Economics Bulletin

Volume 30, Issue 1

A Short Note on the Nowcasting and the Forecasting of Euro-area GDP Using Non-Parametric Techniques

Dominique Guégan
PSE CES--MSE University Paris1 Panthéon-Sorbonne

Patrick Rakotomarolahy
CES--MSE University Paris1 Panthéon-Sorbonne

Abstract

The aim of this paper is to introduce a new methodology to forecast the monthly economic indicators used in the Gross Domestic Product (GDP) modelling in order to improve the forecasting accuracy. Our approach is based on multivariate k-nearest neighbors method and radial basis function method for which we provide new theoretical results. We apply these two methods to compute the quarter GDP on the Euro-zone, comparing our approach, with GDP obtained when we estimate the monthly indicators with a linear model, which is often used as a benchmark.

Citation: Dominique Guégan and Patrick Rakotomarolahy, (2010) "A Short Note on the Nowcasting and the Forecasting of Euro-area GDP Using Non-Parametric Techniques", *Economics Bulletin*, Vol. 30 no.1 pp. 508-518.
Submitted: Dec 30 2009. **Published:** February 10, 2010.

1 Introduction

The aim of this paper is to introduce a new methodology to forecast the monthly economic indicators used in the Gross Domestic Product (GDP) modelling in order to improve its forecasting accuracy.

The GDP is only available on quartely basis with a time span of 2 or 3 months, and sometimes with significant revisions. Thus, governments and central banks need to have accurate tools to update consistently the information used to revise and provide good forecasts for GDP. Monthly economic indicators are routinely used to assess the current economic conditions before GDP figures are made available. There exist different methods for modelling monthly indicators in order to provide a forecast of the quartely GDP. They appear, for instance, in dynamic factor models (Kapetanios and Marcellino, 2006), MIDAS regressions (Marcellino and Schumacher, 2008), structural models (Clements and Hendry, 1999) or bridge equations (Baffigi *et al.* 2004, and Darne 2008). In all cases, the estimation of the monthly indicators is determinant.

We know that GDP is published with a delay and is subject to revision. Bridge equations (BE) are a tool for estimating such quantity, which is measured at a quartely basis, by means of suitable shorter-term indicators. In this paper we forecast monthly indicators used in BE via a non-parametric approach and we show in application to Euro area GDP that this new approach outperforms linear models that are often used in practice. Several BE are considered, which are meant to illustrate the trade-offs between timeliness, tightness of the relationship of the variables with GDP and sizes of revisions. Thus, some equations use data on actual activity, such as industrial production and retail sales, which are subject to significant revisions and are published with significant delays, while some use more indirect indicators of activity, such as confidence surveys and financial variables data, which are typically not revised and available on a timely basis. The real time information set starts in January 1990 when possible and ends in November 2007. The vintage data set for a given month takes the form of an unbalanced set at the end of the sample and we use non-parametric methods to forecast the monthly variables in order to complete their values until the end of the current quarter for GDP nowcasts and until the end of the next quarter for GDP forecasts. Then we aggregate the monthly data to quartely frequencies that we plug in the eight BE proposed by Diron (2008). Our methodology is based on non-parametric techniques.

The non-parametric techniques that we use in this paper work in a multivariate setting: they are the multivariate nearest neighbors (NN) method and the radial basis function (RBF) method. Assuming that we observe an economic indicator X, on a given period, say $X_1, \cdots, X_n$, we embed this information set in a space of dimension $d \in \mathbb{N}^*$ to build NN or RBF forecasts $\hat{X}_{n+h}$, $h \geq 1$, that we use finally in the GDP equations. We provide the algorithm that we use for both methods and also theoretical results proving the accuracy of the forecasts under very smooth assumptions. We apply these methods to compute the quarter GDP on the Euro-zone, comparing our approach, with GDP obtained when we estimate the monthly indicators with a linear model, which is often used as a benchmark.

We describe in section two our methodology providing also asymptotical results. In section three, we exhibit the nowcasting and forecasting of the GDP.

2 The methodology

We describe two methods we consider here, the multivariate NN and RBF methods, (Yatchew, 1998) and references therein. The problem consists in estimating a regression function $m(.)$ linking two random variables $Y = m(X)$. This estimate will have the following representation, $m_n(x) = \sum_{i=1}^{n} \omega_{i,n}(x) Y_i$, where $\omega_{i,n}$ are weights to be specified (Silverman 1986, and Guégan 2003).

Assuming that we observe a time series in $\mathbb{R}$, we transform such original data set by embedding it in a space of dimension d, building vectors $(\underline{X}_n)_n \subset \mathbb{R}^d$. The embedding is interesting because it permits different features of the data to be taken into account which are not observed on the trajectory. Working with NN method, we get an estimate of $m(\underline{x})$, $\underline{x} \in \mathbb{R}^d$, using the k closest vectors of $\underline{X}_n$ inside the training set $S \subset \mathbb{R}^d$. Working with RBF approach, we estimate $m(.)$ by a set of k clusters through a radial basis functions ϕ. We detail now these methodologies.

1. Multivariate k-NN estimate for $m(.)$. After embedding, we determine the k closest vectors of $\underline{X}_n = (X_{n-d+1}, \cdots, X_n)$ inside the training set $S \subset \mathbb{R}^d$:

$$S = \{\underline{X}_{\ell+d} = (X_{\ell+1}, \cdots, X_{\ell+d}) \mid \ell = 0, ..., \ell = n-d-1\}.$$

Denoting by $\underline{X}_{(i)}$ $i = 1, ..., k(n)$ the ith nearest neighbor of $\underline{x}$, then the k-NN estimate of $m(\underline{x})$ is:

$$m_n(\underline{x}) = \sum_{\underline{X}_{(i)} \in S, i=1}^{k(n)} w(\underline{x} - \underline{X}_{(i)}) X_{(i)+1}. \tag{2.1}$$

A general form for the weights is:

$$w(\underline{x}-\underline{X}_{(i)})=\frac{\frac{1}{nR_n^d}K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n})}{\frac{1}{nR_n^d}\sum_{i=1}^{n}K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n})},$$

where $K(.)$ is a given weighting function vanishing outside the unit sphere in $\mathbb{R}^d$ and R_n is the distance between the kth NN of $\underline{x}$ and $\underline{x}$ itself. In this multivariate NN method, we need to detect the neigbors, and to choose the weights. In practice, we often restrict to exponential weight, $K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n})=exp(-||\underline{x}-\underline{X}_{(i)}||^2)$ or to uniform weight, $K(\underline{x}-\underline{X}_{(i)})=\frac{1}{k}$.

2. RBF estimate for $m(.)$. As soon as the data have been embedded in a space of dimension d, we create $(n-d+1)$ vectors. Then, we use a k-means method to partition these vectors providing k clusters, denoted by $\mathcal{C}_i$, $i=1,...,k$. Each vector belongs to the cluster such that its distance to the cluster's center is minimal, then the RBF estimate of $m(\underline{x})$ is:

$$m_n(\underline{x})=w_0+\sum_{i=1}^{k}w_i\phi(\|\underline{x}-c_i\|,r_i). \tag{2.2}$$

The parameters $c_i=(c_1^i,...,c_d^i)\in\mathbb{R}^d$, $r_i\in\mathbb{R}$ and $w_i\in\mathbb{R}$ have to be estimated. The radial basis function $\phi(.)$ can be chosen among gaussian, multiquadric or inverse multiquadric functions (Guégan, 2003). As soon as the function ϕ and the parameters (c_i,r_i), $i=1,...,k$ are known, then $\phi(\|\underline{x}-c_i\|,r_i)$ is known and the function $m(\underline{x})$ is linear in w_i thus w_i is estimated by ordinary least squares method.

For both methods, all the parameters are determined in a space of dimension d. The properties of these estimates are given in the theorem below which provides new results. We now specify the assumptions needed to establish these properties.

We assume that we observe a strictly stationary time series $(X_n)_n$ that is characterized by an invariant measure with density f, the random variable $X_{n+1}\mid(\underline{X}_n=\underline{x})$ has a conditional density $f(y\mid\underline{x})$, and the invariant measure associated to the embedded time series $\{\underline{X}_n=(X_{n-d+1},\cdots,X_n)\}$ is h. On the other hand, we make the following assumptions:

H_0: The time series $(X_n)_n$ is ϕ-mixing.

H_1: $m(\underline{x}), f(y\mid\underline{x})$ and $h(\underline{x})$ are p continuously differentiable functions.

H_2: The function $f(y\mid\underline{x})$ is bounded,

H_3: There exists a sequence $k(n) < n$ such that $\sum_{i=1}^{k(n)} w_i \to 1$ as $\quad n \to \infty$.

H_4: For k-NN method, if $w(\underline{x}) = \frac{1}{k(n)}$, then $\sigma^2 = Var(X_{n+1} \mid \underline{X}_n = \underline{x})$; if $w(\underline{x}) \in \mathbb{R}$ depending (or not) on $(X_n)_n$, then $\sigma^2 = \gamma^2(Var(X_{n+1} \mid \underline{X}_n = \underline{x}))$, where $\gamma \in R^+$.

H_5: For RBF method, ϕ is Gaussian and the estimated weights w_i in (2.2) satisfy $(w_i - \frac{c^i_{d+1}}{A}) \to 0$ in distribution as $n \to \infty$, where $c^i_{d+1} = \frac{1}{N_i}\sum_{j,\underline{X}_j \in \mathcal{C}_i} X_{j+1}$, $N_i = \#\mathcal{C}_i$ and $A = \sum_{j=1}^{k} \exp(-\frac{\|\underline{x}-c_j\|^2}{2r_j^2})$.

Theorem 2.1. *We assume that $\{X_n\}$ is a stationary time series and that the assumptions H_0-H_2 are verified. Moreover, for the multivariate NN estimate (2.1), we assume that the assumptions H_3-H_4 are verified, and for the RBF estimate (2.2) we assume that the assumption H_5 is verified, then :*

$$\sqrt{n^Q}(m_n(\underline{x}) - Em_n(\underline{x})) \to_{\mathcal{D}} \mathcal{N}(0, \sigma^2), \tag{2.3}$$

with $0 \leq Q < 1$, and $Q = \frac{2p}{2p+d}$.

Proof: In case of multivariate NN approach, the convergence has been proven in Guégan and Rakotomarolahy (2009). In case of the RBF estimate, we can prove the same result remarking that $\phi(y,r) = exp(-\frac{y^2}{2r^2})$ and $w_i = \frac{c^i_{d+1}}{A}$ with $A = \sum_{j=1}^{k} \exp(-\frac{\|\underline{x}-c_j\|^2}{2r_j^2})$. The theorem is still true for inverse multiquadric radial basis function using the approximation $\frac{1}{\sqrt{y^2+r^2}} = \exp(-\frac{1}{2}\log(y^2 + r^2)) \approx \frac{1}{r}\exp(-\frac{y^2}{2r^2})$, when the centers c_i are close to $\underline{x}$.

This theorem provides results on robust estimation of regression, justifying the use of nonparametric methods to construct estimates from dependent variables. It extends the known results in the independent case for k-NN estimate to dependent variables (Stute, 1984) and in quadratic mean square error for uniform weight to more general weights (Yakowitz, 1987). It provides new results for RBF estimation. Thanks to the asymptotic normality, it is also possible to exhibit confidence intervals (Guégan and Rakotomarolahy, 2009), and to build density forecasts which can be used as back testing procedure.

3 Carrying out of the method

Information on the current state of economic activity is a crucial ingredient for policy making. Economic policy makers, international organisations and private sector forecasters commonly use short term forecasts of real gross domestic product (GDP) growth based on monthly indicators.

For users, an assessment of the reliability of these tools and of the source of potential forecast errors is essential. In the present exercise, we show that beyond the model chosen to calculate the GDP in the end, the forecasts of monthly economic indicators used in the final model are fundamental and may be considerably misleading if they are not properly estimated.

We therefore consider the approach of bridge equations to calculate the GDP in the final stage, limiting ourselves to the eight BE introduced in the paper of Diron (2008) where each equation provides a model of GDP, denoted $Y_t^j, j = 1, \cdots, 8$. They are finally aggregated consistently to provide a final value of GDP, denoted Y_t. Each equation is calculated from thirteen monthly economic indicators denoted by $X_t^i, i = 1, \cdots, 13$, which are listed in table 1. To realize this objective, we use the real-time data base provided by EABCN through their web site [1].

The real-time information set starts in January 1990 when possible (exceptions are the confidence indicator in services, that starts in 1995, and EuroCoin, that starts in 1999) and ends in November 2007. The vintage series for the OECD composite leading indicator are available through the OECD real-time data base [2]. The EuroCoin index is taken as released by the Bank of Italy. The vintage data base for a given month takes the form of an unbalanced data set at the end of the sample. To solve this issue, we apply the non-parametric methodology to forecast the monthly variables in order to complete the values until the end of the current quarter for GDP nowcasts and until the end of the next quarter for GDP forecasts, then we aggregate the monthly data to quarterly frequencies. We use four various ways to forecast the monthly variables: an ARIMA(p,d,0) approach, the k-NN procedure ($d = 1$ and $d > 1$) with exponential weights and the radial basis function method with various couples (d, k), and various functions $\phi(\cdot, \cdot)$. We present now the four procedures. For the three first methods, the economic indicators have been made stationary in presence of trend.

1. Concerning the ARIMA(p,d,0) procedure, for each economic indicator we use the Akaike criterion AIC for the selection of the lag p. Model's parameters are estimated by least square method.

2. Regarding the NN method when d = 1, we determine the number of neighbors k by min-

[1] www.eabcn.org

[2] http://stats.oecd.org/mei/

imizing : $\sqrt{\frac{1}{n-k}\sum_{t=k}^{n-1}||\hat{X}_{t+1}^i - X_{t+1}^i||^2}$, $i = 1, \cdots, 13$, where n is the sample size, $\hat{X}_{t+1}^i$ is the estimate of the i-th economic indicator X_{t+1}^i obtained from (2.1) with $d = 1$. When the number k is determined for the horizon h = 1, we used it to calculate the forecasts for $h > 1$. This work is done for each economic indicator, and therefore the number of neighbors k may not be the same for all indicators.

3. The multivariate k-NN method: (i) we embed the initial serie $X_1, ..., X_n$ in a space of dimension d building vectors $\{\underline{X}_d, \underline{X}_{d+1}, ..., \underline{X}_n\}$ in $\mathbb{R}^d$, where $\underline{X}_i = (X_{i-d+1}, ..., X_i)$; (ii) we determine the k nearest vectors of $\underline{X}_n$ inside these vectors, and denote $r_i = \|\underline{X}_n - \underline{X}_i\|$, $i = d, d+1, ..., n-1$, the distance between these vectors. We order the sequence $r_d, r_{d+1}, ..., r_{n-1}$ such that $r_{(d)} < r_{(d+1)} < ... < r_{(n-1)}$, and we detect the vectors $\underline{X}_{(j)}$ corresponding to $r_{(j)}$, $j = d, d+1, ..., d+k-1$; (iii) to compute $m_n(\underline{X}_n) = \hat{X}_{n+1}$, we use the expression (2.1). It may be noted that we obtain the one step ahead forecast. Finally we use the information set: $X_1, ..., X_n, \hat{X}_{n+1}$ instead of $X_1, ..., X_n$ and redo step (i)-(iii), to get the two steps ahead forecast, and so on. We keep the couple (d, k) which minimizes $\sqrt{\frac{1}{n-k-d}\sum_{t=k+d}^{n-1}||\hat{X}_{t+1}^i - X_{t+1}^i||^2}$ for each indicator.

4. The RBF method: given a d-dimensional space, (i) we determine k clusters using a k-means clustering: this method permits to determine the centers and the radii (c_i, r_i), $i = 1, ..., k$ characterizing the clusters; (ii) for a given function ϕ, the vectors in $\mathbb{R}^d$ are then grouped inside the k clusters. We estimate the width r_i using the r centers c_j $(r \leq k)$ which are closest to c_i, such that, for $i = 1, ..., k$, $r_i = \frac{1}{r}\sqrt{\sum_{j=1}^r \|c_i - c_j\|^2}$ and finally the weights w_i are estimated by ordinary least squares method; (iii) the one-step-ahead value obtained with the RBF method is given by the relationship (2.2); (iv) It is well known that k-means clustering provides local minimum, thus, we repeat this algorithm many times keeping parameters wich minimize the RSS $\sum_{j=d}^n (X_{j+1} - m(\underline{X}_j))^2$. Then, we consider the information set: $X_1, ..., X_n, \hat{X}_{n+1}$ instead of $X_1, ..., X_n$ and redo step (i)-(iv) to get the two steps ahead forecast, and so on. Again, the RBF method does not need to make the series stationary, which constitutes a great advantage of the method in comparison with the three other ones. Here, the parameter d vary between 2 and 5, and k between 3 and 7, and can be different for each economic indicator.

As soon as the four modellings are retained, we compute the GDP flash estimates that were released in real-time by Eurostat from the first quarter of 2003 to the third quarter of 2007 using the previous forecasts of the monthly indicators. According to this scheme, the monthly series have to be forecast for an horizon h varying between 3 and 6 months in order to complete the data set at the end of the sample. Recall that the h-step-ahead predictor for $h > 1$ is estimated recursively starting from the one-step-ahead formula.

Using five years of vintage data, from the first quarter 2003 to the third quarter 2007, we provide RMSEs for the Euro area flash estimates of GDP growth in genuine real-time conditions. We have computed the RMSEs for the quarterly GDP flash estimates, obtained with the four forecasting methods used to complete adequately in real-time the monthly indicators, that is ARIMA, k-NN ($d = 1$ and $d > 1$) and RBF methods. More precisely, we provide the RMSEs of the combined forecasts based on the arithmetic mean of the eight BE of Diron (2008). Thus, for a given forecast horizon h, we compute $\hat{Y}_t^j(h)$ which is the predictor stemming from Diron's equations $j = 1, \cdots, 8$, in which we have plugged the forecasts of the monthly economic indicators, and we compute the final estimate GDP at horizon h: $\hat{Y}_t(h) = \frac{1}{8}\sum_{j=1}^{8} \hat{Y}_t^j(h)$. The RMSE criterion for the final GDP is $RMSE(h) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t(h) - Y_t)^2}$, where T is the number of quarters between Q1 2003 and Q4 2007 (in our exercise, $T = 19$) and Y_t is the Euro area flash estimate for quarter t. The RMSE errors for final GDP are provided in table 2 and comments follow.

The accuracy of the nowcasting and forecasting increases as soon as the information set grows. Indeed, for all the four methods, if the forecast horizon reduces from $h = 6$ to $h = 1$, the RMSEs becomes lower. Few days before the publication of the flash estimate (around 13 days with $h = 1$), the lowest RMSE is obtained with the RBF method (RMSE=0.170). For all horizons h from 6 to 1, we can see that we have smaller RMSE using non parametric methods (RBF and k-NN) than using linear modelling. Moreover, with the nonparametric procedures, we obtain smaller error if we work in multivariate setting than in univariate approach. Concerning RBF and k-NN methods in multivariate setting, the k-NN method provides lower RMSE for $h = 3, 5$, and for $h = 1, 2, 4, 6$ we get better results using RBF method: thus these two approaches appear competitive to predict the GDP Euro area. Finally, they give always smallest error than the methods developed in the univariate setting. This last result confirms the importance to work in a multivariate framework for GDP computing. Such remark has already been done by authors using factor models (Kapetanios and Marcellino, 2006). The next step will be to compare both multivariate settings: parametric and non-parametric modellings, although there exists a big difference between these two modellings due to the fact that factor models typically use a large number of factors to be efficient, which is not the case here. Nevertheless this work has to be done and will be the purpose of a companion paper.

References

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

Clements, M.P. and D.F. Hendry (1999) *Forecasting non-stationary economic time series*, Cambridge: MIT Press.

Darne, O. (2008) "Using business survey in industrial and services sector to nowcast GDP growth: The French case" Economics Bulletin **3**, 1-8.

Diron, M. (2008) "Short-term forecasts of Euro area real GDP growth: an assessment of real-time performance based on vintage data" Journal of Forecasting **27**, 371-390.

Guégan, D. (2003) *Les Chaos en Finance: Approche Statistique*, Economica Série Statistique Mathématique et Probabilité: Paris.

Guégan, D. and P. Rakotomarolahy (2009) "The Multivariate k-Nearest Neighbor Model for Dependent Variables: One-Sided Estimation and Forecasting" Centre d'Economie de la Sorbonne working paper number 50.

Kapetanios, G. and M. Marcellino (2006) "A parametric estimation method for dynamic factors models of large dimensions" IGIER working paper number 305.

Marcellino, M. and C. Schumacher (2008) "Factor-MIDAS for now and forecasting with ragged-edge data: A model comparison for German GDP" CEPR working paper number 6708.

Stute, W. (1984) "Asymptotic normality of nearest neighbor regression function estimates" Annals of Statistics **12**, 917-926.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapmann and Hall: London.

Yakowitz, S. (1987) “Nearest neighbors method for time series analysis” Journal of Time Series Analysis **8**, 235-247.

Yatchew, A.J. (1998) “Nonparametric regression techniques in economics” Journal of Economic Literature **36**, 669-721.

We provide the list of the monthly economic indicators used in this study for the computation of the GDP using the bridge equations (BE).

Table 1: Summary of the thirteen economic indicators of Euro area used in the eight GDP bridge equations.

Short Notation	Notation	Indicator Names	Sources	Period
I^1	IPI	Industrial Production Index	Eurostat	1990-2007
I^2	CTRP	Industrial Production Index in Construction	Eurostat	1990-2007
I^3	SER-CONF	Confidence Indicator in Services	European Commission	1995-2007
I^4	RS	Retail sales	Eurostat	1990-2007
I^5	CARS	New passenger registrations	Eurostat	1990-2007
I^6	MAN-CONF	Confidence Indicator in Industry	European Commission	1990-2007
I^7	ESI	European economic sentiment index	European Commission	1990-2007
I^8	CONS-CONF	Consumers Confidence Indicator	European Commission	1990-2007
I^9	RT-CONF	Confidence Indicator in retail trade	European Commission	1990-2007
I^{10}	EER	Effective exchange rate	Banque de France	1990-2007
I^{11}	PIR	Deflated EuroStock Index	Eurostat	1990-2007
I^{12}	OECD-CLI	OECD Composite Leading Indicator, trend restored	OECD	1990-2007
I^{13}	EUROCOIN	EuroCoin indicator	Bank of Italy	1999-2007

Table 2: RMSE for the estimated mean quarterly GDP, using AR, k-NN ($d = 1$ and $d > 1$), and RBF predictions for the monthly indicators.

h	ARIMA	RBF	k-NN(1)	k-NN(d>1)
6	0.249	0.194	0.198	0.214
5	0.221	0.196	0.203	0.192
4	0.216	0.186	0.202	0.196
3	0.195	0.178	0.186	0.177
2	0.191	0.175	0.176	0.177
1	0.175	0.170	0.174	0.171