# Alternative methods for forecasting GDP

Dominique Guegan, Patrick Rakotomarolahy

# Alternative methods for forecasting GDP

Dominique Guégan[*], Patrick Rakotomarolahy[‡]

**Abstract**

An empirical forecast accuracy comparison of the non-parametric method, known as multivariate Nearest Neighbor method, with parametric VAR modelling is conducted on the euro area GDP. Using both methods for nowcasting and forecasting the GDP, through the estimation of economic indicators plugged in the bridge equations, we get more accurate forecasts when using nearest neighbor method. We prove also the asymptotic normality of the multivariate k-nearest neighbor regression estimator for dependent time series, providing confidence intervals for point forecast in time series.

**Keywords:** Forecast - Economic indicators - GDP - Euro area - VAR - Multivariate k nearest neighbor regression - Asymptotic normality.

**JEL:** C22 - C53 - E32.

## 1 Introduction

Forecasting macroeconomic variables such as GDP and inflation play an important role for monetary policy decisions and for assessment of future state of the economics. Policy makers and economic analysts either adapt their theoretical analysis of economic conditions according to the macroeconomic variable forecasts or even probably use them as a support and a justification of their theoretical analysis. Better forecast performance for macroeconomic variables will lead to

[*]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.

[‡]This paper has been presented at the following conferences: $3^{rd}$ CFE in october 2009 Limassol Cyprus, $1^{st}$ ISCEF in february 2010 Sousse Tunisia and $9^{th}$ GOCPS in mars 2010 Leipzig Germany. We would like to thank the organizers of these conferences and participants for their remarks.

better decisions. Two types of methods exist on producing such forecasts, parametric (like the linear Autoregressive models in Box and Jenkins (1970) and the non-linear SETAR-STAR or Markov switching models in Tong (1990) and Pena, Tiao and Tsay (2003)) and non-parametric ones (such as the kernel, the nearest neighbor, the neural network and the wavelet methods in Silverman (1986) and Härdle *et al.* (2004)). The former has had great consideration on economic forecasting which might come from the speed development of its theoretical results on consistent, asymptotic properties and robustness to be in accordance with the existing problems in the method. Although, different problems still have raised concerning strong hypothesis on model specification, estimation and building asymptotic properties among others. While the latter method tries to overcome some of these problems, they avoid on making a priori specification on the distribution of the time series and on the link function. In another word, it is based on the fact that it lets the datas speak to themselves. However there is the cost of more complicated mathematical arguments such as the selection of smoothing parameters. Nevertheless recent studies help to avoid these problems and also the speed of computers that can develop search algorithms from appropriate selection criteria, Devroye and Gyorfi (1985), and Becker, Chambers and Wilks (1988). We favor here these nonparametric tools for time series study. We favor in this work the nonparametric methods for time series study.

We address here the forecasting GDP. This has been studies for along, starting from the growing use of linear autoregressive in this field, Sims (1980) introducing forecasting of macroeconomic variables with linear VAR model on US datas. Later one extension of VAR modelings in Litterman (1986) the Bayesian VAR aiming on reducing VAR's parameters for forecasting US GNP; another extension in Engle and Granger (1987) the Vector Error Correction tackling the presence of cointegration between macroeconomic variables, pointing out possible cointegration of GDP and M2. Recent application on South African GDP using these extensions of VAR modeling is presented in Gupta (2006). The development of methods for forecasting GDP has enlarged first on capturing non-linearity by the use of non-linear modelings like Markov Switching in Hamilton (1989) considering 2-regime for US GNP growth with a conclusion on sharpness/shortness of contractions than expansions and SETAR model in Clements and Krolzig (1998) where they conclude that forecasting performance of such model is conditional on regimes in their application on US GNP growth; second on combining various forecasts from different point of views through the Bridge Equations, Baffigi, Golinelli and Parigi (2004) have showed the superiority of this method on their experiment on euro area GDP, and Diron (2008) has proposed small number of economic indicators in bridge equations for euro area GDP modelling. We know that these studies focused on few number of economic indicators. Another direction consists on forecasting GDP from large number of economic indicators, factor models have been considered for such interest, in Stock and Watson (2002) with a conclusion improving forecast from linear model using large number of monthly US economic indicators, extension to more general dynamic factor models in Bernanke and Boivin (2003), Forni *et al.* (2005) they found that their predictor can provide a substantial improvement of Stock and Watson (2002) model, and Kapetanios and Marcellino (2006). Recently forecasting GDP based on microeconomic foundation comes also in the literature the so called dynamic stochastic general equilibrium models, in Smets and Wouters (2004) using euro area datas, they suggest that the DSGE models compare well with VAR models in terms of forecasting accuracy. In general in the literature the linear univariate ARIMA or multviariate VAR have served as benchmark. We think these models are highly used in practice or should be the starting point on testing forecasting performance of alternative methods. We follow such literature by considering here the VAR modeling as benchmark.

When the final objective of time series analysis is prediction, it is of interest to study the conditional means and conditional variances in some period, given the past of the process. When a point prediction is the final objective, an estimate of some conditional mean is desired, while the conditional variances are needed if interval forecasts are desired.

In this paper, we focus on a non-parametric estimate of the conditional mean. There are numerous non-parametric techniques used in time series analysis to estimate the conditional mean: the kernel methods, the wavelet techniques, the neural networks, the spline functions and the nearest neighbor method ($k$-NN) among others, Prakasa Rao (1983), Donoho and Johnstone (1992), Kuan and White (1994), Friedman (1988), and Mack (1981) for instance.

Given a time series $(X_n)_n$, we consider the following representation for the regression function $m(\cdot)$ associated to this time series:

$$m(x) = E[X_{n+1}|X_n = x]. \tag{1.1}$$

Model (1.1) has the type of a nonlinear regression problem for which many smoothing methods can be used for estimating purpose, Hart (1997).

In this paper, we reconstruct the function $m(\cdot)$ using multivariate $k$-nearest neighbors. We limit ourselves to this method because apart from its advantage risk on model miss-specification or on some strong hypotheses of parametric method, first it could handle in one way the non-linearity on GDP growth from different issues such as the unbalanced economic data sets and publication lags due to the delay of the avalaibility of the GDP in real time, as its ability to capture non-linearity in finance Nowman and Saltoglu (2003) and in other way the presence of asymetries from the complexity of economic dimension, difficulity on exploiting all available informations. Second it has many advantages in practice such as fewer parameters in the model resulting on computational time gain and is certainly the easiest to understand and implement. Moreover kernel smoothing methods are covered when the number of neighbors in nearest neighbors coincides with length of time series. It should be compared with Neural Networks (NN), due to the lack of economic structure of NN methods. Working in a multivariate environment allows us to discover and take into account the structural behavior which cannot always be discerned on a path. Recent results have also made available a method for selecting the number of neighbors within a given space, Ouyang, Li and Li (2006).

Our theoretical result is a contribution to the general problem concerning the non-parametric estimate of a regression with $k$-NN method, extending well known results obtained for independent and identically distributed random variables, Stone (1977), Mack (1981), Devroye (1982), and Stute (1984). In case of dependent variables, Collomb (1984) provides piecewise convergence for univariate variables, and Yakowitz (1987) gets the quadratic mean error for uniformly weighted $k$-NN estimates for univariate samples. Here, working with multivariate time series, we control the bias of a general multivariate $k$-NN estimate, using several weights, and we establish the asymptotic normality of this estimate from which we can construct confidence intervals.

The method of nearest neighbors has been used in finance (forecasting exchange rates, interest rates and index returns) and has raised lots of benefits, Mizrach (1992), Nowman and Saltoglu (2003), and Guégan and Huck (2005) among others, they found forecasts of some European

Monetary System exchange rates superior to random walk, better forecast performance for US interest rates compared with two-factor continuous interest rate models and an improvement of the forecast performance of Dow Jones index returns by changing the weighting schemes, respectively. In economics this method is still less known and therefore rarely used, an interesting review is Yatchew (1998). An interesting area of application in economics would be on forecasting GDP. Indeed, predicting GDP is an important challenge for many institutions, particularly central banks. In these latest institutions, many studies have been developed to solve the problem of nowcasting and forecasting the GDP. But few studies use a non-parametric method due to its recent venue on economic forecasting, it could be relied on long tradition consideration of parametric method in central banks with its availability in many softwares for economic analysis and with the speed development of its asymptotic properties, so the researchers concentrated on improving such method. Apart in our knowledge Tkacz and Hu (1999), and Blake (1999) with neural networks resulting on better forecasts for Canadian GDP and for output growth of six European countries compared with univariate linear model, respectively and Ferrara, Guégan and Rakotomarolahy (2010) and Guégan and Rakotomarolahy (2010) with Nearest Neighbors, and radial basis function methods finding more accurate forecast for euro area GDP with these two methods than with the linear ARIMA modelling.

In this paper, in first step we extend previous works providing a new theoretical result. We explore asymptotic results for the non-parametric estimate of $m(\cdot)$ providing new results which concern the asymptotic normality of the estimate of $m(\cdot)$ for dependent variables, with respect to the bias-variance fit dilemma inherent in this kind of methodology. Our result permits construction of confidence interval which can be helpful for discriminating in practice between different classical methods. Second through an application on GDP we compare forecasts obtained using our methodology with a competitive parametric method based on VAR (Vector Autoregressive) modelling which has shown some benefits for modelling macroeconomic variables, Sims (1980), Webb (1995), Gupta (2006), and inside a Special issue of Journal of Forecasting (1995, vol. 14) different application in economic analysis of VAR families and also because it is a multivariate modelling.

To estimate the GDP, we focus on the method developed by Diron (2008). Her method uses a limited number of economic indicators which are plugged in eight linear equations from which an estimate of GDP is obtained. Some features of Diron's method lie first on bridge equations based methods which have been used alot by economic analysts and forecasters to estimate the GDP and second on forecast combination. Moreover, it incorporates a large dimension of economic activities including different single forecasts based on production sectors, survey datas, financial variables and leading index constructed from large number of economic indicators. In most of the works, the economic indicators are estimated by ARIMA processes, Runstler and Sedillot (2003), and Darne (2008). Our methodology is slightly different of theirs in the sense that we estimate the economic indicators with multivariate nearest neighbors and then comparing it with VAR estimates.

The paper is organized as follows. In Section 2, we establish our theoretical result: the asymptotic normality of the multivariate $k$-NN regression estimate for mixing time series. In Section 3, an empirical forecast exercise is provided. It permits to compare non-parametric and parametric approaches for monthly indicators and their impact in the final GDP estimate. Section 4 concludes and Section 5 is devoted to the proofs.

## 2 Theoretical result

We consider a real time series $(X_n)_n$, and we transform the original data set by embedding it in a space of dimension $d$, building $\underline{X}_n = (X_{n-d+1}, \cdots, X_n) \in \mathbb{R}^d$. The embedding concept is important because it allows to take into account some characteristics of the series which are not always observed on the trajectory in $\mathbb{R}$.

We are interested on getting an estimate of $m(\underline{x})$, $\underline{x} \in \mathbb{R}^d$, using the $k$ closest vectors to $\underline{X}_n = \underline{x}$ inside the training set $S = \{\underline{X}_t = (X_{t-d+1}, \cdots, X_t) \mid t = d, ..., n-1\} \subset \mathbb{R}^d$. We define a neighborhood around $\underline{x} \in \mathbb{R}^d$ such that $N(\underline{x}) = \{i \mid i = 1, \cdots, k(n)$ whose $\underline{X}_{(i)}$ represents the $i^{th}$ nearest neighbor of $\underline{x}$ in the sense of a given distance measure$\}$. Then the $k$-NN regression estimate of $m(\underline{x})$, $\underline{x} \in \mathbb{R}^d$ is given by:

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

where $w(\cdot)$ is a weighting function associated to neighbors and it is noteworthy that the parameter $k$ has to be estimated. A general form for the weights is:

$$w(\underline{x} - \underline{X}_{(i)}) = \frac{\frac{1}{nR_n^d} K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n})}{\frac{1}{nR_n^d} \sum_{i=1}^{n} K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n})}, \tag{2.2}$$

where $R_n$ corresponds to the distance between $\underline{x}$ and the further neighbors, and $K(\cdot)$ is a given weighting function. Two weighting functions have been mostly used, the exponential function $K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n}) = exp(-||\underline{x} - \underline{X}_{(i)}||^2)$, and the uniform function $K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n}) = \frac{1}{k}$.

The result established in Theorem 2.1 proves the asymptotic convergence of the NN regression estimate belonging to $\mathbb{R}^d$ for dependent variables extending Yakowitz result (1987). Therefore in practice we can leave the independent framework without having to filter the observed data. This result is important since it guarantees the consistence of the estimate $m_n(\cdot)$, and therefore the conditional mean forecast will asymptotically coincide with the expected true value. More precisely, the knowledge of the bias and speed rate of the variance of the estimates provides consistent estimates, and their asymptotic normality provides confidence intervals. The building of confidence intervals can be used to compare the quality of point forecasts obtained from different methods, and enhances comparison of several methods (parametric and non-parametric methods), beyond point forecast. Indeed, no rigorous test is available to discuss the choice between the parametric and the non-parametric approaches, and predictive methodology can be used for that objective.

To establish our main result, we assume that the time series $(X_n)_n$ is strictly stationary. It 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 $(\underline{X}_n)_n$ is $h$.

**Theorem 2.1.** *Assuming that $(X_n)_n$ is a stationary time series, and that the following assumptions are verified:*

*(i) $(X_n)_n$ is $\phi$-mixing.*

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

*(iii)* $f(y \mid \underline{x})$ *is bounded,*

*(iv) the sequence* $k(n) < n$ *is such that* $\sum_{i=1}^{k(n)} w_i = 1$,

*then k-NN regression function* $m_n(\underline{x})$ *defined in (2.1) verifies:*

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

*with*

$$E[(m_n(\underline{x}) - m(\underline{x}))^2] = O(n^{-Q}), \tag{2.4}$$

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

$$\sigma^2 = \gamma^2(Var(X_{n+1} \mid \underline{X}_n = \underline{x}) + B^2),$$

*with* $B = O(n^{-\frac{(1-Q)p}{d}})$, *and* $\gamma$ *a positive constant which is equal to 1 when we use uniform weights.*

The proof of this theorem is postponed to the end of the article.

Some points can be mentioned:

1- As soon as the number of neighbors $k$ is different from one, we remark that $\forall u$, $0 < w_i(u) < 1$, whatever the weighting function used, uniform or exponential function.

2- The main difference between $k$-NN method and kernel method (Silverman, 1986) lies on the information set that we use to estimate the function $m(\cdot)$ at a given point $\underline{x}$. In the latter case the information set is fix and in the former case, it is flexible with respect to the choice of the number of neighbors $k$. In this case, such a flexibility has an impact on the values of the weights. Indeed, when the number of neighbors $k$ increases the weights $(w_i)_{i=1}^k$ decrease, then the product $(k.w_i)_{i=1}^k$ turn around a constant $\gamma$ which belongs to $\mathbb{R}$. For uniform weights, $w_i = \frac{1}{k}$ and $\gamma = 1$. This last property implies that the asymptotic variance of the estimate $m_n(\cdot)$ does not depend on the true density nor on the quantity $\int w^2(u)du$. This asymptotic property is not verified when we work with the kernel method, details are provided in Section 5.

3- The mixing conditions characterize different behaviors of dependent variables. Parametric processes like the bilinear models including ARMA models, the related GARCH processes and the Markov switching processes are known to be mixing, Guégan (1983) and Carrasco and Chen (2002). Thus, in practice this condition is not too restrictive.

4- The condition (iv) in theorem 2.1 is verified in particular for the weights introduced in equation (2.2). The parameter $\gamma$ introduced before entails the correlations between the vectors $\underline{X}_n$. Finally the theorem (2.1) providing asymptotic normality for the estimate $m_n(\underline{x})$ under regular conditions permits to build confidence interval whose expression is provided in the following corollary.

**Corollary 2.1.** *Under the assumptions of theorem 2.1, a general form for the confidence interval around $m(\underline{x})$, for a given risk level $0 < \alpha < 1$, is:*

$$m(\underline{x}) \in [m_n(\underline{x}) - B - \frac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\sqrt{k}}, m_n(\underline{x}) + B + \frac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\sqrt{k}}] \tag{2.5}$$

*where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the Student law, $\hat{\sigma}$ is an estimate for $\sigma$ and $B$ is such that:*

1. *$B$ is negligible, if $\frac{k(n)}{n} \to 0$, as $n \to \infty$,*
2. *If not, $B = O(r^p)$, with $r = \left(\frac{k(n)}{(n-d)\hat{h}(\underline{x})c}\right)^{\frac{1}{d}}$ where $c = \frac{\pi^{d/2}}{\Gamma((d+2)/2)}$, and $\hat{h}(\underline{x})$ is an estimate for the density $h(\underline{x})$.*

The proof of this corollary is postponed at the end of the article.

# 3 Forecasting Euro-area GDP

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. There exists many studies proposing real-time modelling in order to take into account some complexity inherent to the computation of the GDP which are: the number of economic indicators, the modelling for GDP and the impact of data revisions, Koenig *et al.* (2003), Baffigi *et al.* (2004), and Schumacher and Breitung (2008). The first paper focuses on the choice of vintage datas and found a substantial improvement of GDP forecast when using real-time vintage datas, the second paper deals on the modelling for GDP by comparing euro area GDP forecast from linear ARIMA and VAR models with bridge equations and concludes that the latter does better for modelling GDP where the last paper suggested the use of large factor models for mixed-frequency datas supported by experiment on monthly German GDP with large numbers of economic indicators pointed out a minor impact of data revision in forecasting performance. In the exercise that we present below, 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 misleading not negligible if they are not properly estimated.

We therefore consider the approach of bridge equations to calculate the GDP in the final stage[1]. We limit ourselves to the eight equations introduced in the paper of Diron (2008), each equation providing a model of GDP, denoted $Y_t^i, i = 1, \cdots, 8$. They are finally aggregated to provide a final value of GDP, denoted $Y_t$. Each equation is calculated from thirteen monthly economic indicators, denoted $X_t^i, i = 1, \cdots, 13$. We focus here in the forecasting of these indicators. We estimate and forecast these indicators from two models: the unrestricted VAR modelling and the multivariate NN approach. For the latter method we distinguish forecasts obtained without embedding data sets (d = 1) from forecasts obtained when d$>$ 1. The thirteen economic indicators that we consider are listed in Table 2.

For this exercise, we use the real-time data base provided by EABCN through their web site[2]. 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 [3]. 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 two previous methodologies 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.

Alternatively, we use a stationary VAR (Vector Autoregressive) methodology for forecasting

[1]More details on this bridge modelling can be found in Runstler and Sedillot (2003)

[2]www.eabcn.org

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

economic indicators and combine later with bridge equations proposed by Diron (2008) to get estimates for GDP, making first all the data stationary using first difference. Among the thirteen indicators used in Diron equations, three indicators (Economic Sentiment Indicator (ESI), Composite Leading Indicator (CLI) and EuRoCoin (ERC)) appear redundant in the sense that they are built from the ten others. Thus, in order to avoid variables repetition in the model which could produce extra contribution on the variance through correlation, we consider only ten indicators as endogenous variables in the VAR modelling, building 10-variates VAR for these remaining ten indicators. Finally, using AIC and Schwartz criteria for order selection we retain a 10-variates VAR(1), Akaike (1974) and Schwartz (1978). Nevertheless, we need estimates for the previous three indicators to finalize the computation of GDP with the Diron equations. We adjust a specific ARIMA modelling for each three variables using AIC criterion for determining the orders $p$ and $q$ . Finally, we retain ARIMA(3,1,0), ARIMA(10,1,0) and ARIMA(1,1,0) respectively for ESI, CLI and ERC indicators. In all parametric models the parameters are estimated by least squares method. We use recursive forecasts when computing the forecast beyond one step ahead.

Regarding the method of NN, $d$ being given, we determine the number of neighbors $k$ by minimizing the mean square error criterion (RMSE):

$$\sqrt{\frac{1}{n-k-d}\sum_{t=k+d+1}^{n}||\hat{X}^i_{t+1}-X^i_{t+1}||^2} \quad i=1,\cdots,13 \tag{3.1}$$

where $n$ is the sample size, $\hat{X}^i_{t+1}$ is the estimate of the i-th economic indicator $X^i_{t+1}$ calculated from the expression (2.1). The number $1 \leq k \leq 5$ of nearest neighbors determined by this criterion at the horizon h=1 is used to calculate the forecasting capabilities for $h > 1$.

In the case of the multivariate approach ($d > 1$), we describe below the algorithm used to determine the embedding dimension $d$ and the number of neighbors $k$ used to obtain the best predictor for $X^i_{n+h}$ in the sense of the previous RMSE. We present the method for all indicators, and thus for simplicity we drop the index $i$ in the algorithm. We mention that we work on stationary process $X_t$ i.e keeping the same transformation all indicators as in VAR modeling, this transformation is done to reflect the consistent result under strict stationarity. We assume that we observe a data $X_1, ..., X_n$ in $\mathbb{R}$.

1. We embed this data set in a space of dimension $d$, $2 \leq d \leq 10$, getting a sequence of vectors in $\mathbb{R}^d$: $\{\underline{X}_d, \underline{X}_{d+1}, ..., \underline{X}_n$, where $\underline{X}_i = (X_{i-d+1}, ..., X_i)\}$.

2. Ranking the vectors, we determine the $k$ nearest vectors of $\underline{X}_n$. Denoting $r_i = \|\underline{X}_n - \underline{X}_i\|$, $i = d, d{+}1, ..., n{-}1$, the distance between these vectors, we build the sequence $r_d, r_{d+1}, ..., r_{n-1}$ ordered in an increasing way: $r_{(d)} < r_{(d+1)} < ... < r_{(n-1)}$, which provides the $k$ nearest vectors $\underline{X}_{(j)}$ corresponding to these $r_{(j)}$, $j = d$, $d+1$,..., $d+k-1$.

3. The one step ahead forecast $m_n(\underline{X}_n) = \hat{X}_{n+1}$, is obtained from:
$$\hat{X}_{n+1} = \sum_{j=d}^{k+d-1} w(\left\|\underline{X}_n - \underline{X}_{(j)}\right\|)X_{(j)+1}. \tag{3.2}$$

4. Considering now the new information set: $X_1, ..., X_n, \hat{X}_{n+1}$, redo step 1 to step 4, we get the two step ahead forecast. We obtain the forecast of third step ahead in a similar way as for the two step ahead forecast. And so on $\cdots$. Our choice of maximal value for $d$ to be ten is a priori choice, we limit to this value as we do not have large number of observation in our time series.

We consider exponential weighting function since it reflects the local behavior of nearest neighbor method giving more weight to closest neighbors. We favor this kind of weights rather than the uniform weights which give the same importance to all neighbors. Now, for each indicator $X_t^i, i = 1, \cdots, 13$, the best pair $(d, k)$ is determined again by minimizing the RMSE criterion defined in (3.1). Once the pair $(d, k)$ is found, it is used for all prediction horizons.

As soon as we get the estimates for the monthly indicators with the two previous methods (VAR and $k$-NN methods), 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 $\hat{Y}_t$ in genuine real-time conditions. We have computed the RMSEs for the quarterly GDP flash estimates, obtained with the forecasting methods used to complete adequately in real-time the monthly indicators, that is VAR modelling and $k$-NN methods ($d = 1$ and $d > 1$). More precisely, we provide the RMSEs of the combined forecasts based on the arithmetic mean of the eight Diron equations. Thus, for a given forecast horizon $h$, we compute $\hat{Y}_t^j(h)$ which is the predictor stemming from these 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 used for the final GDP is

$$RMSE(h) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t(h) - Y_t)^2}, \tag{3.3}$$

where $T$ is the number of quarters between Q1 2003 and Q2 2007 (in our exercise, $T = 18$) and $Y_t$ is the Euro area flash estimate for quarter $t$. The RMSE errors for final GDP are provided in table 1 and comments follow.

| h | VAR | k-NN(1) | k-NN(d>1) |
|---|---|---|---|
| 6 | 0.225 | **0.198** | 0.214 |
| 5 | 0.224 | 0.203 | **0.192** |
| 4 | 0.214 | 0.202 | **0.196** |
| 3 | 0.192 | 0.186 | **0.177** |
| 2 | 0.181 | 0.176 | **0.177** |
| 1 | 0.173 | 0.174 | **0.171** |

Table 1: RMSE for the estimated mean quarterly GDP $Y_t$ computed from equation (3.3), using VAR(p) modelling (column 2) and k-NN predictions ($d = 1$ (column 3), and $d > 1$ (column 4)) for the monthly economic indicators $X_t^i, i = 1, \cdots, 13$, $h$ is the monthly forecast horizon. Values in boldface correspond to the smallest error for a given forecast horizon.

For both methods, VAR modelling and $k$-NN method, the RMSE becomes lower when the forecast horizon reduces from $h = 6$ to $h = 1$, illustrating the accuracy of the nowcasting and forecasting which increases as soon as the information set becomes more and more efficient, thanks to the released monthly data. This is the strengthen of GDP forecasting based on monthly economic indicators, instead of considering only GDP itself since each month, new true values of economic indicators are available. We remark that few days before the publication of the flash estimate (around 13 days with $h = 1$), the lowest RMSE is obtained with the multivariate $k$-NN method (RMSE=0.171).

Looking at forecast errors by comparing column 2 on one side with columns 3 and 4 on the other side, we find that forecast errors are always lower with the method of NN than with VAR modelling (except at horizon $h = 1$ where VAR modelling gives better forecast error than univariate $k$-NN). One source of such gain comes from the use of nearest neighbor method which is adapted even with small samples, which is not the case when working with VAR modelling requiring large samples to be robust.

Lastly, if we focus on nonparametric procedures, we obtain smaller errors when working with multivariate setting $d > 1$ than with univariate one $d = 1$. This result shows the gain of the method developed in a space of higher dimension. Indeed, we expect that in terms of predictions, any method developed using information not only on the path improves the forecast accuracy. This is confirmed when we compare, for the same method, the forecast errors obtained only in $\mathbb{R}$ with the error calculated from a treatment in $\mathbb{R}^d$: in this case the errors are always smaller (e.g. comparing columns 3 and 4 of Table 1). This idea has already developed in other empirical works considering multivariate methods, Kapetanios and Marcellino (2006) with factor models, and Guégan and Rakotomarolahy (2010) with multivariate non-parametric techniques.

To see the evolution of the trajectory of both forecasting parametric and non-parametric methods, we provide the graphs of the observed and estimated GDP growth from $k$-NN methods and VAR modeling for forecast horizons varying from one to six in figures 1 and 2. Some points can be mentioned from these graphs, we get very similar shape for horizons H=1,2 with (in blue) for H=2. While, for greater forecast horizons $H > 3$, the shape of GDP growth forecasts from VAR starting to have straight line converging to the sample means where the evolution of the GDP growth forecasts from $k$-NN keep staying on following the observed trajectory.

## 4 Conclusion

We know the importance of the nowcast and the forecast of macroeconomic variables (such as GDP or inflation) when analysing the current state of the economics and setting policy for the future economic conditions, thus we suggest in this paper alternative methods based on non-parametric multivariate nearest neighbor to improve the accuracy of GDP forecast.

We particularly focus on detecting the best predictor for economic indicators using a RMSE criterion. During the period of estimation, we look after the couple of parameters $(d, k)$ that achieve the best result, and in this phase of our work, we are quite close to philosophy developped inside the works around data mining approach, focusing on the relevant set of data permitting to solve a specific problem with respect to an appropriate criteria, Han et al. (1997) and Hoover and Perez (1999).

Concerning the non-parametric multivariate nearest neighbor method considered here, we have extended the $L^2$ consistence obtained in Yakowitz (1987) with uniform weigthing function to asymptotic normality with any weighting functions.

A possible extension of this work is to perform the way on aggregating monthly economic indicators to match quarterly GDP: the debate seems always open. Another extension concerns the definition of an appropriate test for deciding between parametric and non-parametric methods. Questions around stationarity for data sets working with non-parametric techniques need also to be deeply discussed.

## References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control,* AC.19.

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

Becker, R.A., J.M. Chambers and A.R. Wilks (1988). *The new S language*. Chapman and Hall: New York.

Blake, A.P.(1999). An Artificial Neural Network System of Leading Indicators. *National Institute of Economic and Social Research, unpublished paper.*

Bernanke, B.S. and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics 50,* 525-546.

Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control.* Holden Day: New York.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and Stochastic Volatility models. *Econometric Theory 18,* 17-39.

Michael P. Clements M.P. and H.M. Krolzig (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal 1,* C47-C75.

Collomb, G. (1984). Nonparametric time series analysis and prediction: Uniform almost sure convergence of the window and k-NN autoregression estimates. *Math Oper. Stat., Ser. Statistics 1984.*

Darne, O. (2008). Using business survey in industrial and services sector to nowcast GDP growth: The French case. *Economics Bulletin 3*, No 32, 1-8.

Devroye, L.P. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrscheinlichkeneitstheorie Verw Gebiete 61*, 467-481.

Devroye L.P. and L. Györfi (1985). *Nonparametric Density Estimation: the L1 View.* Wiley: New York.

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.

Donoho, D.L. and I.M. Johnstone (1992). Minimax estimation via wavelet shrinkage. Technical report 402, Dept. Stat. Stantford university.

Engle R.F. and C.W.J. Granger (1987). Co-integration and error correction: Representation, Estimation, and Testing. *Econometrica 55,* 251-276.

Ferrara, L., D. Guégan and P. Rakotomarolahy (2010). GDP nowcasting with ragged-edge data: a semi-parametric modeling. *Journal of Forecasting 29*, 186-199.

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

Friedman, J.H. (1988). Multivariate adaptative regression splines (with discussion). *Annals of Statistics 19,* 1-141.

Guégan, D. (1983). Une condition d'ergodicité pour des modèles ilinéaires à temps discret. *CRAS Série 1, 297-301.*

Guégan, D. and N. Huck (2005). On the use of nearest neighbors in finance. *Revue de Finance 26,* 67-86.

Guégan, D. and P. Rakotomarolahy (2010). A Short Note on the Nowcasting and the Forecasting of Euro-area GDP Using Non-Parametric Techniques. *Economics Bulletin 30,* No.1, 508-518.

Gupta, R. (2006) . Forecasting the South African economy with VARs and VECMs. *South African Journal of Economics 74,* 611–628.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57,* 357–384.

Han, J., Y. Fu, W. Wang, K. Koperski and O. Zaine (1997). DMQL: a data mining query language for relational databases. WP Simon Fraser University, B.C. Canada.

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

Hart, J.D. (1994). *Nonparametric smoothing and lack-of-fit tests*. Springer verlag, New York.

Hoover, K.D. and S.J. Perez (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal 2,* 167–191.

Kapetanios, G. and M. Marcellino (2006). A parametric estimation method for dynamic factors models of large dimensions. *IGIER WP, No 305.*

Koenig E.F., S. Dolmas and J. Piger (2003). The use and abuse of real-time data in economic forecasting. *The Review of Economics and Statistics 85,* 618-628.

Kuan, C.M. and H. White (1994). Artificial neural networks : an econometric perspective. *Econometric Reviews 13,* 1-91.

Litterman R.B. (1986). Forecasting with Bayesian vector autoregressions, five years of experience. *Journal of Business Economic Statistics 4,* 25-38.

Mack, Y.P. (1981). Local properties of $k$-NN regression estimates. *SIAM Journal on Algebraic and Discrete Methods 2,* 311-323.

Mizrach, B. (1992). Multivariate Nearest-Neighbour Forecasts of EMS Exchange Rates. *Journal of Applied Econometrics 7,* Supplement: Special Issue on Nonlinear Dynamics and Econometrics (Dec., 1992), S151-S163.

Ouyang, D., D. Li and Q. Li (2006). Cross-validation and nonparametric k nearest neighbor estimation. *Econometrics Journal 9,* 448–471.

Nowman, B. and B. Saltoglu (2003). Continuous time and nonparametric modelling of U.S. interest rate models. *International Review of Financial Analysis 12,* 25-34.

Pagan A. and A. Ullah (1999). *Nonparametric Econometrics.* Cambridge University Press: Cambridge.

Peligrad, M. and S.A. Utev (1997). Central limit theorem for Linear Processes. *The Annals of Probability 25,* 443-456.

Pena, D., G.C. Tiao and R.S. Tsay (2003). *A course in time series analysis.* Wiley Series in Probability and Statistics: New York.

Prakasa Rao, B.L.S. (1983). *Nonparametric functional estimation.* Orlando FL, Academic press.

Rünstler, G. and F. Sedillot (2003). Short-term estimates of Euro area real GDP by means of monthly data. *European Central Bank WP, No 276.*

Schwartz, G. (1978). Estimating the dimension of ARIMA models. *Annals of Statistics 6,* 461-464.

Silverman, B.W. (1986). *Density estimation for statistics and data Analysis.* Chapmann and Hall: London.

Sims C.A. (1980). Macroeconomics and reality. *Econometrica 48*, 1–48.

Smets, F. and R. Wouters (2004). Forecasting with a Bayesian DSGE model: An application to the Euro area. *Journal of Common Market Studies 42,* 841–867.

Stock, J.H., and M.W. Watson (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics 20,* 147–162.

Stone, C. (1977). Consistent non parametric regression. *Annals of Statistics 5,* 595-645.

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

Tkacz, G. and S. Hu (1999). Forecasting GDP growth using artificial neural networks. *Bank of Canada WP, No 99.*

Tong, H. (1990). *Non-linear time series: a dynamical system approach.* Oxford University Press: Oxford.

Webb R.H. (1995). Forecasts of Inflation from VAR Models. *Journal of Forecasting 14,* 267-285.

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.

## 5 Proofs of Theorem 2.1 and Corollary 2.1

We start giving the proof of theorem 2.1. We first establish a preliminary lemma.

**Lemma 5.1.** *Under the hypotheses of theorem 2.1, either the estimate $m_n(\underline{x})$ is asymptotically unbiased or*

$$E[m_n(\underline{x})] = m(\underline{x}) + O(n^{-\beta}) \tag{5.1}$$

*with $\beta = \frac{(1-Q)p}{d}$.*

**Proof 5.1.** *We denote $B(\underline{x}, r_0) = \{z \in \mathbb{R}^d, \|\underline{x} - z\| \leq r_0\}$ the ball centered at $\underline{x}$ with radius $r_0 > 0$. We characterize the radius $r$ insuring that $k(n)$ observations fall in the ball $B(\underline{x}, r)$; indeed, since the function $h(.)$ is $p-$continuously differentiable, for a given $i$ the probability $q_i$ of an observation $\underline{x}_i$ to fall in $B(\underline{x}, r)$ is:*

$$\begin{aligned} q_i &= P(\underline{x}_i \in B(\underline{x}, r)) & (5.2)\\ &= \int_{B(\underline{x},r)} h(\underline{x}_i) d\underline{x}_i = h(\underline{x}). \int_{B(\underline{x},r)} d\underline{x}_i + \int_{B(\underline{x},r)} (h(\underline{x}_i) - h(\underline{x})) d\underline{x}_i & (5.3)\\ &= h(\underline{x}) c r^d + o(r^d), & (5.4) \end{aligned}$$

*where $c$ is the volume of the unit ball and $\underline{x} = dx_1 dx_2 \cdots dx_d$. Thus, $q_i - q_j = o(r^d)$ for all $i \neq j$. We consider now the k-NN vectors $\underline{x}_{(k)}$ and we denote $q$ the probability that they are in the ball $B(\underline{x}, r)$, that is $q = P(\underline{x}_{(k)} \in B(\underline{x}, r))$, then :*

$$q_i = q + o(r^d). \tag{5.5}$$

*Being given $N(r, n)$, the number of observations falling in the ball $B(\underline{x}, r)$, for a given $r > 0$, we characterize $r$ such that $k(n)$ observations fall in $B(\underline{x}, r)$. We proceed as follows. We denote $\mathcal{S}_i^n$ all non ordered combinations of the $i-$uple indices from $(n - d)$ indices, then:*

$$\begin{aligned} E[N(r,n)] &= \sum_{i=0}^{n-d} iP(N(r,n) = i) = \sum_{i=0}^{n-d} i \sum_{(j_1,\cdots,j_i)\in\mathcal{S}_i^n} \prod_{j=j_1}^{j_i} q_j \prod_{\substack{\ell=1 \\ \ell\notin\{j_1,\cdots,j_i\}}}^{n-d} (1 - q_\ell) \\ &\geq \sum_{i=0}^{n-d} i \sum_{(j_1,\cdots,j_i)\in\mathcal{S}_i^n} \underline{q}^i (1 - \overline{q})^{n-d-i} = \sum_{i=0}^{n-d} i \begin{pmatrix} n-d \\ i \end{pmatrix} \underline{q}^i (1 - \overline{q})^{n-d-i} & (5.6) \end{aligned}$$

$$= \underline{q}(n - d)(1 + \underline{q} - \overline{q})^{n-d},$$

*where $\underline{q}$ and $\overline{q}$ are respectively the smallest and largest probabilities $q_i$ $i = 1, \cdots, n - d$. Thus, we obtain a lower bound for $E[N(r, n)]$. If $E[N(r, n)] = k(n)$, using (5.4) - (5.6), we obtain:*

$$r \leq \left( \frac{k(n)}{(n-d)} \right)^{\frac{1}{d}} D(\underline{x}), \tag{5.7}$$

*with* $D(\underline{x}) = \left(\frac{1}{h(\underline{x})c}\right)^{\frac{1}{d}}$.

*Now, using the relationship (2.1), we get:*

$$E[m_n(\underline{x})] \;=\; \sum_{i\in N(\underline{x})} E[w(\underline{x}-\underline{X}_{(i)})Y_i], \tag{5.8}$$

*where* $Y_i = X_{(i)+1}$. *We can remark that* $E[w(\underline{x}-\underline{X}_{(i)})Y_i] = \int_{\mathbb{R}^d}\int_{\mathbb{R}} w(\underline{x}-\underline{x}_i)y_i f(y_i,\underline{x}_i)d\underline{x}_i dy_i$. *Since* $f(y_i,\underline{x}_i) = f(y_i \mid \underline{x}_i)h(\underline{x}_i)$, *then we obtain* $E[w(\underline{x}-\underline{X}_{(i)})Y_i] = \int_{\mathbb{R}^d}\int_{\mathbb{R}} w(\underline{x}-\underline{x}_i)y_i f(y_i \mid \underline{x}_i)h(\underline{x}_i)d\underline{x}_i dy_i$. *Thus, as soon as the weighting function* $w(\cdot)$ *is vanishing outside the ball* $B(\underline{x},r)$*:*

$$E[w(\underline{x}-\underline{X}_{(i)})Y_i] \;=\; \int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)\left(\int_{\mathbb{R}} y_i f(y_i \mid \underline{x}_i)dy_i\right)h(\underline{x}_i)d\underline{x}_i \tag{5.9}$$

$$\;=\; \int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)m(\underline{x}_i)h(\underline{x}_i)d\underline{x}_i. \tag{5.10}$$

*To compute the bias we need to evaluate:* $E[m_n(\underline{x})] - m(\underline{x})$. *We begin to evaluate :*

$$\sum_{i\in N(\underline{x})}\int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)m(\underline{x})h(\underline{x}_i)d\underline{x}_i = m(\underline{x})E[\sum_{i\in N(\underline{x})} w(\underline{x}-\underline{X}_{(i)})] = m(\underline{x}). \tag{5.11}$$

*Then,*

$$E[m_n(\underline{x})] - m(\underline{x}) \;=\; \sum_{i\in N(\underline{x})}\int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)(m(\underline{x}_i)-m(\underline{x}))h(\underline{x}_i)d\underline{x}_i. \tag{5.12}$$

*The equation (5.12) holds because* $\sum_{i\in N(\underline{x})}\int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)h(\underline{x}_i)d\underline{x}_i = 1$, *(Assumption (iv) in Theorem 2.1). Then,*

$$|E[m_n(\underline{x})] - m(\underline{x})| \le \sum_{i\in N(\underline{x})}\int_{B(\underline{x},r)} w(\underline{x}-\underline{x}_i)a\left\|\underline{x}_i-\underline{x}\right\|^p h(\underline{x}_i)d\underline{x}_i. \tag{5.13}$$

*We get this last expression since the constant* $a$ *is known and* $m(\cdot)$ *is* $p-$*continuously differentiable. The inequality (5.13) implies that:*

$$|E[m_n(\underline{x})] - m(\underline{x})| \le ar^pE[\sum_{i\in N(\underline{x})} w(\underline{x}-\underline{X}_{(i)})]. \tag{5.14}$$

*The relationship in (5.14) holds because* $\|\underline{x}_i-\underline{x}\|^p < r^p$, *as soon as* $\underline{x}_i \in B(\underline{x},r)$. *Now, both cases be considered:*

1. *When* $r$ *is very small, than the bias is negligible and* $E[m_n(\underline{x})] = m(\underline{x})$ .

*2. If the bias is not negligible, using (5.7) and (5.14), we get:*

$$|E[m_n(\underline{x})] - m(\underline{x})| \le a\left(\frac{k(n)}{(n-d)}\right)^{\frac{p}{d}} D(\underline{x})^p. \tag{5.15}$$

*If we choose $k(n)$ as in integer part of $n^Q$, and knowing that $\frac{k}{n-d} \sim \frac{k}{n}$, then $|E[m_n(\underline{x})] - m(\underline{x})| = O(n^{-\beta})$ with $\beta = \frac{(1-Q)p}{d}$.*

*The proof of lemma 5.1 is complete.*

Now, we prove theorem 2.1.

**Proof 5.2.** *1. We begin to establish the relationship (2.4). In the following, we denote $Y_i = X_{(i)+1}$. We rewrite the left part of equation (2.4) as follows:*

$$E[(m_n(\underline{x}) - m(\underline{x}))^2] = Var(m_n(\underline{x})) + (E[m_n(\underline{x})] - m(\underline{x}))^2. \tag{5.16}$$

*We first compute the variance of $m_n(\underline{x})$, considering two cases:*

*a) First case: The weights $w_i$, $i = 1, ..., k$, are independent of $(X_n)$. In that case the variance of $m_n(x)$ is equal to:*

$$Var(m_n(\underline{x})) = A + B, \tag{5.17}$$

*where $A = \sum_{i=1}^{k(n)} w_i^2 Var(Y_i)$ and $B = \sum_{i=1}^{k(n)} \sum_{j \ne i} w_i w_j cov(Y_i, Y_j)$. Using the assumption (ii) of theorem 2.1, we get $|B| \le \sum_{i=1}^{k(n)} \sum_{j \ne i} |cov(Y_i, Y_j)|$. This last term is negligible due to Yakowitz' result (1987) on the sum of covariances. Now , $A = \frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_i)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2)$. Using the fact that the weights are decreasing with respect to the chosen distance, $w_k \le \cdots \le w_1$, we get:*

$$\frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_k)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2) \le A \le \frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_1)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2). \tag{5.18}$$

*As soon as $k(n) \to \infty$ the product $k(n)w_i$ converges to one in case of uniform weights, and can be bounded for exponential weights for all $i$ and for all $n$, thus there exist two positive constants $c_0$ and $c_1$ such that (5.18) becomes :*

$$\frac{c_1^2}{k(n)^2} \sum_{i=1}^{k(n)} (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2) \le A \le \frac{c_0^2}{k(n)^2} \sum_{i=1}^{k(n)} (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2). \tag{5.19}$$

*where* $v(\underline{x}) = Var(X_{n+1} \mid \underline{X}_n = \underline{x})$*. Using the assumption (iv) of Theorem 2.1, we remark that* $E[Y_i] = E[m_n(\underline{x})]$*. Now again, if* $k(n) = [n^Q]$ *where* $[\cdot]$ *corresponds to the integer part of a real number, then* $A = O(n^{-Q})$ *from lemma 5.1 when* $n \to \infty$*. It follows that the relationship (5.17) becomes:*

$$Var(m_n(\underline{x})) = O(n^{-Q}), \tag{5.20}$$

*and*

$$(E[m_n(\underline{x}) - m(\underline{x})])^2 = O(n^{-2\beta}). \tag{5.21}$$

*Plugging equations (5.20) and (5.21) inside equation (5.16), we get* $2\beta = Q$ *or* $Q = \frac{2p}{2p+d}$ *and the proof is complete.*

*b) Second case: the weights* $w_i$*,* $i = 1, ..., k$*, depend on* $(X_n)$*. We use again the relationship (5.17) with* $A = \sum_{i=1}^{k(n)} Var(w(x - \underline{X}_{(i)})Y_i)$ *and* $B = \sum_{i=1}^{k(n)} \sum_{j \neq i} cov(w(x - \underline{X}_{(i)})Y_i, w(x - \underline{X}_{(j)})Y_j)$*. Remarking that* $(w(x - \underline{X}_{(j)})Y_j)$ *are* $\phi$*-mixing since* $(X_j)$ *and* $(Y_j)$ *are* $\phi$*-mixing Pagan and Ullah (1999), then* $B$ *is negligible from Yakowitz' result (1987). We remark also that* $A = \sum_{i=1}^{k(n)} (E[(w(x - \underline{X}_{(i)})Y_i)^2] - (E[w(x - \underline{X}_{(i)})Y_i])^2)$*, then*

$$A = \sum_{i=1}^{k(n)} [\int_{\mathbb{R}^d} \int_{\mathbb{R}} w(\underline{x} - \underline{x}_i)^2 y_i^2 f(y_i, \underline{x}_i) d\underline{x}_i dy_i - (\int_{\mathbb{R}^d} \int_{\mathbb{R}} w(\underline{x} - \underline{x}_i) y_i f(y_i, \underline{x}_i) d\underline{x}_i dy_i)^2]. \tag{5.22}$$

*When* $k$ *increases, the weights* $w_i$ *decrease, and* $k(n)w_i \sim \gamma$ *where* $\gamma$ *is a real constant, then*

$$\begin{aligned} A &= \frac{\gamma^2}{k(n)^2} \sum_{i=1}^{k(n)} [\int_{\mathbb{R}^d} \int_{\mathbb{R}} y_i^2 f(y_i, \underline{x}_i) d\underline{x}_i dy_i - (\int_{\mathbb{R}^d} \int_{\mathbb{R}} y_i f(y_i, \underline{x}_i) d\underline{x}_i dy_i)^2] & (5.23) \\ &= \frac{\gamma^2}{k(n)^2} \sum_{i=1}^{k(n)} (E[Y_i^2] - E[Y_i]^2). & (5.24) \end{aligned}$$

*Under stationary conditions for* $(X_n)$ *and recalling that* $Y_i = X_{(i)+1}$*, then equation (5.24) is equivalent to* $A = \frac{\gamma^2}{k(n)}(E[X_1^2] - E[X_1]^2)$ *and* $A = \frac{\gamma^2}{k(n)} Var(X_1)$*. Finally expression (5.17) becomes:*

$$Var(m_n(\underline{x})) = \frac{\gamma^2}{k(n)} Var(X_1). \tag{5.25}$$

*Moreover, when we take* $k(n) = n^Q$*, thus equation (5.25) is equal to:*

$$Var(m_n(\underline{x})) = O(n^{-Q}). \tag{5.26}$$

*Plugging equations (5.26) and (5.21) in equation (5.16), gives* $2\beta = Q$*, and* $Q = \frac{2p}{2p+d}$*, and the proof is complete.*

2. We prove now the asymptotic normality of $m_n(\underline{x})$. We assume that the variance $\sigma_n = var[m_n(\underline{x})]$ exists and is non null, thus:

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} \frac{w_i Y_i - E w_i Y_i}{\sigma_n}. \tag{5.27}$$

*To establish the asymptotic normality of $m_n(\underline{x})$, we distinguish three cases corresponding to the choice of the weighting functions.*

*i) The weights are uniform, $w_i = \frac{1}{k(n)}$, then equation (5.27) becomes:*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} \frac{1}{k(n)} Z_i, \tag{5.28}$$

*where $Z_i = \frac{Y_i - EY_i}{\sigma_n}$. The asymptotic normality of equation (5.28) is obtained using theorem 2.2 in Peligrad and Utev (1997) . To compute the variance, we follow Yakowitz's work (1987): $var(m_n(\underline{x})) = \frac{1}{k(n)^2} var(\sum_{i=1}^{k(n)} Y_i) = \frac{1}{k(n)}[var(Y \mid \underline{X} = \underline{x}) + O(n^{-\frac{2(1-Q)p}{d}})]$, then equation (5.28) becomes,*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sqrt{n^Q} \sum_{i=1}^{k(n)} \frac{w_i Y_i - E w_i Y_i}{\sigma}, \tag{5.29}$$

*when $k(n) = [n^Q]$ and $\sigma^2 = var(Y \mid \underline{X} = \underline{x})$, and the proof is complete.*

*ii) The weights $w_i$ are real numbers and do not depend on $(X_n)_n$, then*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} w_i Z_i, \tag{5.30}$$

*where $Z_i = \frac{Y_i - EY_i}{\sigma_n}$. Now, applying again the theorem 2.2 in Peligrad and Utev (1997), we get the asymptotic normality remarking that $E[\sum_{i=1}^{k(n)} w_i Z_i] = 0$ and $Var[\sum_{i=1}^{k(n)} w_i Z_i] = 1$. To compute $\sigma_n^2 = Var[m_n(x)]$, we use the stationary condition of the time series $(X_n)_n$, thus:*

$$Var[m_n(\underline{x})] = \sum_{i=1}^{k(n)} w_i^2 Var[Y_i] = \sum_{i=1}^{k(n)} w_i^2 [Var[Y_{n+1} | \underline{X}_n = \underline{x}] + B^2],$$

*where $B$ is given in lemma 3.1. Remarking that $\frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n) w_i)^2 < \infty$, then $\sum_{i=1}^{k(n)} w_i^2 < \infty$ and*

$$Var[m_n(\underline{x})] = [Var[Y_i | \underline{X}_i = \underline{x}] + B^2] \sum_{i=1}^{k(n)} w_i^2.$$

*As soon as* $\sum_{i=1}^{k(n)} w_i^2 \sim \frac{\gamma^2}{k(n)}$, *and* $k(n) = [n^Q]$, *we get the result.*

*iii) Finally, we assume that* $w_i = \frac{w(\underline{x}-\underline{X}_{(i)})}{\sum_{i=1}^{K} w(\underline{x}-\underline{X}_{(i)})}$ *where* $w(.)$ *is a given function. In that latter case, the weights depend on the process* $(X_n)_n$*. In the following, we denote by* $N(i)$ *the order of the* $i^{th}$ *neighbor. We rewrite the neighbor indices in an increasing order such that* $M(1) = min\{N(i), 1 \leq i \leq K\}$ *and* $M(k) = min\{N(i) \notin \{M(j), \forall j < k\}, 1 \leq i \leq K\}$ *for* $2 \leq k \leq K$, *and* $K = k(n)$ *is the number of neighbors. We introduce a real triangular sequence* $\{\alpha_{Ki}, 1 \leq i \leq K$ *and* $\alpha_{Ki} \neq 0$ $\forall i\}$ *such that*

$$\underset{K}{Sup} \sum_{i=1}^{K} \alpha_{Ki}^2 < \infty \quad and \quad \underset{1 \leq i \leq K}{max} |\alpha_{Ki}| \underset{n \to \infty}{\longrightarrow} 0. \tag{5.31}$$

*Now using the sequences* $M(j), j = 1, \cdots, K$ *and* $(\alpha_{Ki}), 1 \leq i \leq K$, *we can rewrite expression (5.27) as:*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{K} \alpha_{Ki} S_i, \tag{5.32}$$

*with* $S_i = \frac{w_{M(i)} X_{M(i)+1} - Ew_{M(i)} X_{M(i)+1}}{\alpha_{Ki}\sigma_n}$. *The sequence* $(S_i^2)$ *is uniformly integrable and* $S_i$ *is function only of* $(X_j, j \leq M(i)+1)$, *thus if we denote* $\mathcal{F}_i$, $\mathcal{G}_i$, $\mathcal{F}_i^j$ *and* $\mathcal{G}_i^j$, *the sigma algebras generated by* $\{X_r\}_{r \leq i}$, $\{S_r\}_{r \leq i}$, $\{X_r\}_{r=i}^{j}$ *and* $\{S_r\}_{r=i}^{j}$ *respectively, then* $S_i \in \mathcal{F}_{M(i)+1}$, *and* $\mathcal{G}_i \subset \mathcal{F}_{M(i)+1}$. *For a given integer* $\ell$, *we have also* $\mathcal{G}_{n+\ell}^{\infty} \subseteq \mathcal{F}_{n+M(\ell)+1}^{\infty}$ *since* $M(1) < M(1)+1 \leq M(2) < \cdots \leq M(n+\ell) < M(n+\ell)+1 \leq M(n+\ell+1)$. *Then:*

$$\underset{\ell}{sup} \underset{A \in \mathcal{G}_1^{\ell}, B \in \mathcal{G}_{n+\ell}^{\infty}, P(A) \neq 0}{Sup} |P(B \mid A) - P(B)| \leq \underset{\ell}{sup} \underset{A \in \mathcal{F}_1^{M(\ell)+1}, B \in \mathcal{F}_{n+M(\ell)+1}^{\infty}, P(A) \neq 0}{Sup} |P(B \mid A) - P(B)|. \tag{5.33}$$

*Under the* $\phi-$*mixing assumption on* $(X_n)_n$, *the right hand part of the expression (5.33) tends to zero as* $n \to \infty$ *and the lelf hand part of (5.33) converges to zero, hence the sequence* $(S_i)_i$ *is* $\phi-$*mixing. Moreover, for all i:*

$$S_i \quad is \quad centered \quad and \quad var(\sum_{i=1}^{K} \alpha_{Ki} S_i) = var(\frac{m_n(\underline{x})}{\sigma_n}) = 1. \tag{5.34}$$

*Then, using expressions (5.31) - (5.34), and the theorem 2.2 in Peligrad and Utev (1997), we get:*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} \to_{\mathcal{D}} \mathcal{N}(0,1) \tag{5.35}$$

*The variance of* $m_n(x)$ *is given by the relation (5.25). The proof of the theorem 2.1 is complete.*

We provide now the proof of Corollary 2.1.

**Proof 5.3** (Proof of corollary 2.1). *From theorem 2.1, a confidence interval, for a given $\alpha$ can be computed, and has the expression:*

$$-z_{1-\frac{\alpha}{2}} \leq \frac{m_n(\underline{x}) - Em_n(\underline{x})}{\hat{\sigma}_n} \leq z_{1-\frac{\alpha}{2}} \tag{5.36}$$

*where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of Student law. Previoulsy, we have established that the estimate $m_n(\underline{x})$ can be biased, thus the relationship (5.36) becomes:*

$$m_n(\underline{x}) + B - \hat{\sigma}_n z_{1-\frac{\alpha}{2}} \leq m(\underline{x}) \leq m_n(\underline{x}) + B + \hat{\sigma}_n z_{1-\frac{\alpha}{2}} \tag{5.37}$$

*When the bias is negligible, the corollary is established. If the bias is not negligible, we can bound it. The bound is obtained using expressions (5.7) and (5.38):*

$$B = O\Bigg(\Big(\frac{k(n)}{(n-d)\hat{h}(\underline{x})c}\Big)^{\frac{p}{d}}\Bigg) \tag{5.38}$$

*with $c = \frac{\pi^{d/2}}{\Gamma((d+2)/2)}$, $\hat{h}(\underline{x})$ being an estimate of the density $h(\underline{x})$. Introducing this bound in expression (5.37) completes the proof.*

# 6 APPENDIX

## 6.1 Euro Area Monthly Indicators

We provide in table 2 the list of the monthly economic indicators used in this study for the computation of the GDP using the bridge equations.

## 6.2 The bridge equation

We specify the bridge equations we use, details can be found in Diron (2008). Let us define $Y_t$ as: $Y_t = (\log GDP_t - \log GDP_{t-1}) \times 100$, where $GDP_t$ is the GDP at time $t$. The final GDP $Y_t$ used in the paper is the mean of the eight values computed below.

1. EQ1. $Y_t^1 = a_0^1 + a_1^1(\log X_t^1 - \log X_{t-1}^1) + a_2^1(\log X_t^2 - \log X_{t-1}^2) + a_3^1 X_{t-1}^3 + \varepsilon_t.$

2. EQ2.

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

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

Table 2: Summary of the thirteen economic indicators of Euro area used in the eight GDP bridge equations.

3. EQ3. $Y_t^3 = a_0^3 + a_1^3 X_t^7 + a_2^3 X_{t-1}^7 + \varepsilon_t$.

4. EQ4. $Y_t^4 = a_0^4 + a_1^4 (X_t^6 - X_{t-1}^6) + a_2^4 X_t^3 + \varepsilon_t$.

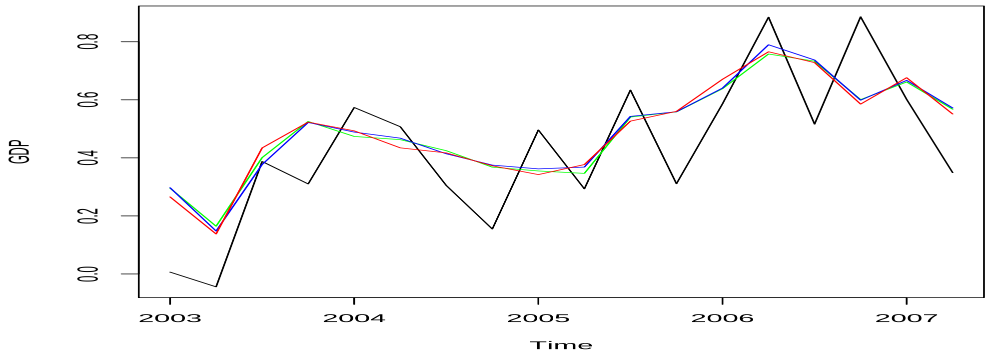
5. EQ5. $Y_t^5 = a_0^5 + a_1^5 (X_t^6 - X_{t-1}^6) + a_2^5 X_t^9 + a_3^5 X_t^8 + \varepsilon_t$.

6. EQ6. $Y_t^6 = a_0^6 + a_1^6 (\log X_{t-2}^{10} - \log X_{t-3}^{10}) + a_2^6 (\log X_{t-1}^{11} - \log X_{t-2}^{11}) + \varepsilon_t$.

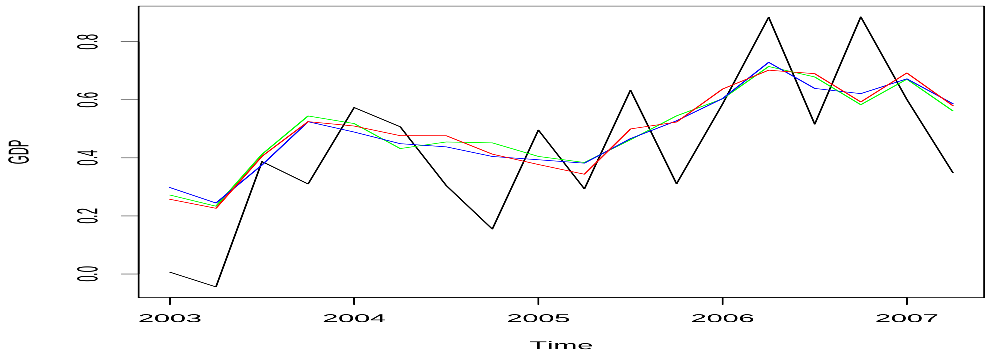
7. EQ7. $Y_t^7 = a_0^7 + a_1^7 (\log X_t^{12} - \log X_{t-1}^{12}) + a_2^7 (\log X_{t-2}^{12} - \log X_{t-3}^{12}) + a_3^7 Y_{t-1}^7 + \varepsilon_t$, and

8. EQ8.$Y_t^8 = a_0^8 + a_1^8 X_t^{13} + \varepsilon_t$.

(a) GDP growth rate forecast at horizon H=1 using VAR and $k$-NN methods.




(b) GDP growth rate forecast at horizon H=2 using VAR and $k$-NN methods.




(c) GDP growth rate forecast at horizon H=3 using VAR and $k$-NN methods.

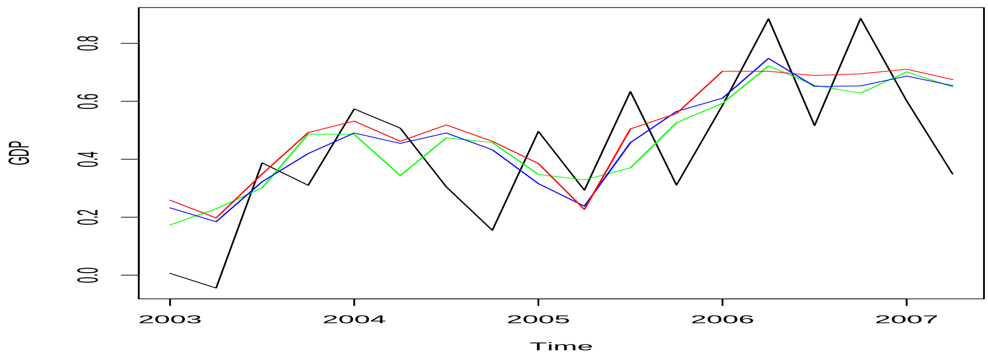
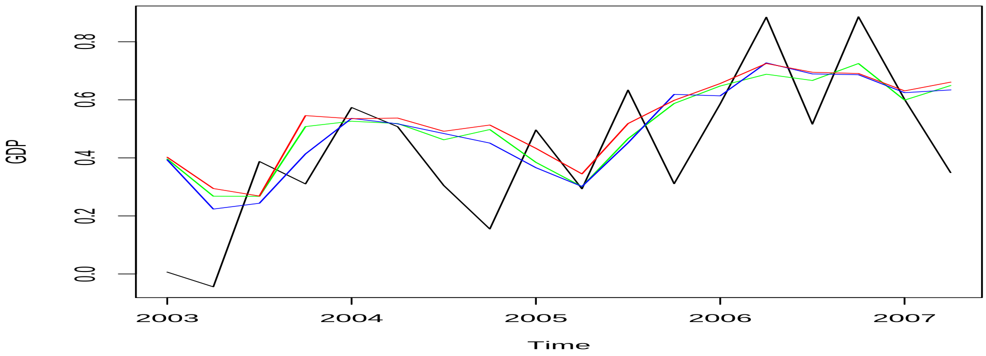
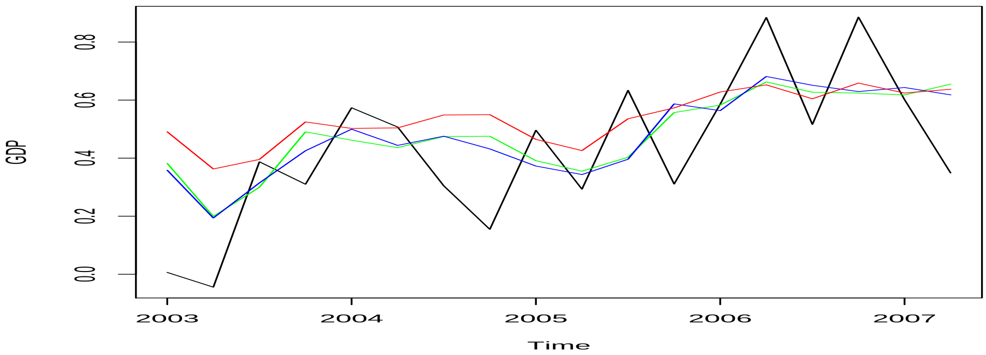



Figure 1: Quarterly observed (in black) and forecasted GDP growth rate computed from $k$-NN with d=1 (in green), $k$-NN with d>1 (in blue) and VAR (in red) models between 2003Q1 and 2007Q2 for different forecast horizons in panel: (a) for horizon $H = 1$, (b) for horizon $H = 2$ and (c) for horizon $H = 3$.

(a) GDP growth rate forecast at horizon H=4 using VAR and $k$-NN methods.




(b) GDP growth rate forecast at horizon H=5 using VAR and $k$-NN methods.




(c) GDP growth rate forecast at horizon H=6 using VAR and $k$-NN methods.

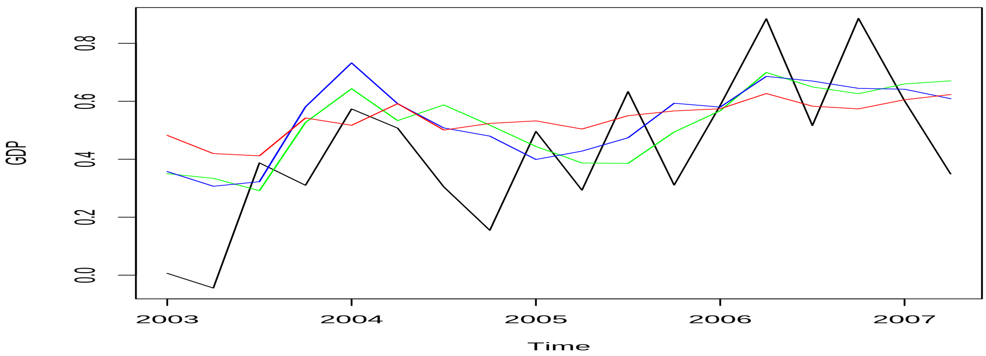



Figure 2: Quarterly observed (in black) and forecasted GDP growth rate computed from $k$-NN with d=1 (in green), $k$-NN with d>1 (in blue) and VAR (in red) models between 2003Q1 and 2007Q2 for different forecast horizons in panel: (a) for horizon $H = 4$, (b) for horizon $H = 5$ and (c) for horizon $H = 6$.