Forecasting Regional GDP in Italy

Luciano Gutierrez*

February 27, 2007

Abstract

This paper explores the usefulness of factor and bootstrap aggregation forecasting in predicting regional GDP in Italy. We use methods designed to target the set of potential predictors. We compute the mean square forecasting error (MSE) by using direct multi-step forecasts for the period 2004-2005. Our findings can be summarized as follows. First, factor and bagging forecasts generally show lower mean square forecasting error than the mean square error of the autoregressive AR(3) model used as a benchmark. Secondly, bagging methods seem to produce similar MSE as factor augmented models, especially for predicting aggregate GDPs. In synthesis, our analysis shows that using factors and bagging methods reduces the prediction mean squared error relative to standard forecasting methods.

JEL Classification: E37, C53 .

Keywords: Factor models, Bootstrap aggregation, Forecast method selection, Pre-testing.

1 Introduction

A problem in out of sample prediction is that a great number of predictors $(X_{1t}, X_{2t}, ..., X_{Nt})$ are potentially relevant when forecasting the variable y_{t+h} h steps ahead, but few of these predictors are likely to have high predictive power. If a researcher uses only one of these predictors the forecasts tend to be generally unreliable and unstable (see, e.g., Cecchetti *et al.* (2000), Stock and Watson (2003)) and if all predictors are included this may lead to overfitting and poor out-of-sample forecast accuracy or may indeed be unfeasible, as in the case of regional forecasts where the number of observations of series is usually small.

*Corresponding author: Department of Agricultural Economics, University of Sassari, Via E. De Nicola 1, Sassari 07100, Italy. Telephone: +39 079 229 256; Telefax: +39 079 229 356; E-mail address: lgutierr@uniss.it.

There are many strategies that can be used in this situation. One strategy combines forecasts from many models with alternative subsets of predictors. In this case one can use the mean, median or trimmed mean of these forecasts as final forecasts, Stock and Watson (2003). A second strategy uses factor augmented forecasts. The methodology allows for factors from all predictors to be estimated by the principal components method, and includes them in a linear forecasting equation for y_{t+h}. This method has been fruitfully used by Delle Monache *et al.* (2006) for predicting real GDP for North-Center and South Italian macro regions.

The previous strategies do not take into account the predictive ability of each X_{it} for y_{t+h}. However a third strategy uses a testing procedure to decide which predictors to include in the forecasting regression. In this case, we conduct a pre-test analysis for fitting a regression, where we include only those predictors that are significant in predicting y. This method has been labeled in Bai and Ng (2006a) "targeted predictors strategy".

The previous pre-test strategy may lead to unstable decision rules, in that small alterations in the data set may cause a predictor to be added or to be dropped. In such cases the variance of forecasts may increase and undermine the accuracy of pre-test forecasts in applied work. However, the predictive accuracy of pre-test strategy can be enhanced by application of the bagging technique. This leads to a further strategy.

Bagging is a statistical method designed to reduce the out-of-sample prediction mean-squared error of forecast models which are selected by unstable decision rules such as pre-tests. The term *bagging* is short for *bootstrap aggregation*, Breiman (1996). Bagging involves analyzing a model that includes all potential predictors, generating a large number of bootstrap resamples, applying the pre-test rules to each of resample, and averaging the forecasts from the models selected by the pre-test on each bootstrap sample. By averaging across resamples, after variable selection bagging may reduce the prediction mean squared error of the regression model.

The factor model and bagging strategies have been used to predict the regional GDP for 20 Italian regions. The data are annual and available for the period 1980-2005. Given the large amount of potential predictors, well above the number of observations available, fitting a model which includes all predictors and then discarding those that are insignificant is unfeasible.

Thus we concentrate on three methods. The first method predicts the GDP for each region using the factors computed from the full set of predictors. The second method selects the predictors using a selection criteria based on a standard t-test. The selected predictors are then used to compute principal components. The last method uses bagging predictors. We introduce the predictors with the highest significance in the forecasting regression and compute the forecasts of y using the bagging method.

In the following two sections we illustrate the methods. In section 4, we use both approaches to forecast regional Italian GDP two years ahead. The relative performance of the methods are then compared in section 4 in order to see if the factor method is more accurate than the bagging procedure. We conclude in section 5.

2 Factor Augmented Regression

Assume that data for a large number of predictors $X_t = (X_{1t}, X_{2t}, ..., X_{Nt})'$, $i = 1, 2, ..., N$ and $t = 1, 2..., T$ are available. We want to forecast the variable y_{t+h}, h step ahead. If $N < T$ the forecast $\hat{y}_{T+h} = \hat{\alpha} W_T + \hat{\beta} X_T$ that uses all predictors and $W_t = (1, y_{t+h-1}, ..., y_{t+h-p_W})$, with p_W the maximum number of lags of the dependent variable, may be feasible. However, when $N > T$ we need a method to reduce the number of predictors. One strategy uses a factor approach to model y. Basically the method estimates the following factor-augmented regression using data for $t = 1, ..., T - h$

$$y_{t+h} = \alpha' W_t + \beta(L)' \hat{f}_t + \varepsilon_{t+h} \tag{1}$$

where $\beta(L)$ is a polynomial of coefficients pertaining to $\hat{f}_t$ with p_f lags, and $\hat{f}_t \subset \hat{F}_t$, denotes a vector of M largest factors extracted from the principal component estimates $\hat{F}_t$ of the model

$$X_{it} = \lambda_i' F_t + e_{it}. \tag{2}$$

Equations (1) and (2) constitute the well known diffusion index forecasting method of Stock and Watson (2002). The principal components estimates $\hat{f}_t$ are $\sqrt{T}$ times the M eigenvectors corresponding to the first M eigenvalues of the matrix $(X_t X_t')$. Thus in order to obtain feasible estimates, the factor-augmented procedure requires the number M of factors in (1). The task may be accomplished or fixing M to some value or using the methodology proposed in Bai and Ng (2002).

The error forecast variance in 1) depends on the parameters and factor estimation uncertainty. In an influential paper Bai and Ng (2006b) show that if $\sqrt{T}/N \to 0$ in large panels, the uncertainty is dominated by parameter uncertainty and factors can be treated as observable. This result highlights the importance of the index N, i.e. the number of indicator or predictor variables, when estimating the factor process F_t. Thus for panels with large N but small T, as is generally the case for regional panels, the forecast error variance is likely to depend more on parameter uncertainty than factor uncertainty.

2.1 Targeted Predictors

Boivin and Ng (2006) find that augmenting the number of predictors in X_t does not necessarily improve forecasts. Thus instead of extracting the factors from a large data set which includes, for example, all the potentially predictors, one can isolate only a subset of the N series in X_t using certain screening methods. We use the following method

a. For each of the N potential predictors, we run the following regression

$$y_t = \alpha_i W_{t-h} + \beta_i X_{it-h} + \varepsilon_{it+h}$$

and compute the t-statistic t_i associated to each predictors X_{it-h} .

b. We rank in descending order the t-statistics associated to each predictor X_{it-h} by sorting $t_1, t_2, ..., t_N$. [1]

c. Choosing a threshold significance level α, for example 5%, we define the number of predictors N^h whose t_i exceeds the significance level.

d. let $X_t^h = (X_{1t}, X_{2t}, ..., X_{N^h t})'$ be the new set of N^h predictors with $X_t^h \subset X_t$. We estimate F_t from X_t^h by the method of principal components. Using Bai and Ng's (2002) method we then compute the optimal number M of factors $\hat{f}_t$ to include in the forecasting equation (1).

e. The h period ahead forecast is computed as $\hat{y}_{T+h} = \hat{\alpha} W_T + \hat{\beta}'(L)\hat{f}_T$.

Thus the method differs from the standard one which uses all the predictors because it includes only variables whose predictive power for y is significant at the level of significance α. Note that in constructing the t-statistic there may be problems if there is serial correlation and/or conditional heteroskedasticity especially when computing the standard errors of β_i. The β_i problem may be addressed using White (1980) and West (1997) robust standard error methods.

3 Bagging method

Bagging is a statistical method designed to reduce out-of-sample mean squared error of forecasts selected by unstable decision rules such as pre-tests. The method was introduced by Breiman (1996) to reduce the variance of a predictor. Bühlmann and Yu (2002) have analyzed the performance of the bagging algorithm for a linear model with a single regressor. They show that the bagging predictor in many cases has better mean square error than other pre-test predictors. Further insights are contained in Inoue and Kilian (2005).

The pre-test procedure presented in the previous section can be easily adapted to obtain bootstrap aggregate or bagging predictions. In brief the method requires

[1] Note that, unlike other procedures which use the absolute value, $|t_i|$, see Bai and Ng (2006a), we focus only on the predictors which have a positive relationship with y.

a. Arrange the set of tuples (y_{T+h}, X_T), $t = 1, 2, ..., T - h$ in the form of a matrix of dimension $(T - h) \times (N + 1)$ and construct a bootstrap samples $(y^*_{1+h}, X^*_1), ..., (y^*_{T+h}, X^*_T)$ by drawing with replacement blocks of k rows of this matrix, where the block size k is chosen to capture the dependence in the error term, see Gonçalves and White (2004).

b. For each of the N potential predictors, run the regression

$$y^*_t = \alpha_i W_{t-h} + \beta_i X^*_{it-h} + \varepsilon_{it+h}$$

and compute the t-statistic t_i associated to each predictors X^*_{it-h} .

c. Rank in descending order the t-statistics associated to each predictor X^*_{it-h} by sorting $t_1, t_2, ..., t_N$.

d. Choose a threshold significance level α, for example 5%, and define the number of predictors N^{h*} whose t_i exceeds the significance level.

e. Compute the h period forecast using the N^{h*} predictors X^{h*}_t defined in the previous step, that is $\hat{y}_{T+h} = \hat{\alpha} W_T + \hat{\gamma}'(L) \hat{X}^{h*}_T$.

f. Compute steps 1-5 B times. Thus the bagged predictor y^{BA}_{T+h} is the expectation of the bootstrap predictor across bootstrap samples

$$y^{BA}_{T+h} = \frac{1}{B} \sum_{i=1}^{B} y^{*i}_{T+h}$$

where while $B = \infty$ in theory, in practice $B = 100$ tends to provide a good approximation.

Note that steps from b. to d. are similar to those presented in the previous section, i.e. the pre-test estimation strategy mimics the strategy used to target predictors in factor analysis.

4 Empirical Analysis

Our interest is in forecasting the variable y_{it+h}, which is the real GDP for each of $i = 1, ..., 20$ Italian regions. These are annual series available for the period 1980-2005 for a total of $T = 26$ observations. [2] Clearly, by aggregation, the method allows the Italian GDP, $Y_{t+h} = \sum_{i=1}^{20} y_{it+h}$, and the GDPs of the five macro-areas North-West, North-East, Center, North-Center and South Italy to be forecasted.

In the analysis we grouped the predictors into three data sets. The first data set groups only regional predictors. As in the case of regional GDPs series, this

[2] Rencently, ISTAT has published the new regional GDP chain-linked volumes data set which covers the years 2000-2005. We have linked the old series 1980-2004 with the new one.

set of predictors is available for the period 1980-2005, although for some series, for example the investment series, the data are available only for the period 1980-2005. Thus for example to predict the regional GDPs two steps ahead, a set of regional predictors which span the full period 1980-2005, are lagged fixing $h = 2$. For the second set of regional predictors which are available only for the period 1980-2004, we fix $h = 3$. The second data set includes only national predictors. Given the relative weight of some regions, such as Lombardia or Emilia Romagna in the national economy, national predictors may forecast regional GDP better. This dataset covers the period 1980-2005. Finally a third data set mixes national and regional predictors. We label these predictors "mixed predictors". The full set of predictors is contained in Appendix 1. If not otherwise stated, both y_{it+h} as well as the set of $(X_{1t}, X_{2t}, ..., X_{Nt})$ predictors are assumed to be I(1) nonstationary series and thus the data are transformed by taking logs and first differences.

For $h = 1, 2, 3$ the factor augmented forecast given information in time t is

$$\hat{y}_{t+h|t} = \hat{\alpha_f} W_t + \hat{\beta'_f}(L)\hat{f}_t, \tag{3}$$

and the bagging forecast equation is

$$\hat{y}_{t+h|t} = \hat{\alpha_b} W_t + \hat{\beta'_b}(L) X_t^{h*}, \tag{4}$$

where the number of lags in W_t and $\hat{f}_t$ as well as in X_t^{h*} may determined by BIC criterion after defining a maximum number of lags. In the analysis, given the small number of observations , we fix $\hat{\beta'_f}(L) = \hat{\beta'_f}$ and $\hat{\beta'_b}(L) = \hat{\beta'_b}$ respectively in (3) and (4), i.e. we only observe contemporaneous effects of factors and bagging predictors on regional GDPs. In equation (3) we allows for up to five factors, i.e M is chosen in the interval $[1, 5]$. A similar strategy is used in bagging regression. In this case we provide estimates for a maximum of five "best" predictors, i.e. N^{*h} assumes the values included in the interval $[1, 5]$. Thus both strategies allow for a minimum number of degree of freedoms in the forecasting equations (3) and (4). The block size is $k = 3$ to take into account possible dependence in error.

In the following tables we report the results for $h = 2$. Similar results are obtained for $h = 1, 3$. For $h = 2$ the last observation used in the estimation is 2003. We employ a standard practice to compute the average of the forecast errors in order to evaluate the different methods. To be precise, we refer to the ratio of the mean squared error (MSE) for a given method to the MSE of an AR(3) as relative mean squared error (RMSE). That is

$$RMSE(method) = \frac{MSE(method)}{MSE(AR(3))}.$$

Thus an entry less than one indicates that the specified method is superior to the benchmark forecast.

4.1 Results

The forecasting exercise uses the data from 1980-2003. A $h = 2$ period ahead forecast is formed by direct multi-step forecast using values of regressors at 2003 to give 2004 and 2005 GDP forecast. Specifically, when $h = 2$ first we estimate factors and bagging predictors, then the forecasting equations are estimated and forecasts for the period 2004-2005 are obtained. For the factor pre-testing strategy and bagging procedure the cutoff point was set at the level of significance of 5%. Thus only predictors with $t > 1.65$ have been included as effective predictors in the set X_t^{h*} .

Table 1 reports the RMSE of the factor augmented model for various values of p, i.e. the lags of dependent variables. Thus when $p = 0$ the forecast equation (3) is a static regression. In the first three columns factors are estimated using all the regional predictors. In the last three columns we present the RMSE computed adopting Bai and Ng's (2006b) BIC(3) procedure, fixing the maximum number of factors to five. In the last two rows we show respectively the percentage of regions and the percentage of aggregates with the RMSE less than one.

Overall the improvements of the factor forecasts over the benchmark forecasts are effective. When $p = 1$ and $M = 5$ 80% of regional forecasts show a RMSE of less than one. Moreover all the aggregates have $RMSE < 1$. Using the pre-testing strategy we obtain quite the same improvements over the benchmark model, as highlighted in Table 2.

A quick glance at the results in Table 3 and at Table 4, where we report RMSE for the national predictors, shows that most of the entries are below 1.This indicates that there are efficiency gains, especially in pre-testing strategy. These results are confirmed by the aggregate. Interestingly, if we use the mixed predictors, that is the set of the national plus regional data sets, we obtain worse results, see Table 5 and 6. Thus factor models which use national variables to predict regional GDPs seems to provide better forecasts for the selected period 2004-2005.

In the following three tables, we report the RMSE computed using the bagging procedure. In the first three columns of the tables we show the RMSE computed when the number of selected predictors is $N^{*h} = 1$. In the remaining three columns we fix $N^{*h} = 5$. As before the lag p of the dependent variable ranges from $[0, 2]$. Overall the method seems to work fine as the factor augmented method. As previously reported, using the set of national predictors improves the number of regions with $RMSE < 1$.

5 Conclusions

In the paper we explore the usefulness of factor forecasting and bootstrap aggregation forecasting, or bagging forecasting, in predicting regional GDP in Italy.

Due to lack of observations we were only able to compare these methods by looking at the mean squared forecasting errors computed for the period 2004-2005. In brief the analysis suggests that using pre-testing strategies to select predictors gives some improvements relatively to methods that use all the available predictors. Both factor methods and bagging procedures reduce regional GDP forecast errors when compared to a standard autoregressive method.

Because of parameter instability, a method that forecasts well in one sample may not forecast well in another sample period. Thus it is possible that factor models will outperform bagging procedures or viceversa in different sample periods. This problem will be analyzed in future works. At the moment we think that both methods should be used to produce reliable regional GDP forecasts.

An interesting avenue for future research would be the use of non linear factors and methods that allow for structural change and spatial autocorrelation across the regions. Another interesting area for future research would be to compare forecasts from single equations with those from panel estimation.

Table 1: RMSE, $h = 2$, Factor model, All predictors, Regional data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.309	0.160	0.167	0.092	0.062	0.047
Valle d'Aosta	0.735	0.766	0.654	0.673	0.707	0.582
Lombardia	2.957	0.977	1.522	2.666	0.589	1.015
Trentino-Alto Adige	0.655	0.714	0.479	0.655	0.714	0.457
Veneto	0.685	0.634	1.060	0.783	0.731	1.348
Friuli	1.355	0.309	1.246	1.833	0.522	2.267
Liguria	0.167	0.093	1.002	0.186	0.107	1.124
Emilia Romagna	0.723	0.439	1.050	0.730	0.394	1.054
Toscana	0.447	0.337	1.084	0.329	0.276	1.192
Umbria	1.624	1.178	5.188	1.790	1.134	3.882
Marche	1.188	1.793	1.763	1.174	2.606	3.563
Lazio	0.382	0.368	0.381	0.455	0.515	0.521
Abruzzo	1.628	0.997	0.643	1.241	0.899	0.620
Molise	0.454	0.557	0.861	0.488	0.716	1.074
Campania	0.397	0.292	0.328	0.851	0.715	0.766
Puglia	5.613	2.704	1.445	3.986	1.800	1.367
Basilicata	0.851	0.788	1.732	0.836	0.774	1.730
Calabria	0.451	0.398	0.320	0.591	0.441	0.328
Sicilia	1.205	0.615	1.053	1.126	0.587	1.041
Sardegna	3.766	3.357	3.322	4.145	4.077	4.072
Italy	1.028	0.560	0.994	0.969	0.522	1.018
North-West	1.279	0.462	0.900	1.010	0.260	0.556
North-East	0.682	0.424	0.920	0.761	0.467	1.105
Center	1.419	1.083	2.277	0.944	0.890	2.313
North-Center	1.041	0.541	1.147	0.892	0.431	1.067
South	0.994	0.624	0.660	1.172	0.811	0.898
Perc. Regions with RMSE < 1	**0.60**	**0.80**	**0.40**	**0.60**	**0.80**	**0.35**
Perc. Aggreg. with RMSE < 1	**0.33**	**0.83**	**0.68**	**0.68**	**1.00**	**0.33**

Table 2: RMSE, $h = 2$, Factor model, Pre-test strategy, Regional data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.257	0.143	0.151	0.245	0.229	0.242
Valle d'Aosta	0.744	0.780	0.676	0.744	0.780	0.377
Lombardia	2.657	0.600	1.036	2.271	1.242	0.022
Trentino-Alto Adige	0.611	0.729	0.505	0.751	0.792	0.578
Veneto	0.754	0.851	1.451	0.622	0.541	1.058
Friuli	1.312	0.300	2.283	1.691	0.245	2.283
Liguria	0.116	0.079	0.907	0.116	0.079	0.907
Emilia Romagna	0.757	0.503	1.139	0.637	0.839	1.698
Toscana	0.533	0.408	0.891	0.371	0.395	0.868
Umbria	2.204	1.848	6.048	2.204	1.848	6.048
Marche	1.434	2.334	0.814	1.434	2.334	0.814
Lazio	0.436	0.499	0.397	0.420	0.711	1.202
Abruzzo	1.629	1.007	0.652	1.657	0.338	0.226
Molise	0.752	1.044	1.406	0.313	0.278	0.529
Campania	0.410	0.356	0.373	0.365	0.356	0.373
Puglia	1.443	0.799	0.774	1.443	0.799	0.774
Basilicata	0.470	0.455	0.900	0.470	0.455	0.900
Calabria	0.484	0.450	0.374	0.120	0.192	0.182
Sicilia	1.279	0.522	0.954	1.279	0.570	0.954
Sardegna	3.349	2.989	3.073	3.349	2.989	3.073
Italy	0.954	0.508	0.888	0.695	0.422	0.456
North-West	1.108	0.314	0.677	0.603	0.308	0.042
North-East	0.715	0.525	1.183	0.642	0.491	1.282
Center	1.522	1.045	1.427	1.327	0.869	0.588
North-Center	1.016	0.516	1.046	0.745	0.446	0.481
South	0.813	0.509	0.550	0.605	0.377	0.398
Perc. Regions with RMSE < 1	**0.60**	**0.75**	**0.65**	**0.60**	**0.80**	**0.70**
Perc. Aggreg. with RMSE < 1	**0.50**	**0.83**	**0.50**	**0.83**	**1.00**	**0.83**

Table 3: RMSE, $h = 2$, Factor model, All predictors, National data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	1.009	0.944	1.000	0.986	0.864	0.731
Valle d'Aosta	1.141	1.276	1.229	1.147	1.253	1.198
Lombardia	1.850	0.368	0.418	1.864	0.207	0.287
Trentino-Alto Adige	0.010	0.037	0.034	0.010	0.037	0.035
Veneto	0.301	0.291	0.802	0.291	0.280	0.744
Friuli	0.288	0.005	0.218	0.290	0.006	0.218
Liguria	0.093	0.047	0.656	0.092	0.045	0.982
Emilia Romagna	0.278	0.147	0.502	0.281	0.143	0.489
Toscana	0.419	0.369	0.853	0.417	0.381	0.941
Umbria	0.060	0.102	0.134	0.071	0.073	0.516
Marche	0.804	0.918	0.850	0.805	0.910	0.827
Lazio	0.507	0.716	0.520	0.524	0.714	0.518
Abruzzo	1.183	0.638	0.975	1.182	0.533	0.932
Molise	0.072	0.098	0.248	0.066	0.046	0.217
Campania	0.719	0.533	0.556	0.719	0.681	0.871
Puglia	3.132	1.857	2.975	3.111	0.787	0.710
Basilicata	0.157	0.156	0.331	0.159	0.137	0.287
Calabria	1.402	0.819	0.776	1.407	0.858	0.890
Sicilia	0.387	0.210	0.474	0.373	0.208	0.666
Sardegna	0.421	0.334	0.494	0.436	0.349	0.480
Italy	0.651	0.342	0.668	0.651	0.310	0.661
North-West	1.079	0.402	0.616	1.082	0.292	0.512
North-East	0.204	0.121	0.489	0.204	0.118	0.467
Center	1.127	0.896	1.334	1.148	0.932	1.463
North-Center	0.611	0.310	0.677	0.610	0.272	0.639
South	0.765	0.470	0.648	0.762	0.455	0.713
Perc. Regions with RMSE < 1	**0.70**	**0.90**	**0.85**	**0.75**	**0.95**	**0.95**
Perc. Aggreg. with RMSE < 1	**0.67**	**1.00**	**0.83**	**0.67**	**1.00**	**0.83**

Table 4: RMSE, $h = 2$, Factor model, Pre-test strategy, National data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.179	0.109	0.133	0.179	0.109	0.133
Valle d'Aosta	1.619	1.578	1.676	1.619	1.578	1.676
Lombardia	0.139	0.023	0.068	0.075	0.023	0.068
Trentino-Alto Adige	0.090	0.014	0.139	0.090	0.014	0.139
Veneto	0.081	0.064	0.133	0.081	0.064	0.133
Friuli	0.058	0.243	0.011	0.103	0.243	0.011
Liguria	1.060	0.309	6.578	1.060	0.309	7.343
Emilia Romagna	0.128	0.046	0.266	0.128	0.009	0.266
Toscana	0.421	0.398	0.857	0.421	0.398	0.857
Umbria	0.013	0.008	0.004	0.041	0.017	0.009
Marche	0.166	0.210	0.200	0.166	0.210	0.200
Lazio	6.026	7.065	7.006	4.546	5.144	5.027
Abruzzo	0.539	0.424	0.562	0.539	0.424	0.562
Molise	0.368	0.277	0.420	0.318	0.330	0.535
Campania	1.663	1.532	1.745	1.663	1.532	4.617
Puglia	0.154	0.031	0.047	0.154	0.031	0.047
Basilicata	0.017	0.015	0.030	0.050	0.099	0.030
Calabria	12.457	15.742	19.216	12.457	14.206	17.525
Sicilia	0.295	0.517	1.129	0.736	2.100	0.213
Sardegna	1.132	0.102	0.368	0.519	0.088	0.106
Italy	0.154	0.071	0.185	0.117	0.046	0.201
North-West	0.162	0.043	0.123	0.112	0.043	0.127
North-East	0.081	0.029	0.160	0.085	0.017	0.160
Center	0.177	0.176	0.211	0.133	0.128	0.170
North-Center	0.109	0.043	0.135	0.091	0.037	0.136
South	0.480	0.352	0.524	0.304	0.100	0.665
Perc. Regions with RMSE $<$ 1	**0.70**	**0.80**	**0.70**	**0.75**	**0.75**	**0.75**
Perc. Aggreg. with RMSE $<$ 1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 5: RMSE, $h = 2$, Factor model, All predictors, Mixed data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.077	0.052	0.064	0.004	0.012	0.050
Valle d'Aosta	0.576	0.649	0.452	1.118	1.377	0.976
Lombardia	0.246	0.096	0.121	0.024	0.058	0.006
Trentino-Alto Adige	0.488	0.473	0.348	0.484	0.465	0.337
Veneto	0.222	0.222	0.317	0.138	0.127	0.125
Friuli	1.680	0.675	2.431	3.349	1.770	3.353
Liguria	2.457	1.962	9.301	4.636	4.206	18.205
Emilia Romagna	0.695	0.551	0.951	0.478	0.257	0.852
Toscana	0.615	0.589	1.115	0.422	0.358	1.648
Umbria	0.063	0.054	0.208	0.068	0.057	0.650
Marche	0.350	0.487	0.469	0.463	0.646	0.716
Lazio	2.721	2.940	2.902	2.588	3.364	3.331
Abruzzo	0.635	0.401	0.282	0.515	0.380	0.285
Molise	0.446	0.641	1.239	1.066	2.091	3.184
Campania	1.277	1.011	1.097	1.283	1.009	1.122
Puglia	0.072	0.033	0.048	0.079	0.044	0.109
Basilicata	0.389	0.432	0.584	0.364	0.406	0.609
Calabria	7.452	6.058	6.018	7.032	5.802	5.329
Sicilia	2.543	1.521	2.495	3.475	2.108	3.868
Sardegna	7.264	6.474	6.294	12.417	10.655	10.696
Italy	0.362	0.235	0.363	0.196	0.072	0.215
North-West	0.222	0.102	0.160	0.023	0.018	0.002
North-East	0.414	0.333	0.558	0.322	0.211	0.395
Center	0.369	0.298	0.476	0.212	0.159	0.612
North-Center	0.299	0.194	0.321	0.114	0.025	0.134
South	0.728	0.484	0.585	0.850	0.604	0.825
Perc. Regions with RMSE < 1	**0.65**	**0.70**	**0.55**	**0.55**	**0.55**	**0.55**
Perc. Aggreg. with RMSE < 1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 6: RMSE, $h = 2$, Factor model, Pre-test strategy, Mixed data set.

	$M = 1$			$M_{max} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.031	0.021	0.035	0.060	0.043	0.005
Valle d'Aosta	0.553	0.494	0.521	0.658	0.515	0.447
Lombardia	0.395	0.059	0.097	0.617	0.068	0.097
Trentino-Alto Adige	0.449	0.522	0.370	0.443	0.175	0.383
Veneto	0.134	0.140	0.241	0.130	0.104	0.156
Friuli	1.464	0.329	1.299	1.560	0.469	1.907
Liguria	3.214	2.433	6.993	0.289	1.913	8.087
Emilia Romagna	0.499	0.303	0.705	0.413	0.332	1.354
Toscana	0.559	0.410	0.990	0.429	0.345	1.572
Umbria	0.849	0.849	1.652	0.057	0.048	0.060
Marche	0.313	0.437	0.439	0.184	0.298	0.292
Lazio	3.001	3.597	2.034	2.961	3.253	1.842
Abruzzo	0.692	0.430	0.276	1.254	0.525	0.937
Molise	0.520	0.464	1.271	0.534	0.420	1.285
Campania	1.308	1.006	0.886	0.851	0.833	0.845
Puglia	0.041	0.029	0.113	0.035	0.027	0.111
Basilicata	0.079	0.096	0.075	0.026	0.039	0.019
Calabria	7.300	9.541	8.347	6.448	41.070	57.502
Sicilia	1.236	0.381	2.978	0.717	0.424	2.893
Sardegna	4.888	4.465	4.843	5.275	5.132	6.201
Italy	0.327	0.158	0.320	0.311	0.137	0.361
North-West	0.293	0.061	0.115	0.406	0.065	0.084
North-East	0.277	0.186	0.397	0.257	0.124	0.511
Center	0.454	0.419	0.723	0.239	0.373	0.726
North-Center	0.299	0.135	0.276	0.297	0.105	0.282
South	0.490	0.302	0.570	0.371	0.341	0.833
Perc. Regions with RMSE < 1	**0.65**	**0.75**	**0.60**	**0.75**	**0.80**	**0.55**
Perc. Aggreg. with RMSE < 1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 7: RMSE, $h = 2$, Bagging procedure, Regional data set.

	$N^{*h} = 1$			$N^{*h} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.112	0.106	0.085	0.167	0.111	0.057
Valle d'Aosta	0.900	0.919	1.512	1.204	1.249	0.359
Lombardia	0.427	0.026	0.295	0.522	0.226	0.252
Trentino-Alto Adige	0.273	0.488	0.296	0.223	0.479	0.269
Veneto	0.196	0.215	0.238	0.158	0.058	0.289
Friuli	2.282	0.877	2.716	1.756	0.302	1.502
Liguria	1.510	0.440	6.303	1.746	1.173	8.814
Emilia Romagna	0.618	0.411	0.855	0.610	0.400	0.089
Toscana	0.450	0.411	0.819	0.588	0.411	1.011
Umbria	0.127	0.159	0.379	0.025	0.145	0.244
Marche	0.353	0.543	0.479	0.336	0.480	0.518
Lazio	2.957	2.339	3.252	3.732	5.099	2.274
Abruzzo	0.678	0.463	0.814	0.901	0.375	0.497
Molise	1.310	1.033	2.442	1.518	1.579	1.854
Campania	3.263	1.843	0.398	1.605	2.464	0.659
Puglia	0.182	0.104	0.034	0.210	0.093	0.150
Basilicata	0.075	0.152	0.415	0.100	0.053	0.227
Calabria	9.353	9.746	6.525	3.169	7.748	9.357
Sicilia	2.678	3.282	3.329	3.554	1.509	2.842
Sardegna	6.534	7.437	7.584	8.071	7.782	8.265
Italy	0.474	0.238	0.403	0.419	0.227	0.312
North-West	0.348	0.050	0.287	0.429	0.212	0.253
North-East	0.372	0.300	0.479	0.319	0.128	0.220
Center	0.504	0.356	0.483	0.231	0.160	0.801
North-Center	0.370	0.157	0.370	0.334	0.156	0.304
South	1.104	0.861	0.579	0.932	0.752	0.347
Perc. Regions with RMSE $<$ 1	**0.60**	**0.70**	**0.60**	**0.55**	**0.60**	**0.60**
Perc. Aggreg. with RMSE $<$ 1	**0.83**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 8: RMSE, $h = 2$, Bagging procedure, National data set.

	$N^{*h} = 1$			$N^{*h} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.239	0.217	0.227	0.321	0.081	0.185
Valle d'Aosta	0.550	1.318	0.613	0.639	0.662	0.329
Lombardia	0.291	0.058	0.119	0.165	0.023	0.190
Trentino-Alto Adige	0.083	0.074	0.165	0.022	0.232	0.052
Veneto	0.126	0.118	0.188	0.137	0.094	0.188
Friuli	0.597	0.016	0.466	1.151	0.028	0.031
Liguria	1.303	0.933	6.354	0.111	0.550	10.370
Emilia Romagna	0.405	0.188	0.386	0.225	0.136	0.280
Toscana	0.520	0.456	1.104	0.633	0.457	1.153
Umbria	0.003	0.007	0.004	0.014	0.002	0.024
Marche	0.243	0.351	0.311	0.319	0.206	0.277
Lazio	3.526	4.306	3.365	5.317	3.419	7.753
Abruzzo	0.732	0.354	0.522	0.379	0.224	0.374
Molise	0.810	0.705	0.148	0.386	0.118	0.956
Campania	3.311	1.537	2.554	2.608	3.263	1.905
Puglia	0.187	0.127	0.116	0.103	0.013	0.126
Basilicata	0.052	0.042	0.073	0.046	0.016	0.120
Calabria	18.446	17.863	15.774	27.209	13.240	16.926
Sicilia	2.098	0.897	1.500	0.139	0.353	2.150
Sardegna	2.678	1.399	1.686	2.484	0.206	2.073
Italy	0.347	0.179	0.303	0.214	0.097	0.294
North-West	0.291	0.104	0.195	0.195	0.036	0.265
North-East	0.215	0.122	0.263	0.150	0.097	0.193
Center	0.276	0.279	0.299	0.201	0.196	0.279
North-Center	0.242	0.126	0.229	0.163	0.072	0.223
South	1.081	0.575	0.791	0.564	0.280	0.783
Perc. Regions with RMSE < 1	**0.70**	**0.75**	**0.65**	**0.75**	**0.85**	**0.65**
Perc. Aggreg. with RMSE < 1	**0.83**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 9: RMSE, $h = 2$, Bagging procedure, Mixed data set.

	$N^{*h} = 1$			$N^{*h} = 5$		
Regions	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Piemonte	0.058	0.176	0.085	0.131	0.288	0.042
Valle d'Aosta	1.184	0.687	0.408	1.381	0.514	3.454
Lombardia	0.454	0.261	0.164	1.090	0.198	0.020
Trentino-Alto Adige	0.275	0.637	0.318	0.094	0.410	0.170
Veneto	0.243	0.206	0.185	0.203	0.454	0.172
Friuli	1.953	2.761	3.523	4.919	6.863	5.886
Liguria	0.511	4.007	9.137	9.231	0.290	4.765
Emilia Romagna	0.641	0.992	0.669	0.796	0.408	0.897
Toscana	0.785	0.724	1.043	0.168	1.062	1.230
Umbria	0.320	0.106	0.980	0.119	0.170	1.253
Marche	0.470	0.572	0.534	0.175	1.245	0.652
Lazio	2.800	3.253	3.153	3.277	3.456	4.152
Abruzzo	0.648	0.450	0.496	0.591	0.688	0.835
Molise	1.096	0.762	1.157	0.935	1.704	3.163
Campania	1.110	3.385	0.993	0.232	2.109	1.907
Puglia	0.081	0.045	0.044	0.098	0.028	0.069
Basilicata	0.159	0.109	0.214	0.123	0.052	0.023
Calabria	12.424	11.035	7.312	15.240	2.996	5.598
Sicilia	2.787	0.206	3.224	2.635	0.821	9.551
Sardegna	8.510	8.052	7.377	4.588	3.691	7.780
Italy	0.469	0.427	0.368	0.485	0.321	0.296
North-West	0.318	0.284	0.200	0.872	0.049	0.033
North-East	0.404	0.514	0.398	0.399	0.556	0.429
Center	0.800	0.493	0.651	0.085	1.199	0.497
North-Center	0.412	0.379	0.323	0.483	0.309	0.191
South	0.775	0.667	0.618	0.504	0.397	1.148
Perc. Regions with RMSE < 1	**0.60**	**0.70**	**0.60**	**0.60**	**0.60**	**0.45**
Perc. Aggreg. with RMSE < 1	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**0.83**

References

Bai, J., and S. Ng (2002), Determining the Number of Factors in Approximate Factor Models, *Econometrica*, **70**, 191-221.

Bai, J., and S. Ng (2004), A Panic Attack on Unit Roots and Cointegration, *Econometrica*, **72**, 1127-1177.

Bai, J., and S. Ng (2006a), Forecasting Economic Time Series Using Targeted Predictors, mimeo, Department of Economics, University of Michigan.

Bai, J., and S. Ng (2006b), Confidence Interval for Diffusion Index Forecasts and Inference with Factor-Augmented Regressions, *Econometrica*, forthcoming.

Boivin, J., and S. Ng (2006), Are More Data Always Better for Factor Analysis?, *Journal of Econometrics*, **132**, 169-194.

Bühlmann, P., and B. Yu (2002), Analyzing Bagging, *Annals of Statistics*, **30**, 927-961.

Breiman, L. (1996), Bagging Predictors, *Machine Learning*, **36**, 105-139.

Cecchetti, S., R. Chu, and C. Steindel (2000), The Unrealibility of Inflation Indicators, *Federal Reserve Bank of New York Current Issue in Economics and Finance*, **6**, 1-6.

Delle Monache D., L. Gutierrez, and G. Pellegrini (2006), Econometric Analysis of Short Term Models Aimed to Predict GDP in North-Center and South of Italy, mimeo, Department of Cohesion and Development Policies, Ministry of Economic Development, (in Italian).

Gonçalves, S. and H. White (2004), Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models, *Journal of Econometrics*, **119**, 199-220.

Inoue, A., and L. Kilian (2005), How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S, CPI Inflation, mimeo, Department of Economics, University of Michigan.

Stock, J., and M.W. Watson (2002), Forecasting Using Principal Components from a Large Number of Predictors, *Journal of American Statistical Association*, **97**, 1167-1179.

Stock, J., and M.W. Watson (2003), Forecasting Output and Inflation: The Role of Asset Prices, *Journal of Economic Literature*, **41**, 788-829.

West, K. (1997), Another Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator, *Journal of Econometrics*, **76**, 171-191.

White, H. (1980), A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heterogeneity, *Econometrica*, **48**, 817-838.