Euro area GDP forecasting using large survey datasets
A random forest approach

Olivier Biau - Angela D'Elia
Directorate General for Economic and Financial Affairs
European Commission

Views expressed represent exclusively the positions of the authors and do not necessarily correspond to those of the European Commission

EU workshop on recent developments in business and consumers surveys
Brussels, 15-16 November 2010

Introduction

- ✓ **Increasing interest in forecasting methods that utilise large datasets**
- ✓ **Not only an issue of academic interest**

  Eklund and Kapetanios, 2008
- ✓ **Factor methods in the forefront of developments**

  Stock and Watson, 2002; Forni *et al.*, 2005; Giannone *et al.*, 2008
- ✓ **Factor analysis combined with linear modelling**

Introduction

- **A new statistical approach to forecasting macro-economic aggregates, based on the Random Forests technique**

  Breiman, 2001, 2002

- **RF algorithm is widely applied in medical research and biological studies, becomes more and more popular and appears to be very powerful**

  Arun and Langmead, 2006; Díaz-Uriarte and Alvarez de Andrés, 2006; Ward *et al.*, 2006

- **RF is largely unknown in economics**

  Biau. G, Biau. O and Rouvière, 2007

Introduction

- ✓ **RF enjoys good prediction properties, is robust to noise and can handle a very large number of input variables**
- ✓ **RF is considered to be one of the most accurate general-purpose learning techniques available, independent of any functional and distributional assumptions**
- ✓ **However, from a mathematical point of view, the mechanism of RF algorithms remains largely unknown and is not clearly explained**

  Breiman, 2002; Lin and Jeon, 2006; Biau. G *et al.*, 2008, Biau. G and Devroye, 2008; Biau. G, 2010

Introduction

- ✓ **A specific application for short-term GDP forecasting in the euro area is shown using (*only*) the harmonized European Union Business and Consumer surveys dataset**
- ✓ **A typical high-dimensional regression problem ($n << p$)**
- ✓ **The RF technique is explored with two aims in mind:**
  1. to obtain a non-parametric forecast of GDP
  2. to obtain a ranking of the explanatory variables, and then select those variables to build a linear model to forecast GDP
- ✓ **The forecast performance is assessed through an out-of-sample exercise**

Outline

- ✓ Data
- ✓ RF algorithm
- ✓ Benchmark
- ✓ Results

Data

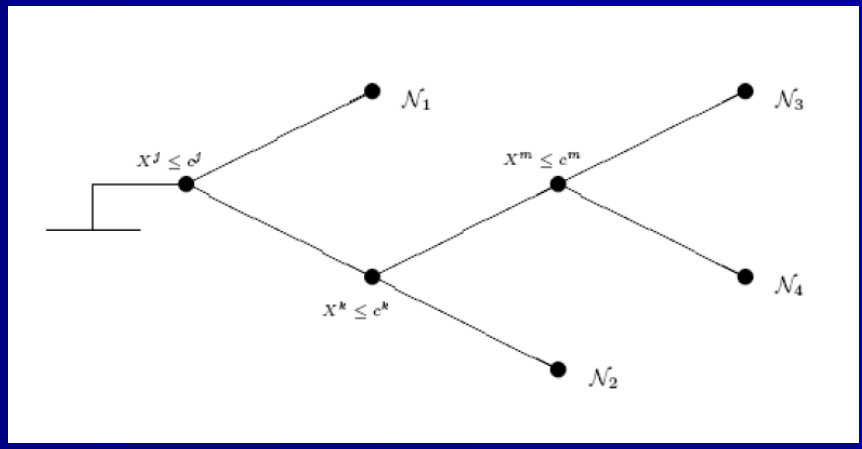
- ✓ Based on the Joint Harmonized EU Programme of Business and Consumer Surveys (manufacturing industry, services, retail trade, construction, and consumers)
- ✓ Qualitative surveys
- ✓ It covers all the 27 Members States, Croatia, the FYROM and Turkey
- ✓ More than 125 000 firms and over 40 000 consumers surveyed every month

Data

- The dataset mainly consists of the euro area balances of opinion (% positive - % negative)
- The time series used in the analysis are those available at the end of the third month of each quarter:
  - the level series: monthly *St* or quarterly *Sq*,
  - the difference series: (*St* - *St-1*), (*St* - *St-2*), (*St* - *St-3*) and (*Sq* - *Sq-1*)
- The dataset is composed of $p = 172$ 'soft' series: X_i
- The only 'hard' variable is the euro area GDP qoq growth series: Y_i
- Finally, we have a « learning set » $L = \{(X_1, Y_1) \ldots (X_n, Y_n)\}$ $n = 57$ (1995Q3 - 2009Q3)

From binary trees to Random Forest

- ✓ What is a binary tree?



- ✓ How to grow a tree?

- ✓ What is the tree predictor (or tree regressor)?

How to grow a tree? with CART

How to predict the income of a newcomer (entering the room) based on observed characteristics X_i (gender, size, weight, age, …) and income Y_i of people attending the conference?

- **For the first node, we seek the first splitting variable X^j and the first split point s which discriminate the most, by solving:**

$$\min_{j,s}\left[\min_{c_1}\sum_{X_i\in N_1[j,s]}(Y_i-c_1)^2+\min_{c_2}\sum_{X_i\in N_2[j,s]}(Y_i-c_2)^2\right]$$

- **For any choice j and s, the inner minimization is solved by:**

$$c_i = average(Y_i / i\in N_i)$$

$$where \quad N_1=[X^j\le s] \quad and \quad N_2=[X^j>s]$$

- **Having found the best split, we partition the data into the two resulting regions and repeat the splitting process until each node reaches a user-specified minimum *nodesize* and becomes a terminal node**

What is the tree predictor?

Once the tree is built, a new *X* arrives…

… it ‘falls’ into a terminal node N(X),

… the tree regressor *h(X)* is computed by averaging the observed Y_i over the observations *i* ‘falling’ in that node:

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

RF algorithm

- **Breiman's idea: instead of finding the best tree, build a large number (*K*) of simpler regression trees and aggregate them**

- **RF algorithm (Hastie and Tibshirani, 2009)**

  1. For $k = 1$ to ***K***:

  (a) Draw a bootstrap sample from the learning dataset $L = \{ (X_1, Y_1) \ldots (X_n, Y_n) \}$.

  (b) Grow a random-forest tree h_k to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum ***nodesize*** is reached.

  - Select ***mtry*** variables at random from the ***p*** variables
  - Pick the best variable/split-point among the ***mtry***
  - Split the node into two daughter nodes

  2. the predicted outcome (final decision) is obtained as the average value over the ***K*** trees:

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

- **For the free parameters *K*, *nodesize* and *mtry*, we used the default values 500, 5 and *p*/3 of the random forest R-package**

RF algorithm

- ✓ **RF posses one important feature to reduce data dimensionality**
- ✓ **Breiman (2001) suggests a measure called 'variable importance' to discriminate between informative and noninformative variables**
- ✓ **For each variable, the idea is to compare the prediction error with the prediction error where the variable is randomly permuted**
  - Large positive values for a variable indicate that this variable is predictive
  - zero or negative importance values indicate non-predictive variables

Benchmark

- We predict GDP growth for quarter Q (nowcast), based on data available at the end on quarter Q

- Out-of-sample analysis:
  - 2004Q1-2009Q3, using GDP vintage data
  - Criterion: Mean Square Error (MSE)

- Two competitors:
  - AR model
  - *euro zone economic outlook*
    - Quarterly publication, published in the first days of quarter Q+1
    - Nowcast (Q), 2-steps-ahead projections (Q+1 and Q+2) for GDP, IP, Cons., Inflation
    - But also economic links explaining these forecasts

- How does a data-driven model like the RF perform relative to competitors for GDP nowcasting?

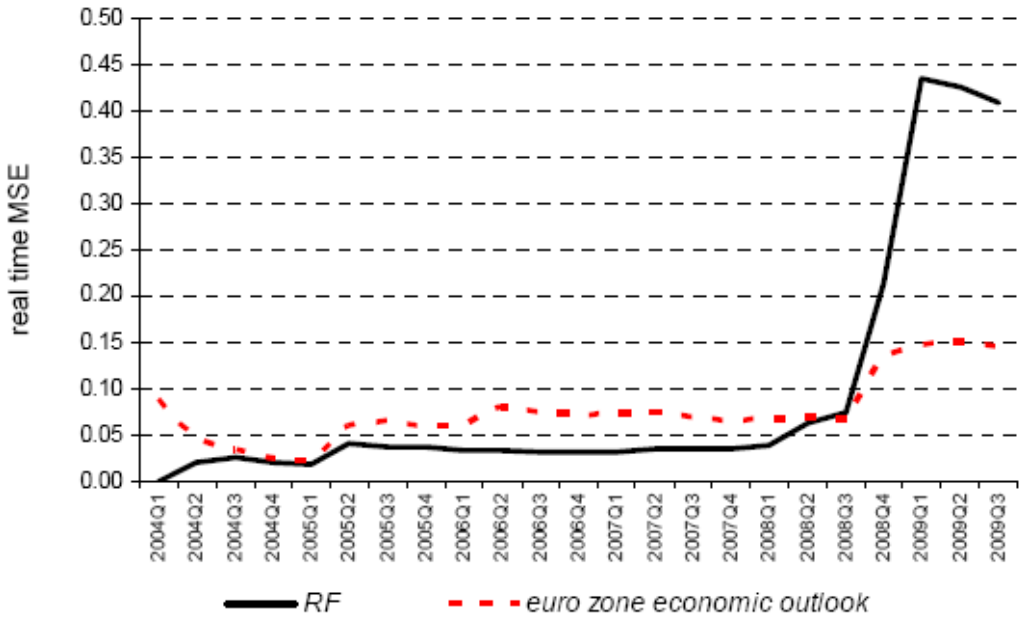
Results (1)

RF: pure Random Forest

- RF outperforms the AR but not the *euro zone economic outlook*
- Good performance of RF before the crisis
- Poor performance during the crisis… however, by construction!

Mean Square Error - 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)



Results (2)

- ✓ **To square the problem of non negative values in the learning set, a two steps procedure:**
  - **Select the 25 most *important variables* (RF feature)**
  - **Choose the « best » linear model to explain GDP (Gets)**

- ✓ **RF_LINMOD: model retained**

$$100.GDP_t = \underset{(12.033)}{0.615} + \underset{(2.683)}{0.011} V12_t + \underset{(3.248)}{0.032} V24_t + \underset{(3.152)}{0.020} V41_t + \underset{(4.153)}{0.044} V62_t + \underset{(1.933)}{0.025} V101_t$$

OLS results — estimation period 1995Q3 -2009Q2 — Standard error of the regression = 0.246

Values of t-statistic in brackets — R^2=0.858, adjusted R^2= 0.844, DW(0) = 2.18.

V12: Orders development over past 3 months - INDU

V24: Expectation about household financial positions over next 12 months - CONS

V41: Orders development expected over next 3 months - RETA

V62: Export orders development expected over next 3 months — difference series (t – t-1) - INDU

V101: Assessment of current order book — difference series (t – t-2) - INDU

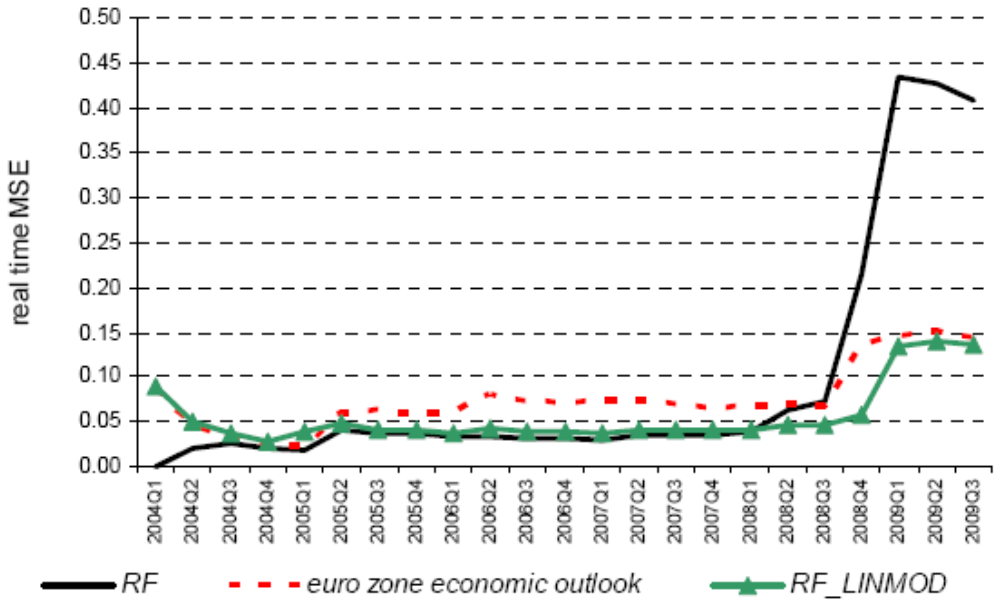
Results (2)

RF_LINMOD

- performs as well as the *euro zone economic outlook*
- outperforms the pure RF and compares well to the *euro zone economic outlook* during the 'crisis'

Mean Square Error - MSE	
RF	*0.43*
euro zone economic outlook	*0.15*
RF_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)



Conclusion

- ✓ RF: a new approach for short-term analysis
  - to forecast macroeconomic variables
  - to reduce data dimensionality
- ✓ RF is fast and easy to implement
- ✓ RF outperforms AR and compares well with reliable forecasts
- ✓ RF is worth adding to the economists' toolbox