample is updated periodically.

firms, and this gives on average 38 responses out of the 44 possible during the period (see Table 1 which presents a summary).

Table 1: Selection of firms.

BTS quarterly data from 1995-1 to 2005-4 (February, May, September, November).
Maximum responses in the period: 44. Total number of firms : 6,336. Average number of responses: 17. Median: 13. Q3 (3rd quartile): 27. **Selection of 1,587 firms** whose number of response is larger than 27. Average of their response: 38. Median of their response: 39. Q3: 43.

Let us consider a BTS, related to time t, in which a sample of $m = 1587$ manufacturing firms are asked whether their production has risen, remained unchanged or fallen. The responses are collected in a $m \times 2$ matrix denoted by X_t:

$$X_t = \begin{pmatrix} x_{1,1}^t & x_{1,2}^t \\ x_{2,1}^t & x_{2,2}^t \\ \vdots & \vdots \\ x_{i,1}^t & x_{i,2}^t \\ \vdots & \vdots \\ x_{m,1}^t & x_{m,2}^t \end{pmatrix}$$

where x_{ij}^t stands for the answer of the firm i regarding the past production ($j = 1$) and the expected production ($j = 2$). As explained earlier, each x_{ij}^t can take four values:

$$x_{ij}^t = \begin{cases} -1 & \text{for the answer “down”} \\ 0 & \text{for the answer “unchanged”} \\ 1 & \text{for the answer “up”} \\ NA & \text{when there is no response.} \end{cases}$$

With this notation, each observation X_t consists of $2m$ variables. Associated with each X_t is the manufacturing production quarterly growth rate

observed at time t, denoted hereafter by Y_t. Thus, given a **new** BTS represented by a generic matrix $X = (x_{ij})$, the statistical problem is to predict the associated manufacturing production quarterly growth rate Y from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where n is the number of data items which are available to make the prediction (to lighten notation, we drop out the symbol t in the generic X and Y). In our problem, $n = 44$.

Despite their qualitative nature, the surveys can be used to make quantitative short-term predictions of the macroeconomic magnitudes. This is a very useful exercise, as it can be carried out well before the national accounts figures become available. The results of the BTS are available about 2 months before the publication of the first estimates of the growth of Gross Domestic Product (GDP), that is at a particularly early point in time from the point of view of forecasters. We are now in a position to present our forecasting algorithms.

3 Forecasting algorithms

3.1 k-Nearest neighbor regression

The k-nearest neighbor regression is among the most popular nonparametric methods used in statistical pattern recognition with over 900 research articles published on the method since 1981 alone! Dasarathy (1991) has provided a comprehensive collection of around 140 key papers. To summarize in our context, given a new observation X, the technique consists in finding the k nearest neighbors of X among the past observations $X_1, \ldots, X_n$. Then the manufactured production growth rate Y associated with X is predicted by the mean of the k-observed rates in the past.

Given the entrepreneur individual responses to the current survey $X_t, t = 1, \ldots, n$, the algorithm prediction is based on the outcomes of the k neighbors of the observation X. Therefore, it is necessary to measure how far a new data item X is from any X_t in the training sample. A natural choice is to consider the Euclidean distance for matrices (also called Schur distance) defined as

$$||X - X_t|| = \sqrt{\sum_{\substack{1 \le i \le m \\ 1 \le j \le 2}} (x_{ij} - x_{ij}^t)^2}\,.$$

However, to take into account the missing observations in the X_t, we propose to modify this distance as follows:

$$||X - X_t||^2 = \begin{cases} \dfrac{1}{\operatorname{Card} \mathcal{C}(X, X_t)} \displaystyle\sum_{(i,j) \in \mathcal{C}(X, X_t)} (x_{ij} - x_{ij}^t)^2 & \text{if } \operatorname{Card} \mathcal{C} \ge m \\ +\infty & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\mathcal{C}(X, X_t)$ denotes the set of indexes of the elements of X and X_t corresponding to an effective response of the firm (*i.e.*, not NA), that is

$$\mathcal{C}(X, X_t) = \Big\{(i,j) : 1 \leq i \leq m, 1 \leq j \leq 2, x_{ij} \neq NA, x_{ij}^t \neq NA\Big\}.$$

Note that the definition (3.1) implies that X_t cannot be among the k nearest neighbors of X if the proportion of common responses between X and X_t does not exceed one half.

Practically speaking, to perform the k-nearest neighbor regression, we first reorder the data

$$\big(X_{(1)}(X), Y_{(1)}(X)\big), \ldots, \big(X_{(n)}(X), Y_{(n)}(X)\big)$$

according to increasing distances $||X - X_t||$ (defined by (3.1)) of the X_t to X. In other words, $X_{(t)}(X)$ is the t-th nearest neighbor of X amongst $X_1, \ldots, X_n$. If distance ties occur, a tie-breaking strategy must be defined. For example, in case of $||X_t - X|| = ||X_{t'} - X||$, X_t may be declared closer to X if $t < t'$, *i.e.*, the tie-breaking is done by indices. The k-nearest neighbor prediction function g is then defined as the average of the k-nearest neighbor outcomes:

$$g(X) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(X).$$

3.2 Random forests

In a series of recent papers, Breiman (2001a,b) has demonstrated that consequential gains in prediction accuracy can be achieved by using a set of trees. Each tree in the set is growing in accordance to a generated random vector. Final predictions are obtained by aggregating over the tree set, typically using equal weights. In this section, we investigate how random forests can be adapted to the prediction problem we are dealing with (see Breiman, 2001b for more material).

3.2.1 How random forests work

Regression trees partition the space into regions, often hyperrectangles parallel to the axes. Among these, the most important are the binary regression trees, since they have just two children per node and are thus easiest to manipulate and update. Many strategies have been proposed for constructing the binary decision tree (in which each internal node corresponds to a cut, and each terminal node corresponds to a set in the partition). For examples

and list of references, we refer the reader to Devroye, Györfi and Lugosi (1996). Given a generic X in R^{2m}

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{i,1} & x_{i,2} \\ \vdots & \vdots \\ x_{m,1} & x_{m,2} \end{pmatrix}$$

a standard strategy is to let each node splits the data set according to a linear decision on one, and only one, variable x_{ij} (see Figure 1 that depicts an example).

The tree regression algorithms are presented in detail in the book of Hastie, Tibshirani and Friedman (2001). In this paragraph, we recall the core concepts and briefly present how to grow a binary regression tree using a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The algorithm CART automatically decides both splitting variables and split points. Suppose for example that we have a partition into M regions, say $R_1, R_2, ..., R_M$, and we model the tree regressors as a constant c_m in each region. Then the best $\hat{c}_m$ is just the average of the Y_t falling in region R_m. Finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible because the potential nimber of regions can be huge. Hence, it is usually done trough the following heuristic. Starting with all observations, consider a splitting variable $x_{i,j}$ and split point s, and define the pair of half-planes

$$R_1\big[(i,j),s\big] = \{x_{i,j} \leq s\} \quad \text{and} \quad R_2[(i,j),s\big] = \{x_{i,j} > s\}.$$

Then we seek the splitting variable j and split point s which solve

$$\min_{j,s}\Big[\min_{c_1} \sum_{X_t \in R_1[(i,j),s]} (Y_t - c_1)^2 + \min_{c_2} \sum_{X_t \in R_2[(i,j),s]} (Y_t - c_2)^2\Big].$$

For any choice j and s, the inner minimization is solved by $\hat{c}_1$ (respectively $\hat{c}_2$) equal to the average of the Y_t associated with the X_t falling in R_1 (respectively R_2). For each splitting variable, the determination of the split point s can be done very quickly. Therefore, by scanning through all the inputs, determination of the best pair $[(i,j),s]$ is feasible. Having found the best split, we partition the data set into two resulting regions, we repeat the splitting process on each of the two regions, and so on. The process continues until each node (*i.e.*, a region) reaches a user-specified minimum node size N_{min} and becomes a terminal node. In our problem, the terminal nodes, taken together, form a partition of R^{2m}, and the tree regressor h is

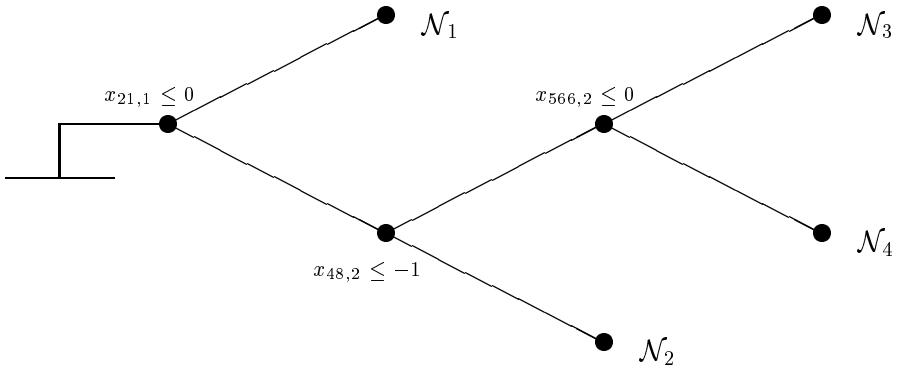

Figure 1: An example of binary tree.

then defined on each terminal region by the mean

$$h(X) = \frac{1}{\text{Card}\,\{t : X_t \in \mathcal{N}(X)\}} \sum_{t:X_t \in \mathcal{N}(X)} Y_t,$$

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

In the present paper, we propose to use a related approach called random forest proposed by Breiman (2001a,b) and which consists in growing a set of smaller size trees. More precisely, a random forest is a collection of tree predictors h_k, $k = 1, \ldots, K$, where each tree is constructed from a bootstrap (Efron and Tibshirani, 1993) sample drawn with replacement from the training data. However, instead of determining the optimal split on a given node by evaluating all possible splits on all covariates, a subset of the covariates, drawn at random, is used. Thus, formally each tree is grown as follows:

1. Construct a bootstrap sample from $(X_1, Y_1), \ldots, (X_n, Y_n)$.
2. Choose N_{min}, the minimum node size.
3. Specify $p << 2m$ such that, at each node, p variables only are selected at random out of the $2m$. The best splits (calculated with the CART algorithm) on these p variables for the bootstrap sample is used to split the node. Note that the value of p is held constant during the growth of the forest.

For the free parameters K, N_{min} and p, we used the default values $K = 500$, $N_{min} = 5$ and $p = \sqrt{n}$ of the random forest R-package[6].

[6] http://lib.stat.cmu.edu/R/CRAN/src/contrib/Descriptions/randomForest.html

Finally, the prediction is the unweighted average over the tree collection, that is

$$h(X) = \frac{1}{K} \sum_{k=1}^{K} h_k(X). \tag{3.2}$$

Breiman (2001a,b) argues that random forests enjoy exceptional prediction accuracy, and that this accuracy is achieved for a wide range of settings of the tuning parameters. In addition, random forests possess a number of interesting features, including measures of proximities between the observations and measures of covariate importance. In the next paragraph, we investigate how these features can be used to deal with the problems of missing values and variable selection.

3.2.2 Missing values and variable selection

The random forest predictor (3.2) does not support missing values in the X_t. As suggested by Breiman (2001b), missing values can be estimated by constructing proximities between the observations in the training sample. To this aim, after a tree is grown, we put all the data items X_t, $t = 1, \ldots, n$, down the tree. If t and t' are in the same terminal node, we increase the proximity between X_t and $X_{t'}$ by one. To finish, we normalize by dividing by the number of trees. Thus, if K stands for the number of tree predictors, the proximity $P(X_t, X_{t'})$ between X_t and $X_{t'}$ is defined by

$$P(X_t, X_{t'}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\{X_t \in \mathcal{N}_k(X_{t'})\}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\{X_{t'} \in \mathcal{N}_k(X_t)\}}$$

where $\mathcal{N}_k(X)$ is the terminal node of the tree h_k which contains X.

Starting from Breiman's idea of proximity, we discuss now a new algorithm, called **RF1**, which allows the treatment of missing values. For notational convenience, X will be denoted X_{n+1}.

RF1

INPUT: $(X_1, Y_1), \ldots, (X_n, Y_n)$, X_{n+1}.

1. Consider any prediction $\tilde{Y}_{n+1}$ associated with X_{n+1}. Denote by $\mathcal{S}$ the augmented sample $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \tilde{Y}_{n+1})$.

2. Fill in the missing values by the method of your choice. Denote by $\tilde{\mathcal{S}}$ the sample $(\tilde{X}_1, Y_1), \ldots, (\tilde{X}_n, Y_n), (\tilde{X}_{n+1}, \tilde{Y}_{n+1})$ without missing values.

3. Run the random forest algorithm on $\tilde{\mathcal{S}}$ and compute proximities.

4. Replace the missing values in the sample $\mathcal{S}$ by the average of the corresponding variables weighted by the proximities between the relevant cases and the non missing-value cases. More precisely, if $x_{ij}^t = NA$, replace it by

$$\frac{1}{\sum_{\{t':t'\neq t, x_{i,j}^{t'}\neq NA\}} P(\tilde{X}_t, \tilde{X}_{t'})} \sum_{\{t':t'\neq t, x_{i,j}^{t'}\neq NA\}} P(\tilde{X}_t, \tilde{X}_{t'}) x_{i,j}^{t'}.$$

   Denote by $\tilde{\mathcal{S}} = (\tilde{X}_1, Y_1), \ldots, (\tilde{X}_n, Y_n), (\tilde{X}_{n+1}, \tilde{Y}_{n+1})$ the resulting sample.

5. Iterate N times step 3. and step 4.

OUTPUT: the outcome predicted for $\tilde{X}_{n+1}$ by the random forest algorithm based on $(\tilde{X}_1, Y_1), \ldots, (\tilde{X}_n, Y_n)$.

Breiman argues that $N = 5$ iterations are generally enough. In our experiments, we chose for the initial $\tilde{Y}_{n+1}$ the (linear) prediction obtained by the traditional INSEE methodology, which will be described in Section 4.

Recall that each observation X_t takes its values in a space of dimension $2 \times m = 3174$. However, it is well established that in high dimensional spaces, learning suffers from the curse of dimensionality (see for example Abraham, Biau, and Cadre, 2006). Thus, in practice, before applying any learning technique to model real data, a preliminary dimension reduction or model selection step is crucial for appropriate smoothing and circumvention of the dimensionality effect. In this respect, Breiman (2001b) suggests a

measure, called variable importance, to discriminate between informative a noninformative variables. In the algorithm **RF2** below, we include this measure. The general idea is to run the random forest algorithm only on the most important variables (see Breiman, 2001b, for more information).

RF2

INPUT: $(X_1, Y_1), \ldots, (X_n, Y_n)$, X_{n+1}.

1. Run the algorithm $RF1$ with input data $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}$ and compute the variable importance for each of the $2m$ variables.

2. Specify $p_{max} \leq 2m$ and for $t = 1, \ldots, n+1$, denote by $\bar{X}_t$ the vector composed of the p_{max} most important variables of X_t.

OUTPUT: the outcome predicted by $RF1$ with input data $(\bar{X}_1, Y_1), \ldots, (\bar{X}_n, Y_n), \bar{X}_{n+1}$.

In our experiences, we observed that the choice $p_{max} = 150$ variables was enough. Thus, this dimension reduction step means that the algorithm automatically selects the 150 most representative entrepreneur answers out of the 3174 possible ones.

4 Results and comparison with the INSEE methodology

Before presenting the practical results, we briefly describe the traditional INSEE methodology, which is based on linear models on the balances of opinion. These models are the most currently used indicators for short-term analysis.

4.1 INSEE methodology

Balances of opinion are interesting indicators in many respects. Firstly, they are easy to implement. As univariate series, they are simple to read and to track over time, at the price of an acceptable loss of information with respect to the corresponding exhaustive three-dimensional statistics. Secondly, balances of opinion are subject to limited revisions across time. Finally, the main balances of opinion—notably those relating to activity—are highly correlated with the corresponding aggregates of interest, even though they are generally smoother (and therefore easier to read). This

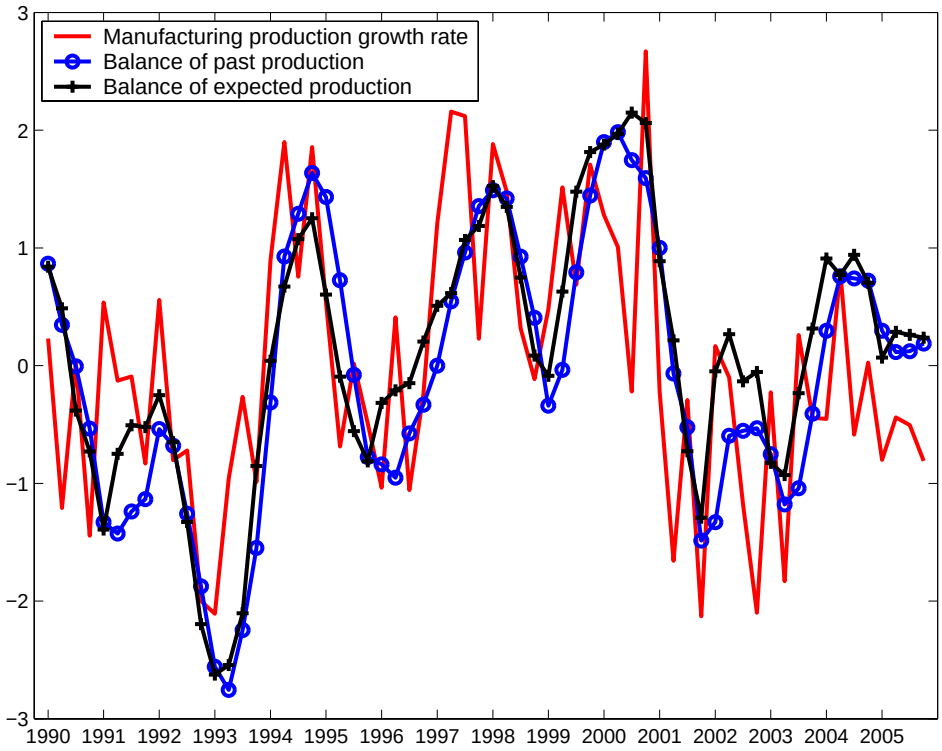

Figure 2: Balances of opinion relating to manufactured production together with the manufacturing production quarterly growth rate. (Note that the data set has been centered and standardized).

is typically the case, for instance, for the balances of opinion relating to past production derived from the INSEE Industry survey (see Figure 2). All these interesting properties explain why the balances of opinion are the main (if not the only) indicators used by short-term analysts as explanatory variables in a linear model. All in all, due to their good empirical properties, the balances of opinion prove to be very useful, as they are well adapted to the quick production and release conditions of BTS.

The most common methodology to predict the quarterly national accounts using business surveys, known as calibrations (see Raynaud and Scherrer, 1996, Buffeteau and Mora, 2000, Dubois and Michaux, 2006), consists in fitting a linear model between the balances of opinion S_j^t (as before, $j = 1$ for the past production, and $j = 2$ for the expected production), and the dependent variable Y_t, which may typically be the manufacturing production growth. In mathematical terms,

$$Y_t = c + a_1 S_1^t + a_2 S_2^t + u_t,$$

where u_t is some random noise.

The quality of this kind of model can be slightly improved by including the past values of Y and by taking into account the variation of the balance of opinion. Nevertheless, in the present paper, we will focus on this simple model, whose validity and robustness has already been established, through the application of several specification tests using the estimated residuals, such as tests of stability of the coefficients (Chow test), tests of homoskedasticty (White test), or test of normality. We finally note that the calibration model uses the balances of opinion as computed and published by the INSEE. These balances are based on the 4,000 firms data items, which are preprocessed to deal with missing values and seasonal adjustement. In the present study, the INSEE approach should be considered as a benchmark.

4.2 Results

The error rate for forecasting new observations is unknown. However, it can be estimated using a simple leave one out methodology. To this aim, we select one item X_t together with its outcome Y_t out of the 44 observations, and we consider it as new observation. Next, we determine the outcome $\hat{Y}_t$ using the procedure under study worked out with the 43 remaining data items (see Figure 3), and we finally compare the estimated outcome with the true one. This process, repeated for each of the 44 observations, provides us with an estimate of the mean square error rate, denoted hereafter by MSE:

$$\text{MSE} = \frac{1}{44}\sum_{t=1}^{44}(Y_t - \hat{Y}_t)^2.$$

We will use the following acronyms:

- LM refers to the linear model on the balances of opinion.
- KNN refers to the k-nearest neighbor regression. Note that the parameter k is chosen so as to minimize MSE.
- RF1 and RF2 stand for the random forest-type algorithms described in Section 3.

The results obtained by the different procedures are presented in Table 2. We note the excellent performance achieved by the algorithm RF2, whose MSE is one third better than the error rate obtained by the traditional INSEE methodology LM. The difference between RF1 and RF2 enlightens the importance of the variable selection step. Similarly, the poor results of the KNN method could undoubtedly be improved with a preliminary variable selection step. To illustrate the superiority of RF2 over its competitors, we performed the bilateral Harvey, Leybourne, and Newbold (1997) tests of equal accuracy in forecast performance. The results are presented in Table 3.

Table 2: Results of the different procedures.

Method	MSE
LM	1.039
KNN	1.194
RF1	1.105
RF2	0.702

Table 3: Results of the Harvey, Leybourne, and Newbold (1997) tests of equal accuracy in forecast performance.

Test of the model involving RF2 versus the model involving	Difference of MSE	Test statistic	p-value
LM	-0.337	-1.411	0.082
KNN	-0.492	-3.425	0.000
RF1	-0.403	-4.108	0.000

At 0.9 confidence level, we reject the null hypothesis that the forecast performance accuracy of the tested method (one per line) is similar to RF2 if the p-value is less than 10%. An inspection of Table 3 reveals that the model RF2 lead to significantly better forecast performances than the other models. We finally note that te RF2 algorithm works fast: using the R-package "RandomForest", our prediction takes less than one minute.

5 Perspectives

To improve the results of the present study, we suggest two research directions. Firstly, it seems important to study the impact of putting weights on the entrepreneur responses: under the assumption that the firm size is correlated with the macro-economic production, an improvement in the relative performances of the nonparametric approaches is possible. Secondly, one could use these new algorithms with other surveys (e.g. using retail trade survey to forecast household consumption) or mix the surveys (eg. industry and services) to forecast the GDP. Finally, it would also be interesting to identify the 150 variables which are automatically selected by the algorithm RF2 (size, sector...). With this preliminary selection step, the calibration model using balances of opinion could also undoubtedly be improved.

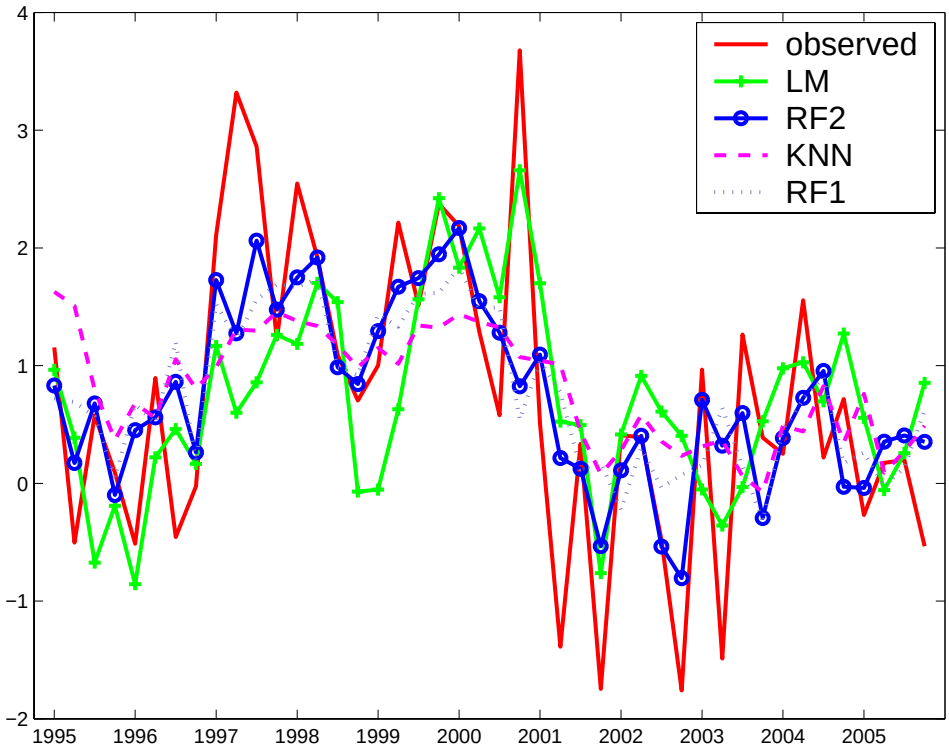

Figure 3: Manufacturing production quarterly growth rate and predictions obtained by the different methods.

References

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 2005. In press.

[2] O. Biau, H. Erkel-Rousse, and N. Ferrari. Individual responses to business tendency surveys and the forecasting of manufactured production: An assessment of the Mitchell, Smith and Weale dis-aggregate indicators on French data. In *European Commission-OECD joint Workshop on international developments of business and consumer tendency surveys*, Bruxelles, November 2005.

[3] L. Breiman. Statistical modeling: the two cultures. *Statistical Science*, 13:119–215, 2001a.

[4] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001b.

[5] L. Breiman, J.H. Friedman, R.A. Olsen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[6] S. Buffeteau and V. Mora. Predicting the national accounts of the euro zone using business surveys. *Conjoncture in France, INSEE*, December 2002.

[7] B.V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, 1991.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[9] E. Dubois and E. Michaux. Etalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats. *Economie et Prévision*, 2006. forthcoming.

[10] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[11] D.I. Harvey, S.J. Leybourne, and P. Newbold. Testing the equality of prediction mean square errors. *International Journal of Forecasting*, 13:273–281, 1997.

[12] T. Hastie, R.J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[13] J. Mitchell, R.J. Smith, and M.R. Weale. Aggregate versus disaggregate survey-based indicators of economic activity. In *27th CIRET conference*, Warsaw, September 2004.

[14] M. Reynaud and S. Scherrer. Une modélisation VAR de l'enquête de conjoncture de l'INSEE dans l'industrie. *Document de travail de la Direction de la Prévision*, 96-12, 1996.