GDP nowcasting with ragged-edge data: a semi-parametric modeling

Laurent Ferrara, Dominique Guegan, Patrick Rakotomarolahy

GDP nowcasting with ragged-edge data: A semi-parametric modelling *

Laurent Ferrara†, Dominique Guégan‡, Patrick Rakotomarolahy§

Abstract

This paper formalizes the process of forecasting unbalanced monthly data sets in order to obtain robust nowcasts and forecasts of quarterly GDP growth rate through a semi-parametric modelling. This innovative approach lies on the use on non-parametric methods, based on nearest neighbors and on radial basis function approaches, to forecast the monthly variables involved in the parametric modelling of GDP using bridge equations. A real-time experience is carried out on Euro area vintage data in order to anticipate, with an advance ranging from six to one months, the GDP flash estimate for the whole zone.

Keywords: Euro area GDP - Real-time nowcasting - Forecasting - Non-parametric methods.

1 Introduction

The monetary policy decisions in real time are enlighted by the assessment of current economic conditions based on incomplete economic data and by the anticipations of future economic fluc-

*This paper has been presented at the *5th Eurostat Colloquium on Modern Tools for Business Cycle Analysis* in October 2008 at Luxembourg. The second author would like to thank G.-L. Mazzi and M. Marcellino for the invitation.

†Banque de France, Business Conditions and Macroeconomic Forecasting Directorate, 31 rue Croix des Petits Champs, 75049 Paris Cedex 01, France, e-mail:laurent.ferrara@banque-france.fr. The views expressed herein are those of the author and do not necessarily reflect those of the Banque de France.

‡Paris School of Economics, CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: dguegan@univ-paris1.fr

§CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: rakotopapa@yahoo.fr

tuations. Since most data are released with a lag and are subsequently revised, the estimation of current and next quarters for GDP growth and other key variables is an important task for central banks.

In short-term economic analysis and forecasting teams of central banks, monthly indicators are routinely used to assess the current economic conditions before GDP figures are made available by statistical offices. For example in the Euro area, the first GDP estimation, referred to as *flash estimate*, is released by Eurostat around 43 days after the end of the reference quarter. Therefore, during this interval of time, econometric models are requested in order to provide a quantitative evaluation of the economic activity. Models able to perform this exercise must exploit timely information, deal with mixed frequency (using monthly data to nowcast quarterly GDP) and ragged-edge data (at the end of the sample, different variables will have missing points corresponding to different dates in accordance with their timeliness).

A well-known particular class of models that possesses those characteristics goes under the name of bridge equations. These are predictive equations that bridge monthly variables with quarterly ones. More precisely, bridge equations are regressions of quarterly GDP growth on a small set of key monthly indicators. This simple modelling strategy has been popular among policy institutions which commonly pooled several GDP forecasts from bridge equation models so as to consider a large number of predictors.

In this paper, we develop an innovative method based on parametric bridge equations associated with non-parametric estimation of the variables in order to nowcast and to forecast quarterly GDP. Indeed, we forecast the monthly indicators involved in equations through non-parametric methods, specifically the k-nearest neighbors (k-NN) method and the radial basic functions (RBF) method. These forecasts will then be plugged inside the bridge equations, permitting to get semi-parametric nowcasts and forecasts for the GDP growth. We validate this semi-parametric approach by carrying out a true real-time experience on the Euro area that uses vintage data in order to nowcast the flash estimate of the current quarter and to forecast the one of the next quarter.

The paper is organized as follows. Section two gives a quick review on the main approaches used to nowcast and forecast GDP growth. Section three is devoted to the original method put forward in this paper to provide non-parametric estimates for monthly indicators. The section four provides the nowcast and forecast values for the GDP growth flash estimates using an innovative semi-parametric modelling and section five concludes.

2 Nowcasting GDP: A review

GDP nowcasting and forecasting is a topic of great interest for monetary policy-makers. Indeed, until the first publication of the GDP figures for a given quarter by national statistical institutes, there is a strong need to assess the current economic climate in order to conduct an optimal monetary policy. To nowcast or forecast GDP, forecasters need to take into account a large amount of information which arrives sequentially. Thus, new information becomes available continously throughout the quarter and the nowcasts and forecasts may be adjusted in response to those changes. However, macroeconomic forecasters and analysts face several constraints inherent to the data.

First, it is not obvious to exploit all the available information, as economic and financial indicators are released in an asynchronous way. Due to these different publication lags, multivariate datasets typically exhibit complicated patterns of missing values at the end of the sample and imply unbalanced samples for estimation. This leads to the so-called 'ragged-edge' data problem in econometrics, originally pointed out by Wallis [1] and forecasting methods are thus necessary to tackle this issue.

Second, we notice that monthly indicators are generally used to explain a quarterly variable, namely GDP growth rate. We face therefore a mixed frequency problem. A common solution in such cases consists in pre-filtering the data so that the left-hand and right-hand side variables in the equation could be available at the same frequency (for example by aggregating monthly variables into quarterly ones). An alternative way is the Mixed Data Sampling (MIDAS) regressions introduced by Ghysels *et al.* [2].

Third, a part of the recent literature discusses the issue of the amount of information which is necessary to get robust GDP estimates. After nearly ten years during which researchers have tried to determine the optimal amount of data which is necessary to get relevant estimates of the GDP, the answer seems mitigated: we refer to Marcellino *et al.* [3], Bernanke and Boivin [4], Forni *et al.* [5], Boivin and Ng [6], d'Agostino *et al.* [7] for deep discussions on this problem.

Recently, several approaches based on dynamic factor models have been developed to tackle those three previous issues, we refer to Forni *et al.* [8], Stock and Watson [9], Doz *et al.* [10] [11], Kapetanios and Marcellino [12], Giannone *et al.* [13] or Barhoumi *et al.* [14], among others. In this paper, we rather focus on linear bridge equations that link quarterly GDP growth rate and monthly economic variables. The classical bridging approach is based on linear regressions of quarterly GDP growth on a small set of key monthly indicators carefully selected or identified through automatic procedures. Applications to this approach are provided in Kitchen and Monaco [15], Rünstler and Sedillot [16], Baffigi *et al.* [17] or Diron [18]. This simple modelling strategy has been popular among policy institutions which commonly pooled several GDP forecasts from bridge equation models so as to consider a large number of predictors. We can distinguish several implementations of this approach. One combines a number of selected bridge equations based on multiple regressors, (Rünstler and Sédillot [16], and Diron, [18]), while another one pools forecasts of GDP based on large number of bridge equations with only one predictor each (Kitchen and Monaco, [15]). An active strand of the literature also focuses on bridging GDP with composite variables estimated through dynamic factor models with different kinds of methodology, see for example Angelini *et al.* [19], Schumacher and Breitung [20] and Marcellino and Schumacher [21].

In the present work, we follow the methodology proposed and developped by Diron [18] to compute *in fine* the quarterly GDP growth for the Euro area. We use the same eight linear equations introduced in her paper and our contribution concerns the way to forecast the auxiliary economic indicators involved in those equations. We put forward an alternative forecasting method based on a non-parametric approach that differs from the linear autoregressive projections generally used in this framework. Two main advantages of this approach are respectively the facts that we are free of the assumption of stationarity and of the knowledge of the distribution function

of the data sets on which we work.

3 An innovative tool for GDP forecasts: A non-parametric approach

In order to provide a monthly flash estimate of the quarterly GDP, we put forward an innovative statistical tool to complete the monthly information set used to nowcast and forecast the quarterly GDP growth.

Here, our approach permits to provide robust forecasts for the monthly variables via non-parametric methods and to use the predictive linear parametric equations to bridge monthly information with quarterly one. *In fine,* this approach will be referred to as *a semi-parametric modelling for quarterly GDP* in the sense that we combine non-parametric methods to forecast the monthly variables and parametric equations to estimate GDP growth rate.

In this section, we present two non-parametric methods that we use to forecast the monthly indicators, namely the nearest neighbors method and the radial basis function method.

3.1 Why the non-parametric approach ?

Working with non-parametric approaches permits to be free from a lot of assumptions concerning the datasets on which we work. First, the methods that we use in the following are close to non-linear modelling and then appear more flexible than the classical linear regressions that are generally used in the literature. Second, it is not necessary to make the data stationary: this means that we can work with data presenting some trend for instance. Third, the way to estimate parameters appearing in the various modelling does not necessitate to know the distribution of the data sets and thus it is free from the usual Gaussian assumption.

The methods that we use to estimate the monthly indicators are based on the estimation of a regression without using parametric modelling. We consider the simple problem which consists to estimate the relationship between two random variables, say X and Y. Regression analysis is concerned with the question of how Y (the dependent variable) can be explained by X (the

explanatory or regressor variable). This means the existence of a relationship of the form:

$$Y = m(X), \tag{3.1}$$

where $m(\cdot)$ is a function in the mathematical sense. We do not assume here any specific hypotheses on $m(\cdot)$. The random variables X, in the expression (3.1), can be the past values of the random variables Y_t. Given $\{(Y_1, X_1), ..., (Y_t, X_t)\}$, a set of observations identically distributed as (X, Y), then, the underlying principle that theoretical laws usually do not hold in every individual case but merely on average is considered here and can be formalized as:

$$E[Y|X = x] = m(x), \tag{3.2}$$

or as:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, ..., t. \tag{3.3}$$

Equation (3.3) says that the relationship $Y = m(X)$ does not need to hold exactly for the ith observation but is disturbed by the random variable ε. Here, we will assume that the estimator of $m(\cdot)$ has the following form:

$$\hat{m}(x) = \sum_{i=1}^{t} \omega_{i,t}(x) Y_i, \tag{3.4}$$

where $\omega_{i,t}$ are weights to specify and to estimate. The choice of these weights has led to many research papers, we refer for instance to Yakowitz [22], Girosi and Anzelatti [23] and Finkenstadt and Kubier [24].

In this paper we focus on k-NN weights and on RBF weights. Other methods could also be considered because of their interesting properties, namely the wavelets and the neural network methods. We do not consider them in this paper, but this issue will be adressed in a companion paper. Another well known weight regressor is the kernel regressor, but we do not consider it here because of its static representation and of strong assumptions on the data set to get robust estimates. Some references on non-parametric methods are, among others, Silverman [25], Wand and Jones [26], Devroye et al. [27], Guégan [28], Härdle *et al.* [29].

3.2 Two non-parametric modelling

In the following we assume that at time t, we observe a random variable X_t and its past values $X_1, \cdots, X_{t-1}$ for estimation purposes. Thus, formally our aim is to estimate the function $m(\cdot)$

such that:

$$X_t = m(\underline{X}_{t-1}) + \varepsilon_t, \tag{3.5}$$

where $\underline{X}_{t-1}$ is a set of variables taken in the past values of X_t. In our framework X_t represents a monthly economic indicator known until date t and our aim is to forecast values at time $t+1, t+2, ..., t+H$ where H is the maximum forecast horizon.

1. **The k-Nearest Neighbors method**

   The k-nearest neighbor (k-NN) estimator of equation (3.4) can be viewed as a weighted average of the response variables in a neighborood around x, with the important point that the neighborood width is not fixed but variable. To be more specific, the values of X_{t+1} used in computing the average estimate, are those which belong to the k observed values of X_t that are nearest to the point x, at which we would like to estimate $m(x)$.

   To estimate $m(\cdot)$ given in (3.5), assuming that we are at time t, among the observations $X_1, \cdots, X_{t-1}$, we are looking at the k closest neighbors of X_t. If we denote those k points by $X_{(i)}$, $i = 1, ..., k$, then an estimate of $m(\cdot)$ by the k-NN method permits to compute the one-step-ahead predictor $\hat{X}_{t+1}$ such that:

$$\hat{X}_{t+1} = \sum_{i=1}^{k} w(X_t - X_{(i)}) X_{(i)+1}$$

   where the weights $w(\cdot)$ are usually either exponential or uniform functions given by:

   - exponential weight : $w(x - X_{(i)}) = \frac{\exp(-(x-X_{(i)})^2)}{\sum_{i=1}^{k} \exp(-(x-X_{(i)})^2)}$
   - uniform weight :$w(x - X_{(i)}) = \frac{1}{k}$.

   If we estimate $m(\cdot)$ at a point x where the data are sparse then it might happen that the k-nearest neigbors are rather far away from x (and each other), thus consequently we end up with a wide neighborood around x for which an average of the corresponding values of X_{t+1} is computed. It is noteworthy that the parameter k needs to be estimated. Note also that the variance of this estimate does not depend on the distribution of the data set, which constitutes a great difference with the kernel method.

2. **The Radial Basis Function method**

The radial basis function (RBF) method is a particular case of what we call the spline smoothing. Spline smoothing provides a natural and flexible approach to curve estimation, which copes well whether or not the design points are regurlarly spaced. The smoothing spline's behaviour is intermediate between fixed kernel smoothing and smoothing on an average of a fixed number of neighboring values. Such intermediate behavior is desirable because, in a certain sense, moving from fixed kernel to nearest neighbors method over-compensates for effects caused by the variability in density of design points.

The motivation of this last method is based on the estimation of $m(x)$ by minimizing the residuals sum of squares (RSS) error, such that $RSS = \sum_{j=d}^{t}(X_{j+1} - m(\underline{X}_j))^2$, interpolating the data without exploiting any structure that might be present in the data. Generally we use spline estimator having good properties like to be twice differentiable with boundary conditions on the second derivatives.

Here, we consider, as spline estimator, the radial basis functions which are defined in the following way. A radial basis function ϕ is an application defined from $\mathbb{R}^d \times \mathbb{R}$ to $\mathbb{R}^d$ and is characterized by its centroid c and its width r. This means that the observations X_t are embedded in a space of dimension d, in which the parameters which characterized this function will be estimated. Using the same information set as before, the estimation of m by a set of k clusters through a radial basis functions ϕ is given by:

$$\hat{X}_{t+1} = w_0 + \sum_{i=1}^{k} w_i \phi(\|\underline{X}_t - c_i\|, r_i),$$

where $\underline{X}_t = (X_t, X_{t-1}, \cdots, X_{t-(d-1)}) \in \mathbb{R}^d$, ϕ is the radial basis function and $||\cdot||$ is the Euclidean norm. The parameters to estimate are c_i, r_i and w_i. The radial basis function can be chosen, for instance, among the following functions:

- Spline function : $\phi(x, r) = \frac{x^2}{r^2}\log(\frac{x}{r})$,
- Gaussian function : $\phi(x, r) = \exp(-\frac{x^2}{2r^2})$,
- Multiquadric function : $\phi(x, r) = \sqrt{x^2 + r^2}$,
- Inverse multiquadric : $\phi(x, r) = \frac{1}{\sqrt{x^2+r^2}}$.

The procedure to estimate the parameters follows the k-means cluster method. k-means method corresponds to a partition of the individuals into clusters such that each indivual belongs to the cluster whose center is closest in Euclidean distance. Then, the centroids are the centres of each cluster. To determine all these parameters, we first embed the data set in a space of dimension d, as the embedding dimension d is not known a priori, we proceed step by step, beginning at $d = 1, ...$ Then given a d-dimensional space, we use a k-means clustering method to partition the points such that each one belongs to the cluster whose centre is the closest in the Euclidean distance sense (the procedure is done for a given function ϕ). Thus, the centroids are the centres of each cluster. At this step, the observations in $\mathbb{R}^d$ are organized around k clusters with their centres $c_i, i = 1, ..., k$. Then, we estimate the width r_i using the r centers c_j $(r \leq k)$ which are closest to c_i, such that, $i = 1, ..., k$:

$$r_i = \frac{1}{r}\sqrt{\sum_{j=1}^{r} \|c_i - c_j\|^2}.$$

This choice permits to avoid that the clusters overlap. As soon as the function ϕ and the parameters (c_i, r_i), $i = 1, ..., k$ are known, then $\phi(\|\underline{X}_t - c_i\|, r_i)$ is known and the function $m(x)$ is linear in w_i, so we can estimate w_i by ordinary least squares method. Finally the one-step-ahead value obtained with the RBF method is given by:

$$\tilde{X}_{t+1} = \hat{w}_0 + \sum_{i=1}^{k} \hat{w}_i \phi(\underline{X}_t - \hat{c}_i, \hat{r}),$$

where $\phi(x, r)$ is one of the previous radial basis function.

4 Semi-Parametric Modelling for GDP growth

In this section we put forward a semi-parametric modelling to estimate nowcasts and one-quarter-ahead forecasts for the Euro area GDP growth, based on the parametric equations proposed by Diron [18] in which we have plugged the non-parametric estimates of the monthly indicators obtained using the two previous methods.

4.1 The principle

In order to nowcast and forecast the quarterly GDP quarter-over-quarter growth rate, we assume that the indicators have already been chosen and we forecast them using the non-parametric

methods presented above. We select the parametric GDP modelling proposed in the paper of Diron [18] and we marginally adapt it in order to take new values into account. This approach has been developed and used by the ECB in their monthly assessment of the GDP in the Euro area [30].

In table 1, we provide the name and the period of the thirteen indicators that we use in this study, as well as their source. To compute the GDP, we use eight equations provided in section 6. The first two bridge equations are close to simple accounting the industrial production in order to reproduce GDP computations. The other bridge equations rely on more indirect indicators, such as opinion surveys or financial data. The equations are various combinations of a small set of selected indicators for the Euro area: industrial production in the manufacturing industry and in construction, retail sales, new car registrations, the European Commission confidence indicators, the OECD leading indicator and the CEPR-Bank of Italy coincident indicator for the Euro area, referred to as EuroCoin. These indicators are used in the individual eight equations that are described in the Annex.

4.2 The monthly data set

4.2.1 Description

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. We use the real-time data base provided by EABCN through their web site [1]. 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. As underlined previously in the paper, 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.

[1] www.eabcn.org

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

4.2.2 Forecasting procedures

We use three various ways to forecast the monthly variables: an ARIMA approach, the k-NN procedure with exponential weights and the radial basis function method with various couples (d, k), and various functions $\phi(\cdot, \cdot)$. Following Diron's approach [18], we have retained for all the indices an ARIMA(6,1,0) model.

Concerning the k-NN procedure, we look after the best value for k where $k = 1, ..., 5$, for each monthly indicator, in the sense of minimizing the mean root mean squared error (RMSE) for a forecast horizon varying from $h = 1$ until $h = 6$. We retain the following values : $k = 1$ for the indicators I^4; $k = 2$ for the indicator I^1; $k = 3$ for the indicators I^7 and I^{11}; $k = 4$ for the indicators I^5, I^{10} and I^{13}; $k = 5$ for the indicators I^2, I^3, I^6, I^8, I^9 and I^{12}. It is noteworthy that significantly more accurate results are obtained for the series I^1, I^2 and I^4, namely IPI, IPI in construction and retail sales, when differencing the data. Indeed, those latter series are clearly non-stationary and k-NN forecasting based on data in differences improves the results.

Considering the RBF method, for each monthly indicator, we search for the best couple (d, k) varying $d = 2, ..., 5$ and $k = 3, ..., 7$, in order to minimize the mean root mean squared error , and finally we retain the following sets of values: $(d = 4, k = 4)$ and multiquadric radial function for the indicator I^1; $(d = 2, k = 4)$ and Gaussian radial function for indicators I^2 and I^{10}; $(d = 5, k = 4)$ and Gaussian radial function for the indicator I^3; $(d = 4, k = 7)$ and Gaussian radial function for indicators I^5, I^4, and I^6; $(d = 3, k = 4)$ and Gaussian radial function for indicators I^7 and I^9; $(d = 2, k = 7)$ and Gaussian radial function for the indicator I^8; $(d = 2, k = 3)$ and Gaussian radial function for the indicator I^{11}; $(d = 4, k = 3)$ and Gaussian radial function for the indicator I^{12}; $(d = 3, k = 7)$ and Gaussian radial function for the indicator I^{13}. Using RBF method, there is no need to make the series stationary which constitutes a great advantage of this method in comparison with the others.

4.3 The real-time GDP growth modelling

Our aim is to estimate 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 non-parametric forecasts of the monthly indicators that we have previously introduced.

In this paper, we provide a true real-time analysis. We assume that GDP nowcasts and forecasts are computed at each end of the month, as soon as the opinion surveys are released by the European Commission, that is around the last working day of the month. Thus, for a given month, we assume that we have only access to the information available at this time. Moreover, GDP values are those that were available at this exact date. For a quarter Q_i, $i = 1, 2, 3, 4$, of a given year six GDP estimates are provided, namely 3 nowcasts and 3 forecasts. The first nowcast is estimated at the end of the second month of this quarter, the second at the end of the third month of this quarter and the last at the the end of first month of the next quarter. It turns out that the last nowcast is done around 13 days before the flash estimate release. It is worth noticing that the last nowcast can also be seen as a *backcast* in the sense that the estimation is done once the quarter is finished (end of April for Q1, end of July for Q2, end of October for Q3 and end of January of next year for Q4). For this given quarter Q_i, the first forecast is done at the end of the second month of the previous quarter, the second is done at the end of the third month of the previous quarter and the last one is done at the end of the first month of the quarter. Technically, this last forecast could be called a nowcast in the sense that the quarter has already started.

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. Note that the h-step-ahead predictor for $h > 1$ is estimated recursively starting from the one-step-ahead formulas.

4.4 Empirical results

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 three forecasting methods used to complete adequately in real-time the monthly indicators, that is ARIMA, k-NN and RBF. More precisely, we provide the RMSEs of the combined forecasts based on the arithmetic mean of these eight equations. Thus, for a given forecast horizon h, the RMSE

criterion is defined as follows:

$$RMSE(h) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{X}_t(h) - X_t)^2}, \tag{4.1}$$

where n is the number of quarters between Q1 2003 and Q4 2007 ($n = 20$), X_t is the Euro area flash estimate for quarter t and $\hat{X}_t(h)$ is the h-months ahead predictor of X_t for a given forecasting method (ARIMA, k-NN or RBF) such that:

$$\hat{X}_t(h) = \frac{1}{8}\sum_{j=1}^{8}\hat{X}_t^j(h), \tag{4.2}$$

where $\hat{X}_t^j(h)$ is the predictor stemming from equation EQj, $j = 1, \ldots, 8$. The result is presented in table 2. The following comments point out the main results:

1. As expected, the precision increases as well as the information set is getting bigger. Indeed, for all the three methods, as the forecast horizon reduces from $h = 6$ to $h = 1$, the RMSEs tend be 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).

2. The non-parametric RBF method permits to improve kindly the ARIMA projection method as regards the accuracy of the GDP nowcasts (from $h = 1$ to $h = 4$), while the results obtained with the k-NN method are less conclusive. GDP forecasts for longer horizon ($h = 5$ and $h = 6$) are less accurate with k-NN and RBF methods than with ARIMA model and need therefore improvement. While we systematically implement an integrated model of order one for the ARIMA method, we let the series unchanged with the RBF approach. It turns out that for equations 1 and 2 involving only non-stationary variables, forecasting with ARIMA models provide lower quadratic errors for all equations.

3. If we make differentiation on some indicators in order to achieve stationarity, we obtain slightly better results with the k-NN approach, while results are roughly similar with the RBF approach.

5 Conclusion

A key feature of this paper is that we examine the forecast performance taking into account the real-time data flow, that is, the non-synchronous release of monthly information throughout the

quarter. In this paper we combine bridge equations, developed to link monthly economic variables with quarterly GDP growth, and non-parametric forecasting of those monthly variables. Using vintages data on activity of the Euro area, this paper assesses the impact of new data release on short-term forecasts of the Euro area quarter-over-quarter GDP growth.

The main contribution of this paper is a new way to complete ragged-edge monthly data which appear in the estimation of the quarterly GDP growth, using non-parametric methods based on nearest neighbors and radial basis function approaches. We show that this new approach provides forecasts whose errors are smaller than errors' forecasts based on linear modelling. The interest of the method is to obtain relevant forecasts being free of classical assumptions of Gaussianity and stationarity.

A number of questions arise from this exercise: the main point is the choice of the model that we use, say the bridging equations. This latter assumption is very important. We observe that the averaging procedure generally enables to get more precise results than each of the individual equation, underlying the interest of this approach for forecasting purposes in spite of its simplicity. This empirical result has been pointed out in many empirical studies. It would be interesting consider more refined statistical procedures to combine the forecasts. Thus, we need to develop further models in order to improve and robustify the model itself: number of equations, choice of indicators, characteristic of the model like for instance the linearity.

Another aspect of this work is to point the influence of non-stationarity for some indicators and to open the discussion on the best way to take this feature into account.

Finally, the non-parametric estimates that we used are consistent (Yakowitz [22]) and then the forecasts too, being obtained through linear equations containing these estimates. Now, if the almost surely consistency and L^2 consistency of these estimates have been obtained in the dependence case, their asymptotic normality has to be proved for any weight. This work is actually in progress and will permit to provide, in the future, confidence intervals for GDP forecasts.

6 Annex: Regression equations

Let us define X_t as:

$$X_t = (\log GDP_t - \log GDP_{t-1}) \times 100.$$

1. EQ1. The first equation combines indicators of production in the main sectors of activity: industrial production index (I^1) and construction production index (I^2) as well as the confidence index in the services (I^3) as a proxy for the production in this sector :

$$X_t = a_0^1 + a_1^1(\log I_t^1 - \log I_{t-1}^1) + a_2^1(\log I_t^2 - \log I_{t-1}^2) + a_3^1 \log I_{t-1}^3 + \varepsilon_t. \tag{6.1}$$

2. EQ2. The second equation uses monthly information on private consumption which includes retail sales (I^4) and new passenger registrations (I^5). No monthly data are available on consumption of services. We add the industrial production index (I^1) and construction production index (I^2):

$$X_t = a_0^2 + a_1^2(\log I_t^1 - \log I_{t-1}^1) + a_2^2(\log I_t^2 - \log I_{t-1}^2) + a_3^2(\log I_t^4 - \log I_{t-1}^4) + a_4^2(\log I_t^5 - \log I_{t-1}^5) + \varepsilon_t. \tag{6.2}$$

3. EQ3. The third equation relates real GDP growth to the European Sentiment Index (I^7), which is an average of 5 confidence indexes:

$$X_t = a_0^3 + a_1^3 I_t^7 + a_2^3 I_{t-1}^7 + \varepsilon_t. \tag{6.3}$$

4. EQ4. The fourth equation relates GDP growth to industry and services confidence indexes (respectively, I^6 and I^3):

$$X_t = a_0^4 + a_1^4(I_t^6 - I_{t-1}^6) + a_2^4 I_t^3 + \varepsilon_t. \tag{6.4}$$

5. EQ5. The fifth equation uses survey data including consumer confidence index (I^8), which aims at capturing developments in consumption, and manufacturing and retail trade confidence index ((I^6 and I^9) :

$$X_t = a_0^5 + a_1^5(I_t^6 - I_{t-1}^6) + a_2^5 I_t^9 + a_3^5 I_t^8 + \varepsilon_t. \tag{6.5}$$

6. EQ6. The sixth equation uses financial variables: effective exchange rate (I^{10}) and stock market price index deflated by the HICP inflation (I^{11}). Note that in opposition to the

paper of Diron (2008), the yield spread between 10-year government bond rates and 3-months interest rates has not been found significant.

$$X_t = a_0^6 + a_1^6(\log I_{t-2}^{10} - \log I_{t-3}^{10}) + a_2^6(\log I_{t-1}^{11} - \log I_{t-2}^{11}) + \varepsilon_t. \tag{6.6}$$

7. EQ7. The seventh and the eighth equations make use of the composite indicators for growth in the Euro area, the OECD leading indicator (I^{12}) and the EuroCoin (I^{13}) respectively:

$$X_t = a_0^7 + a_1^7(\log I_t^{12} - \log I_{t-1}^{12}) + a_2^7(\log I_{t-2}^{12} - \log I_{t-3}^{12}) + a_3^7 X_{t-1} + \varepsilon_t. \tag{6.7}$$

and

8. EQ8.

$$X_t = a_0^8 + a_1^8 I_t^{13} + \varepsilon_t. \tag{6.8}$$

References

[1] Wallis KF. Forecasting with an Econometric model: The ragged edge problem. *Journal of Forecasting* 1986; **5** : 1-13.

[2] Ghysels R, Santa-Clara P, Valkanov R. There is a risk return tradeoff after all. *Journal of Financial Economics* 2005; **76**: 509-548.

[3] Marcellino M, Stock JH, Watson MW. *Macroeconomic forecasting in the Euro area: country specific versus Are-Wide Information*. IGIER WP 2001: No 201.

[4] Bernanke BS, Boivin J. Monetary policy in a data-rich environment. *Journal of Monetary Economics* 2003; **50** : 525-546.

[5] Forni M, Giannone D, Lippi M, Reichlin L. *Opening the black box: structural factor models with large cross-sections*. European Central Bank WP 2005: No 571.

[6] Boivin J, Ng S. Are more data always better for factor analysis. *Journal of Econometrics* 2006; **132**: 169-194.

[7] d'Agostino A, Giannone D, Surico P. *(Un)predictability and macroeconomic stability*. European Central Bank WP 2006: No 605.

[8] Forni M, Hallin M, Lippi M, Reichlin L. The generalized dynamic factor model: Identification and estimation. *The Review of Economics and Statistics* 2000; **82 (4)**: 540-554.

[9] Stock JH, Watson MW. Macro-economic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 2002; **20 (2)** : 147-162.

[10] Doz C, Giannone D, Reichlin L. *A quasi maximum likelihood approach for large approximate dynamic factor models*. CEPR Discussion Paper 2006: No. 5724.

[11] Doz C, Giannone D, Reichlin L. *A two-step estimator for large approximate dynamic factor models based on Kalman filtering*. CEPR Discussion Paper 2007: No 6043.

[12] Kapetanios G, Marcellino M. *A parametric estimation method for dynamic factors models of large dimensions*. IGIER WP 2006: No 305.

[13] Giannone D, Reichlin L, Small D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics* 2008; **55**: 665-676.

[14] K. Barhoumi, S. Benk, R. Cristadoro, A. den Feijer, A. Jakaitiene, P. Jelonek, A. Rua, G. Runstler, K. Ruth, C. van Nieuwenhyyze, Short term forecasting of GDP using large monthly data sets: a pseudo real-time forecast evaluation exercise, European Central Bank, 2008, No 84.

[15] Kitchen J, Monaco R. Real-time forecasting in practice: The US treasury staff's real time GDP forecast system. *Business Economics* 2003; 10-19.

[16] Rünstler G, Sedillot F. *Short-term estimates of Euro area real GDP by means of monthly data*. European Central Bank WP 2003: No 276.

[17] Baffigi AR, Golinelli R, Parigi G. Bridge models to forecast the euro area GDP. *International Journal of Forecasting* 2004; **20** : 447-460.

[18] Diron M. Short-term forecasts of Euro area real GDP growth: an assessment of real-time performance based on vintage data. *Journal of Forecasting* 2008; **27** : 371-390.

[19] Angelini E, Camba-Mendez G, Giannone D, Reichlin L, Runstler G. *Short-term forecasts of Euro area GDP growth*. CEPR Discussion Paper 2008: No 6746.

[20] Schumacher C, Breitung J. Real-time forecasting of GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting* 2008; **24**: 368-398.

[21] Marcellino M, Schumacher C. *Factor-MIDAS for now and forecasting with ragged-edge data: A model comparison for German GDP*. CEPR WP 2008: No 6708.

[22] Yakowitz S. Nearest neighbors method for time series analysis. *Journal of Time Series Analysis* 1987; **8**: 235 - 247.

[23] Girosi F, Anzelotti G. Rates of convergence for radial basis functions and neural networks *Artificial Neural Networks for Speech and Vision* 1993; 97-114.

[24] Finkenstadt B, Kuhbier P. Forecasting nonlinear economic time series: a simple test to accompany the nearest neigbor approach, *Empirical Economics* 1995; **20**: 243 -263.

[25] Silverman BW. *Density Estimation for Statistics and Data Analysis*. Chapmann and Hall: London, 1952.

[26] Wand MP, Jones MC. *Kernel Smoothing*. Chapman and Hall: London, 1995.

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

[28] Guégan D. *Les Chaos en Finance: Approche Statistique*. Economica Série Statistique Mathématique et Probabilité: Paris, 2003.

[29] Hardle W, Muller M, Sperlich S, Werwatz A. *Non-Parametric and Semi-Parametric Models*. Springer Verlag, 2004.

[30] European Central Bank. *S*hort-term forecasts of economic activity in the euro area. Monthly Bulletin 2008: April 2008.

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

h	ARIMA(6,1,0)	RBF	k-NN
6	0.190	0.194	0.198
5	0.191	0.196	0.203
4	0.190	0.186	0.202
3	0.181	0.178	0.186
2	0.177	0.175	0.176
1	0.171	0.170	0.174

Table 2: RMSE for the estimated mean quarterly GDP, using AR, k-NN and RBF predictions for the monthly indicators.

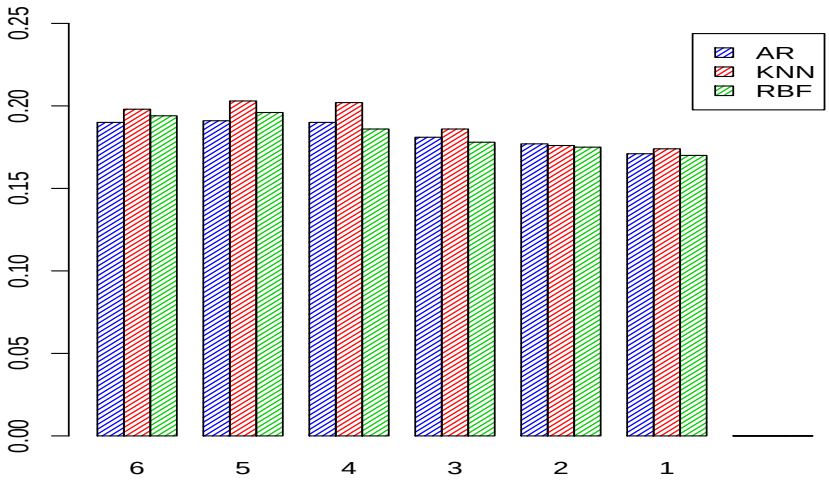

Figure 1: Evolution of RMSEs for the mean of the 8 forecasts according to the forecasting methods for monthly variables.