Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Forecasting EMU Macroeconomic Variables

Massimiliano Marcellino

Working Paper n. 216

June 2002

IGIER – Università Bocconi, Via Salasco 5, 20136 Milano –Italy
http://www.igier.uni-bocconi.it

Forecasting EMU macroeconomic variables *

Massimliano Marcellino **

IEP- Bocconi Univerisity, IGIER and CEPR

June 2002

Abstract

After the creation of the European Monetary Union (EMU), both the European Commission (EC) and the European Central Bank (ECB) are focusing more and more on the evolution of the EMU as a whole, rather than on single member countries. A particularly relevant issue from a policy point of view is the availability of reliable forecasts for the key macroeconomic variables. Hence, both the fiscal and the monetary authorities have developed aggregate forecasting models, along the lines previously adopted for the analysis of single countries. A similar approach will be likely followed in empirical analyses on, e.g., the existence of an aggregate Taylor rule or the evaluation of the aggregate impact of monetary policy shocks, where linear specifications are usually adopted. Yet, it is uncertain whether standard linear models provide the proper statistical framework to address these issues. The process of aggregation across countries can produce smoother series, better suited for the analysis with linear models, by averaging out country specific shocks. But the method of construction of the aggregate series, which often involves time-varying weights, and the presence of common shocks across the countries, such as the deflation in the early '80s and the convergence process in the early '90s, can introduce substantial non-linearity into the generating process of the aggregate series. To evaluate whether this is the case, we fit a variety of non-linear and time-varying models to aggregate EMU macroeconomic variables, and compare them with linear specifications. Since non-linear models often over-fit in sample, we assess their performance in a real time forecasting framework. It turns out that for several variables linear models are beaten by non-linear specifications, a result that questions the use of standard linear methods for forecasting and modeling EMU variables.

J.E.L. Classification: *C2, C53, E30*

Keywords: European Monetary Union, Forecasting, Time-Varying models, Non-linear models, Instability, Non-linearity,

---

* I am grateful to Jim Stock and Mark Watson for several conversations on the topics addressed in this paper, and for making available the GAUSS programs that form the basis of the code used in this paper. ISF2002 participants also provided valuable comments. The usual disclaimers apply.

** Correspondence to: IGIER, Università Bocconi, Via Salasco 3/5, 20136, Milan, Italy.
E-mail: massimiliano.marcellino@uni-bocconi.it Phone: +39-02-5836-3327 Fax: +39-02-5836-3302

1. Introduction

With the creation of the European Monetary Union, the focus of the economic policy debate and of macroeconomic analyses is shifting more and more from the single member countries to the Union as a whole. This poses the question of the proper econometric methods for empirical analyses in the new context, and this paper contributes to the debate by studying the relative forecasting performance of a large set of univariate linear, time-varying, and non-linear models.

In a first stage, both economists and policy makers have analyzed the Union by looking at differences and similarities among the member countries. For example, several authors estimated linear Taylor rules for the largest countries in the union, e.g. Clarida, Gali and Gertler (1998), Gerlach and Schnabel (2000), Favero and Marcellino (2001), to evaluate whether there were major differences across central banks in the conduct of monetary policy in the recent past. Similarly, many studies focused on the analysis of asymmetries across countries in the transmission of monetary policy, and several of them used linear models (VARs) see e.g. Clements, Kontolemis, Levy (2001), Dornbusch, Favero, Giavazzi (1998), Favero and Marcellino (2001). Also, when using linear models, aggregation of country specific forecasts for key macroeconomic variables in general produced better results than direct forecasts of the aggregates, see e.g. Marcellino, Stock and Watson (2001).

Are the same linear techniques appropriate for modeling EMU macroeconomic variables or different econometric methods are required? In particular, is there a role for models with time-varying parameters or non-linear components? These questions can be hardly answered from a theoretical point of view, they require a thorough empirical analysis. In this paper we focus on univariate models, leaving the more complex analysis of multivariate specifications for future research.

In general, more complicated models tend to fit well in sample, better than linear specifications, because of their greater flexibility. Yet, often the resulting models are too specific for the particular estimation sample, and their good performance is not replicated out of sample, in a forecasting context. For example, Stock and Watson (1999) found only minor gains from complicated models for forecasting about 200 macroeconomic variables for the US, on average linear specifications were quite good, and pooled forecasts even better. This suggests running the model comparison not in sample but out of sample, in a pseudo real time forecasting exercise.

Marcellino (2002a, 2002b) conducted a comparison similar to that by Stock and Watson (1999), but focusing on EMU countries. The main result was that on average linear models or pooling methods were still the best, but non-linear specifications were forecasting well for a good percentage of the macroeconomic variables under analysis. Basically, while linear models forecast reasonably well for all the series, non-linear models are quite good for some series but quite bad for others, so that on average they are beaten by simpler methods.

In this paper we evaluate whether this is the case also for EMU aggregate macroeconomic variables. In particular, since the official Eurostat EMU series are too short for fitting complex models, we consider longer series of aggregated data constructed at the ECB by Fagan, Henry and Mestre (2001). We then also focus on two other aggregate datasets, by Marcellino, Stock and Watson (2001), and Beyer, Doornik and Hendry (2001). Section 2 provides more details on the variables under analysis, which include real gdp and its components, prices, interest rates, the real exchange rate and trade variables.

We compare three main forecasting methods. The linear method, which includes autoregressive (AR) models, exponential smoothing and random walk models. The time-varing method, which includes time-varying AR models and smooth transition AR models. The non-linear method, that includes artificial neural network models. Within each method we consider several alternative specifications, for a total of 58 models, and different forecast horizons. The models are described in details in Section 3.

The competing forecasts are compared on the basis of three measures, in increasing level of disaggregation. First, we compute the average value over all variables of several loss functions, including the common mean absolute and mean square forecast error. Second, we rank the models on the basis of the percentage of variables for which they are among the top-N models, for several values of N. Finally, for each forecasting model we compute the empirical distribution function (over variables) of its mean square forecast error, relative to a benchmark AR model, and selected percentiles of this distribution. The results of the comparison are reported in Section 4. We can anticipate that models outside the linear class perform well for about 40% of the variables, including the components of gdp and the real exchange rate, though on average they are outperformed by the linear models.

When none of the models under analysis coincides with the generating mechanism of the series, it can be possible to improve the forecasting performance by combining together the different forecasts. The three most common methods in practice are linear combination, with weights related to the mean square forecast error of each forecast, median forecast selection, and predictive least squares, where a single model is chosen, but the selection is recursively updated at each forecasting period on the basis of the past forecasting performance. The relative merits of these combination methods are analyzed in Section 5.

In Section 6 we address three additional issues. First, we detect and discuss the best forecasting model for each of the variables under analysis. Second we evaluate whether the good performance of models outside the linear class can be explained by the presence of extensive instability in the variables. Third, we evaluate the robustness of the results by repeating the analysis with the datasets by Marcellino, Stock and Watson (2001), and Beyer, Doornik and Hendry (2001).

Section 7 summarizes the main findings of the paper and offers suggestions for further extensions.

2. The data

The construction of aggregate data for the EMU area poses several problems. A partial list includes: interpolation of missing observations and disaggregation of annual figures into quarterly data, which are not available for all variables and countries in the EMU for a long enough time period; seasonal adjustment, working day adjustment and treatment of major redefinitions and institutional changes, such as the German reunification; and choice of the aggregation (over countries) method.

We do not discuss these issues in detail here, or construct our own EMU series. Instead, we analyze the aggregate variables in Fagan, Henry and Mestre (2001), Marcellino, Stock and Watson (2001), and Beyer, Doornik and Hendry (2001), to whom we refer for additional details on the Euro zone data reconstruction. Unfortunately, the official series for the EMU area, released by Eurostat, start in the early ’90s and are too short for the forecasting exercise we wish to conduct.

Fagan, Henry and Mestre (2001) construct an area wide macroeconometric model for the euro area, and estimate it using quarterly data starting in 1970. The aggregation method adopted is the so-called “Index Method” where, for instance, the log of aggregate gdp is a weighted sum of the logs of the country specific gdps. The weight for each country is constant over time, and equal to its real gdp share in 1995, see also Fagan and Henry (1998) for more details. The series are available from the ECB web site, with different ending dates in 1997-1998. We consider the following 15 macroeconomic variables, for the period 1970:1-1997:4: real gdp (YER) and its components, namely, personal and government consumption (PCR and GCR), investment and inventories (ITR and SCR), and imports (MTR) and exports (XTR); consumer prices (HICP) and the gdp deflator (YED); unit labor cost (ULC) and unemployment (URX); short-term and long-term interest rates (STN and LTN); and the real exchange rate (EER) and the trade balance (TBR).

Marcellino, Stock and Watson (2001) follow a similar aggregation method but construct monthly series for the period 1982:1-1997:8. The resulting series are rather close to the official Eurostat variables over the ‘90s. From their dataset, we analyze industrial production (IPMSW) and consumer prices (CPIMSW), while we do not consider the unemployment rate since the series starts later on.

Finally, Beyer, Doornik and Hendry (2001) adopt a more sophisticated aggregation procedure, which is based on the aggregation of growth rates of the variables, with time-varying weights, whose evolution depends also on the behavior of the exchange rate. The Euro zone aggregate growth series are then cumulated to obtain the levels of the variables. The method yields real gdp (GDPBDH), the gdp deflator (YEDBDH) and M3 (M3BDH), monthly, over the period 1980:1-1999:2.

3 Forecasting methods

As in Stock and Watson (1999) and Marcellino (2002a,b), we consider forecasting models of the type

$$y^h_{t+h} = f(Z_t;\theta_{ht}) + \varepsilon_{t+h}, \qquad (1)$$

where y_t is the variable being forecast, *h* indicates the forecast horizon, Z_t is a vector of predictor variables, ε_t is an error term, and θ_h is a vector of parameters, possibly evolving over time. It is worth distinguishing between forecasting methods and forecasting models. Forecasting methods differ for the specification of the f function, i.e., the form of the relationship between y^h_{t+h} and Z_t. Each method contains several models, which differ for the choice of the regressors Z_t and the stationarity transformation applied to y_t.

The h-step forecast can be written as

$$\hat{y}^h_{t+h} = f(Z_t;\hat{\theta}_{ht}), \quad (2)$$

and the forecast error is

$$e_{t+h} = y^h_{t+h} - \hat{y}^h_{t+h}. \quad (3)$$

When y_t is treated as stationary, it is $y^h_{t+h} = y_{t+h}$, while if y_t is integrated then $y^h_{t+h} = y_{t+h} - y_t$. We present results for both cases. Moreover, we also consider a pre-test forecast where the decision on the stationarity of y_t is based on a unit root test, which often improves the forecasting performance, see e.g. Diebold and Kilian (2000). In particular, we use the Elliot, Rothenberg and Stock (1996) DF-GLS statistics, which performed best in the simulation experiments in Stock (1996). Since $e_{t+h} = y_{t+h} - \hat{y}_{t+h}$, independently of whether y_t is treated as stationary or not, the forecast errors from the three different cases (stationary, I(1) and pre-test) are directly comparable.

We consider forecasts 1, 2 and 4 quarters ahead. When the forecast horizon, *h*, is larger than one, the "*h*-step ahead projection" approach in (1), also called dynamic estimation (e.g. Clements and Hendry (1996)), differs from the standard approach of estimating a one-step ahead model, then iterating that model forward to obtain *h*-step ahead predictions. The *h*-step ahead projection approach has two main advantages in this context. First, the potential impact of specification error in the one-step ahead model can be reduced by using the same horizon for estimation as for forecasting. Second, we need not resort to simulation methods to obtain forecasts from non-linear models. The resulting forecasts could be slightly less efficient, see e.g. Granger and Terasvirta (1993, Ch.8), but the computational savings in our real time exercise with several series are substantial.

Because of possible problems in the estimation of non-linear and time-varying methods with the rather short sample available, a few forecast errors can be very large. In order not to bias the comparison against these methods, we automatically trim the forecasts. In particular, when the absolute value of a forecasted change is larger than any previously observed change, a no change forecast is used.

We now list the methods and models we compare, and briefly discuss their main characteristics and estimation issues, see Stock and Watson (1996, 1999) for additional details.

Linear methods

Autoregression (AR). Since the pioneering work by Box and Jenkins (1970), the good performance of these models for forecasting economic variables has been confirmed by several studies, see e.g. Meese and Geweke (1984), or Marcellino, Stock and Watson (2001) for the Euro area. The f function in (1) is linear, and Z_t only includes lags of the y variable and a deterministic component. The latter can be either a constant or also a linear trend. The lag length is either fixed at 4, or it is chosen by AIC or BIC with a maximum of 4 lags. Since the y_t variable can be treated as stationary, I(1), or pre-tested for unit roots, overall there are 18 models in this class.

Exponential smoothing (ES). Makridakis et al. (1982) found this simple method to perform rather well in practice in a forecast comparison exercise, even though from a theoretical point of view the resulting forecasts are optimal in the mean square forecast error sense only when the underlying process follows a particular ARMA structure, see e.g. Granger and Newbold (1986, Ch.5). We consider both single and double exponential smoothing, which are usually adopted for, respectively, stationary and trending series.

Estimation of the parameters is conducted by means of (recursive) non-linear least squares (see e.g. Tiao and Xu (1993)). The third model in this class is given by a combination of the single and double forecasts, based on the outcome of the unit root test.

No change. When a pure random walk model is adopted, the resulting forecast is $\hat{y}_{t+h} = y_t$. Notwithstanding its simplicity, in a few cases it was found to outperform even forecasts from large-scale structural models, see e.g. Artis and Marcellino (2001).

Time-varying methods

Time-varying autoregression (TVAR). The parameters of the AR models are allowed to evolve according to the following multivariate random walk model (see e.g. Nyblom (1989)):

$$\theta_{ht} = \theta_{ht-1} + u_{ht}, \qquad u_{ht} \sim iid(0, \lambda^2 \sigma^2 Q), \quad (4)$$

where σ^2 is the variance of the error term ε in (1), $Q = (E(Z_t Z_t^{'}))^{-1}$, and we inspect several values of λ : 0 (no evolution), 0.0025, 0.005, 0.0075, 0.01, 0.015, or 0.020. We consider a fixed specification with a constant, 2 lags and λ = 0.005, and models where the number of lags (1,2,4) is jointly selected with the value of λ by either AIC or BIC. In each case, y_t can be either stationary, or I(1) or pre-tested, so that we have a total of 9 TVAR models. The models are estimated by the Kalman filter.

Logistic smooth transition autoregression (LSTAR). The generic model is

$$y_{t+h}^{h} = \alpha^{'} \zeta_t + d_t \beta^{'} \zeta_t + \varepsilon_{t+h}, \qquad (5)$$

where $d_t = 1/(1+\exp(\gamma_0 + \gamma_1 \zeta_t))$, and $\zeta_t = (1, y_t, y_{t-1}, ..., y_{t-p+1})$ if y_t is treated as stationary or $\zeta_t = (1, \Delta y_t, \Delta y_{t-1}, ..., \Delta y_{t-p+1})$ if y_t is I(1). The smoothing parameters γ_1 regulate the shape of parameter change over time. When $\gamma_1 = 0$ the model becomes linear, while for large values of γ_1 the model tends to a self-exciting threshold model (SETAR), see e.g. Granger and Terasvirta (1993), Terasvirta (1998) for details. For models specified in levels we consider the following choices for the threshold variable in d_t: $\zeta_t = y_t$, $\zeta_t = y_{t-1}$, $\zeta_t = y_{t-3}$, $\zeta_t = y_t - y_{t-2}$, $\zeta_t = y_t - y_{t-4}$. For differenced variables, it can be $\zeta_t = \Delta y_t$, $\zeta_t = \Delta y_{t-1}$, $\zeta_t = \Delta y_{t-3}$, $\zeta_t = y_t - y_{t-2}$, $\zeta_t = y_t - y_{t-4}$. In each case the lag length of the model was either 1 or 2 or 4. We report results for the following models: 2 lags and $\zeta_t = y_t$ (or $\zeta_t = \Delta y_t$ for the I(1) case); 2 lags and $\zeta_t = y_t - y_{t-2}$; AIC or BIC selection of both the number of lags and the specification of ζ_t. In each case y_t can be either stationary or I(1) or pre-tested, which yields a total of 12 LSTAR models. Estimation is carried out by (recursive) non-linear least squares, using an optimizer developed by Stock and Watson (1999).

Non-linear methods

Artificial neural network (ANN). Artificial neural networks can provide a valid approximation to the generating mechanism of a vast class of non-linear processes, see e.g. Hornik, Stinchcombe and White (1989), and Swanson and White (1997) for their use as forecasting devices. The so called single layer feedforward neural network model with n_1 hidden units (and a linear component) is specified as:

$$y_{t+h}^{h} = \beta_0^{'}\zeta_t + \sum_{i=1}^{n_1} \gamma_{1i} g(\beta_{1i}^{'}\zeta_t) + \varepsilon_{t+h}, \qquad (6)$$

where $g(x)$ is the logistic function, $g(x) = 1/(1+e^x)$. A more complex model is the double layer feedforward neural network with n_1 and n_2 hidden units:

$$y_{t+h}^{h} = \beta_0^{'}\zeta_t + \sum_{j=1}^{n_2} \gamma_{2i} g\left(\sum_{i=1}^{n_1} \beta_{2ji} g(\beta_{1i}^{'}\zeta_t)\right) + \varepsilon_{t+h}. \quad (7)$$

We report results for the following specifications: n_1=2, n_2=0,p=2 (recall that p is number of lags in ζ_t); n_1=2, n_2=1,p=2; n_1=2, n_2=2,p=2; AIC or BIC selection with n_1=(1,2,3), n_2=(1,2 with n_1=2),p=(1,2). For each case y_t can be either stationary, or I(1) or pre-tested, which yields a total of 15 ANN models. The models are estimated by (recursive) non-linear least squares, using an algorithm developed by Stock and Watson (1999).

Overall, there are 58 models in the forecast comparison exercise, 22 belong to the linear class, 21 are time-varying, and 15 are non-linear. They are summarized in Table 1. To mimic real time situations, for each variable, method and model the unit-root tests, estimation and model selection are repeated each month over the forecasting period, which is 1990:1-1997:4.

4. Forecast Evaluation

In this section we compare the forecasting performance of the 58 models for the 15 macroeconomic variables in the ECB dataset from Fagan, Henry and Mestre (2001). Results for the other datasets are summarized in Section 6.3.

To start with, we need to choose a loss function. We define the loss from model *m* for variable *n* as

$$Loss^{h}{}_{n,m} = \frac{1}{T-h}\sum_{t=1}^{T-h}\left|e_{t+h,n,m}\right|^{\rho}, \qquad (8)$$

where e_{t+h} is the h-step ahead forecast error, and ρ can be equal to 1, 1.5, 2, 2.5 or 3. The values $\rho = 1$ and $\rho = 2$ correspond to the familiar choices of, respectively, the mean absolute error (mae) and the mean square forecast error (msfe) as the loss function.

In order to compare the loss over the whole set of N=15 variables, we adopt the following loss function for model *m*:

$$Loss^{h}{}_{m} = \frac{1}{N}\sum_{n=1}^{N}\frac{Loss^{h}{}_{n,m}}{Loss^{h}{}_{n,1}}, \qquad (9)$$

namely, a weighted average of the loss for each variable, with weights given by the inverse of the loss of a benchmark forecast, which makes the magnitude of the losses comparable across variables. As a benchmark, we adopt throughout an AR model with 4 lags and a constant, specified in levels.

In Table 2 we report the ranking of the models, for different values of ρ. Four main results emerge. First, the best models for each value of ρ and for each forecast horizon are linear. Second, it is worth imposing the presence of a unit root and choosing the lag length by BIC or AIC. Third, the AR models with random parameters rank second. Fourth, simple forecasts (ES and No Change) and complex forecasts (STAR and ANN) are never in the top-10 models.

Overall these results indicate that a linear structure, possibly with random parameters, provides a proper representation for the 15 variables under analysis, while more complex models do not yield any forecasting gains. Yet, the bad performance of STAR and ANN models can be due to the averaging over variables underlying the ranking in Table 2. These models could forecast well for a few variables and values of *h*, but very badly for the remaining variables, so that on average they are beaten by simpler specifications. To evaluate whether this is the case, in Table 3a we report the number of series for which a given method yields the lowest msfe (results for each model are available upon request).

The good qualities of constant and time-varying parameter AR models are confirmed, as well as the bad ones of exponential smoothing and random walk forecasts. But STAR and ANN models are now the best for 40% of the variables. A thorough variable by variable analysis is conducted in Section 6 below, but we can anticipate that STAR and ANN models work particularly well for the components of gdp, and for the real exchange rate.

In Table 3b we provide additional results for the models that perform best in the msfe sense within each method. In particular, we compute the number of series for which these models are the best or among the top 5, 10, 15 and 20. It turns out that the best AR and TVAR models are more often also in the top-20 than the best STAR or ANN models. This confirms the intuition that complex models work well for a few series and badly for the other ones.

To provide further evidence on this intuition, and to gather additional information on the relative merits of the competing models, we take an even more disaggregate approach. First, for each variable we compute the relative msfe (rmsfe) of each forecasting model with respect to the benchmark AR(4), so that an rmsfe higher than one indicates that the method under analysis is worse than the benchmark. In formulae, the rmsfe of model *j* for variable *m* is:

$$rmsfe^{h}{}_{j-AR4,m} = \left(\sum_{t=1}^{T-h} e^{2}_{j,t+h,m}\right) \Big/ \left(\sum_{t=1}^{T-h} e^{2}_{AR4,t+h,m}\right). \qquad (10)$$

Then, for each model, we calculate the empirical distribution of the rmsfe over the variables. In Table 4 we report the mean of the distribution and some percentiles for selected models (the best in Table 2 and those in Table 3b, results for all models are available upon request).

Two main comments are in order. First, the mean of the distribution underlies the ranking in Table 2 for $\rho = 2$. Actually, the model ARFT1b has the lowest average msfe for *h*=1,2, and ARFT1a for *h*=4. Their performance is rather similar, with average gains with respect to the benchmark AR4 of about 10% when h=1,2, 20% when h=4. The ARTVFC03 is a close second best for h=1,2, but it is worse when h=4.

Second, the LSP063 and, in particular, the ANN0203 models are substantially worse than the benchmark on average. Yet, looking at the percentiles of the distribution of the relative msfe, they are better than the benchmark for at least 25% of the series. Notice also that the median of the distribution is substantially lower than the mean for these models, while it is very close for the AR and ARTV models. This characteristic, combined with the worse performance on the higher percentiles of the distribution, does confirm the asymmetric performance of the complex models, whose forecasts are quite good for some series and quite bad for the other ones. On the contrary, AR and ARTV models have a rather stable forecasting ability.

5. Forecast pooling

It often happens that a combination of the forecasts from different models performs better than each single forecast, see e.g. Bates and Granger (1969) and Granger and Newbold (1986) for pooling procedures, and Clements and Hendry (2001) for possible explanations for such an outcome. In this section we evaluate whether this is the case for the aggregate variables in the ECB dataset, when using the models in the previous subsections. First we define the pooling procedures we implement, next we discuss results.

5.1 Pooling procedures

Linear combination forecasts (C). These forecasts are obtained by weighted averages of the single forecasts:

$$\hat{y}_{t+h} = \sum_{m=1}^{M} k_{m,h,t} \hat{y}_{t+h,m}, \quad k_{m,h,t} = (1/msfe_{m,h,t})^{w} / \sum_{j=1}^{M} (1/msfe_{j,h,t})^{w}, \qquad (11)$$

where m indexes the models, $k_{m,h,t}$ denotes the weights, and msfe indicates the mean square forecast error. Bates and Granger (1969) showed that the weighting scheme that minimizes the msfe of the pooled forecasts involves the covariance matrix of all the forecast errors, which is unfeasible in our case because M is very large. Hence, following their suggestion, the weight of a model is simply chosen to be inversely proportional to its msfe, which is equivalent to setting w=1 in equation (8). We also consider the cases w=0, equal weight for each forecast, and w=5, more weight for the best performing models. Moreover, we analyze separately pooling the linear models only, the non-linear models only, and all the models. Thus, overall we have 9 linear combination forecasts.

Median combination forecasts (M). These are the median forecasts from a set of models, and are computed because with non-Gaussian forecast errors linear combinations of the forecasts are no longer necessarily optimal. As in the previous method, we distinguish among three groups of models: linear, non-linear, and all models. Thus, we have 3 median combination forecasts.

Predictive least squares combination forecasts (PLS). In this approach the model is selected on the basis of its past forecasting performance over a certain period, three years in our case. Thus, the model that produced the lowest msfe over the past three years is used as the forecasting model, and the choice is recursively updated each quarter over the forecast period. We compute 4 of these forecasts, differing for the set of models compared: all models, all linear models, all non-linear models, all models plus the linear and the median combination forecasts.

The three pooling procedures and the resulting 16 forecasts are summarized in the last panel of Table 1.

5.2 Results

We start by looking at the average performance of the methods, ranking them using the loss function in equation (9). From Table 5, the best pooling method ranks only third for h=1,2, and fourth for h=4. It is a linear combination with equal weights of either all models or the linear models only. Median forecasts are ranked ninth or higher, while predictive least squares does not enter in the top-10 forecasting models.

The picture above emerges also from the more disaggregate analysis in Tables 6a and 6b. Pooling yields the lowest msfe for only 3 variables when h=1, and 1 variable for h=4. Yet, as we will se in more details below, the 3 variables are quite important: gdp, personal consumption and imports. For these variables linear forecasts were the best before. From Table 6b, the only pooling method that yields good forecasts for at least 50% of the series is PA when h=1.

The percentiles of the distribution of the relative msfe in Table 7 yield additional information on the characteristics of the pooling procedures. Focusing in particular on combination and median forecasts, the distribution is more concentrated than for linear and non-linear models. Hence, their performance is more stable for all the variables, but they are beaten by different models for different variables.

Overall, there appears to be limited scope for the adoption of pooling procedures, except for a few variables when h=1. In these cases pooling substitutes linear models as the best forecasting device, while non-linear models are in general not outperformed.

6. A disaggregate analysis

In this section we evaluate the relative merits of the 74 forecasts for each of the EMU variables in the ECB dataset. Next we consider whether stability tests could have detected the problems with constant parameter linear models that emerge from the forecast comparison. Finally, we address these two issues also for the EMU variables in the BDH and in the MSW datasets.

6.1 Variable by variable forecasting analysis

In Table 8 we report the best forecasting model for each of the 15 variables under analysis, using the loss function in equation (8), for different values of ρ and h. Though the results vary substantially with the values of these parameters and for different variables, they can be summarized in six points, focusing on mae and msfe, i.e, on $\rho = 1$ and $\rho = 2$.

First, for real gdp (YER) a combination method is preferred for h=1,4 when using the msfe and for h=2 when using the mae criterion for comparison. Otherwise, a STAR and an ANN model are the best for, respectively, h=1 and h=4 with mae, and the ARFC1b for h=2 and msfe. Moreover, linear specifications do not appear in the top-5 models, except for h=2 and msfe. Overall, there is a clear indication that a linear model does not represent a good forecasting tool for real gdp.

Second, linear models do not perform well also for the components of gdp, namely, personal and government consumption (PCR and GCR), investment and inventories (ITR and SCR), and exports (XTR), with the exception of imports (MTR). AR models are the best only in 5 out of 18 cases when using the msfe, and in 4 cases for mae. They also do not appear often in the top-5 models for these variables. Among the other forecasting methods, ANN works rather well according to the mae, models within this method are the best in 9 out of 18 cases. Using the msfe there is no clear cut ranking of methods.

Third, in general, the best models for prices (consumer prices (HICP) and gdp deflator (YED)) belong to the ARTV method. In this case linear models rank fourth or fifth.

Fourth, for the real exchange rate (EER), ANN models in general work well.

Fifth, for the short-term interest rate (STN) the best model is instead linear, and the result is robust to the choice of ρ and h. A similar result holds for the long-term rate (LTN) when h=4, while ARTV are preferred for h=1.

Finally, the linear model performs rather well also for the unit labor cost (ULC), unemployment (URX), and the trade balance (TBR), though non-linear specifications are the best for certain values of ρ and h.

6.2 Variable by variable stability analysis

We now consider whether it is possible to have an indication of the problematic performance of linear models for aggregate EMU series using standard in-sample tests for parameter stability. We can anticipate that the answer is rather negative, which supports the adoption of our forecasting approach for model evaluation.

Following Stock and Watson (1996), who present a detailed analysis of instability for US macroeconomic variables, we consider three different types of statistics.

First, tests for constant versus randomly time-varying coefficients. This set includes Nyblom's (1989, NY) locally most powerful test against the alternative of random walk coefficients (λ=0 versus λ >0 in equation (4)), and a Breusch and Pagan (1979, BP) Lagrange multiplier test against the alternative of iid random coefficients with constant mean and variance.

Second, tests based on functions of the cumulative sum of OLS residuals from equation (1), see Ploeberger and Kramer (1992). We consider the supremum of the cumulative sum (KP1), and its mean square (KP2).

Third, F-tests for constancy of the parameters against the alternative of a single break at an unknown date. The tests are computed recursively for a range of dates, say $[t_0,\dots,t_1]$, where t_0 and t_1 are selected in order to discard the first and last 15% of the sample. Three functions of the resulting sequence of statistics are considered: the supremum (Quandt (1960, QLR)); the mean (Hansen (1992), Andrews and Ploeberger (1994), MLR); and the so called average exponential (Andrews and Ploeberger (1994), ALR).

Stationarity transformations, i.e. logarithms and differencing, are applied to all series when needed (a detailed list is available upon request), and all series are represented as an AR process in levels, with 4 lags and a constant.

The outcome of the stability tests is summarized in Table 9. According to the figures, a stable model is accepted for 8 variables: YER, PCR, ITR, MTR, YED, URX, STN, and TBR. Yet, from the previous subsection, only for STN, and partly for URX and TBR, a linear model forecasts well. It is also worth noting that for the other 7 variables more than one test rejects the null hypothesis of stability, and that the KP tests reject less often than the F-test based statistics.

6.3 Other datasets

As mentioned in Section 2, the BDH and MSW datasets are monthly. The former includes the gdp (YERBDH), the gdp deflator, and M3 (M#BDH), and the forecasting period is 1991:1-1999:2. The latter includes the consumer price index (CPIMSW) and industrial production (IPMSW), and the forecasting period is 1993:1-1997:8. In both cases the forecasting horizons we consider are h=1,3,6 months.

Table 10a reports the best forecasting models for the BDH variables. For YERBDH there is a clear preference for ANN models, whose performance is also rather good for YEDBDH, while the results for M3BDH are more mixed. The latter is the only variable for which stability of a linear specification is rejected, see Table 10b.

Linear models do not perform well also for CPIMSW and IPMSW. For these variables ANN or pooled forecasts are in general the best, see Table 11a. From Table 11b, stability is rejected for both variables.

Hence, though based on a more limited number of variables, these results also suggest going beyond linear specifications for forecasting EMU variables.

7. Conclusions

In this paper we have provided a thorough analysis of the relative merits of linear, time-varying, non-linear and pooled forecasts for aggregate EMU variables. The main result is that for several variables forecasts from linear models can be substantially improved upon, even though linear specifications perform well on average.

This finding is interesting for policy purposes, but also more generally for empirical macroeconomic analysis. For example, it suggests that measures of persistence based on linear specifications can be inappropriate, as well as impulse response functions. Also, the use of GMM for the estimation of EMU forward looking Taylor rules is questionable, since the relationship between future values of inflation and the instruments can be non-linear or time-varying.

The main limitation of the current analysis is that it lacks a theoretical economic explanation for the results. The many changes that affected the economies now in the EMU can explain the failure of the linear model, but they do not provide a clear indication on the pattern of time variation of the parameter or the type of non-linearity to be included in the statistical models for the EMU variables. Further research in this area is required.

References

Andrews, D.W.K. and Ploeberger, W. (1994), "Optimal tests when a nuisance parameter is present only under the alternative", *Econometrica*, 62, 1383-1414.

Artis, M. and Marcellino, M. (2001), "Fiscal forecasting: the track record of IMF, OECD and EC", *Econometrics Journal*, 4, s20-s36.

Bates, J.M. and Granger, C.W.J. (1969), "The combination of forecasts", *Operations Research Quarterly*, 20, 415-468.

Beyer, A., Doornik, J.A., and Hendry, D.F. (2001), "Constructing historical Euro-zone data", *Economic Journal*, 111, F102-21.

Box, G.E.P. and Jenkins, G.M. (1970), *Time series analysis, forecasting and control*, San Francisco: Holden Day.

Breusch, T.S. and Pagan, A.R. (1979), "A simple test for heteroschedasticity and random coefficient variation", *Econometrica*, 47, 1287-1294.

Clarida, R., Gali, J., and Gertler, M. (1998), "Monetary policy rules in practice: Some international evidence", *European Economic Review*, 42, 1033-67.

Clements, M.P. and Hendry, D.F. (1996), "Multi-step estimation for forecasting", *Oxford Bulletin of Economics and Statistics*, 58, 657-684.

Clements, M.P. and Hendry, D.F. (2001), "Pooling of forecasts", *Econometrics Journal*, (forthcoming).

Clements, B., Kontolemis, Z., and Levy, J. (2001), "Monetary policy under EMU: Differences in the transmission mechanism?", IMF WP 102.

Diebold, F.X. and Kilian, L. (2000), "Unit Root Tests are Useful for Selecting Forecasting Models," *Journal of Business and Economic Statistic*s, 18, 265-273.

Dornbusch, R., Favero, C.A., and Giavazzi, F. (1998), "The immediate challenges for the European Central Bank", *Economic Policy*, 26, 16-52.

Elliott, G., Rothenberg, T.J. and Stock, J.H. (1996), "Efficient tests for an autoregressive unit root", *Econometrica*, 64, 813-36.

Fagan, G. and Henry, J. (1998), "Long-run money demand in the EU: Evidence for area-wide aggregates", *Empirical Economics*, 23, 483-506.

Fagan, G., Henry, J. and Mestre, R. (2001), "An area-wide model for the Euro area", ECB WP 42.

Favero, C.A. and Marcellino, M. (2001), "Large datasets, small models and monetary policy in Europe", CEPR WP 3098.

Gerlach, S and Schnabel, G (2000), "The Taylor rule and interest rates in the EMU area", *Economics Letters*, 67, 165-71.

Granger, C.W.J. and Newbold, P. (1986), *Forecasting economic time series*, San Diego: Academic Press.

Granger, C.W.J. and Terasvirta, T. (1993), *Modelling non-linear economic relationships*, Oxford: Oxford University Press.

Hansen, B. (1992), “Tests for parameter instability in regressions with I(1) processes, *Journal of Business and Economic Statistic*s, 10, 321-336.

Hornik, K., Stinchcombe, M. and White, H. (1989), “Multilayer feedforward networks are universal approximators”, *Neural Networks*, 2, 359-66.

Marcellino, M. (2002a) “Instability and non-linearity in the EMU”, CEPR WP 3312.

Marcellino, M. (2002b) “Forecast pooling for short time series of macroeconomic variables”, CEPR WP 3313.

Marcellino, M., Stock, J.H. and Watson, M.W. (2001), “Macroeconomic forecasting in the Euro area: country specific versus Euro wide information”, *European Economic Review* (forthcoming).

Makridakis, S. Anderson, A., Carbonne, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. (1982), “The accuracy of extrapolation (time series) methods: Results of a forecasting competition”, *Journal of Forecasting*, 1, 111-153.

Meese, R. and Geweke, J. (1984), A comparison of autoregressive univariate forecasting procedures for macroeconomic time series”, *Journal of Business and Economic Statistics*, 2, 191-200.

Nyblom, J. (1989), “Testing for constancy of parameters over time”, *Journal of the American Statistical Association*, 84, 223-230.

Ploeberger, W. and Kramer, W. (1992), “The CUSUM test with OLS residuals”, *Econometrica*, 60, 271-286.

Quandt, R.E. (1960), “Tests of the hypothesis that a linear regression system obeys two separate regimes”, *Journal of the American Statistical Association*, 55, 324-330.

Swanson, N.R. and White, H. (1997), “A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks”, Review of Economics and Statistics, 79, 540-550.

Terasvirta, T. (1998), “Modelling economic relationships with smooth transition regressions” in Ullah, A. and Giles, D.E.A. (eds.), *Handbook of Applied Economic Statistics*, New York: Marcel Dekker, 507-552.

Tiao, G.C. and Xu, D. (1993), “Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case”, *Biometrika*, 80, 623-641.

Stock, J.H. (1996), “VAR, error correction and pretest forecasts at long horizons”, *Oxford Bulletin of Economics and Statistics*, 58, 685-701.

Stock, J.H. and Watson, M.W. (1996), “Evidence on structural instability in macroeconomic time series relations”, *Journal of Business and Economic Statistic*s, 14, 11-30.

Stock, J.H. and Watson, M.W. (1999), “A comparison of linear and non-linear univariate models for forecasting macroeconomic time series”, in Engle, R. and White, R. (eds), *Cointegration, causality, and forecasting: A festschrift in honor of Clive W.J. Granger*, Oxford: Oxford University Press, 1-44.

Table 1 – Forecasting models

A. Linear methods

ARF(X,Y,Z)	*Autoregressive models* (18 models)
	X = C (const.) or T (trend)
	Y = 0 (stationary), 1 (I(1)), P (pre-test)
	Z = 4 (4 lags), a (AIC), b (BIC)
EX(X)	*Exponential smoothing* (3 models)
	X = 1 (single), 2 (double), P (pre-test)

B. Non-linear methods

ARTVF(X,Y,Z)	*Time-varying AR models* (9 models)
	X = C (const.)
	Y = 0 (stationary), 1 (I(1)), P (pre-test)
	Z = 3 (3 lags), a (AIC), b (BIC)
LS(X,Y,Z)	*Logistic smooth transition* (6 models)
	X = 0 (stationary), 1 (I(1)), P (pre-test)
	Y = transition variable, 10 ($\zeta_t = y_t$), 06 ($\zeta_t = y_t - y_{t-6}$)
	Z = 3 (p, lag length)
LSF(X,W)	*Logistic smooth transition* (6 models)
	X = 0 (stationary), 1 (I(1)), P (pre-test)
	W = a (AIC on transition variable and p), b (BIC)
AN(X,Y,Z,W)	*Artificial neural network models* (9 models)
	X = 0 (stationary), 1 (I(1)), P (pre-test)
	Y = 2 (n_1)
	Z = 0, 1, 2 (n_2)
	W = 3 (p, lag length)
ANF(X,S)	*Artificial neural network models* (6 models)
	X = 0 (stationary), 1 (I(1)), P (pre-test)
	S = a (AIC on n_1, n_2, p), b (BIC)

C. No Change

NOCHANGE	*No change forecast* (1 model)

D. Pooling

C(X,Y)	*Linear combination* (9 forecasts)
	X = 1 (combine A,B,C), 2 (A only), 3 (B only)
	Y = 0, 1, 5 (weight, w in equation (8))
M(X)	*Median combination* (3 forecasts)
	X = 1 (combine A,B,C), 2 (A only), 3 (B only)
P(X)	*Predictive least square combination* (4 forecasts)
	X = 1 (combine A,B,C), 2 (A only), 3 (B only), A (A,B,C,D)

Table 2 - Ranking of competing models with different loss functions

Rank/ρ		1	1.5	2	2.5	3
1	h=1	ARFT1b	ARFT1b	ARFT1b	ARFT1b	ARFT1b
	h=2	ARFT1b	ARFT1b	ARFT1b	ARFT1b	ARFT1b
	h=4	ARFT1a	ARFT1a	ARFT1a	ARFT1a	ARFT1a
2	h=1	ARFT14	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03
	h=2	ARTVFC0b	ARTVFC0b	ARTVFC0b	ARTVFC0b	ARTVFC0b
	h=4	ARFT14	ARFT14	ARFT14	ARFT14	ARFT14
3	h=1	ARTVFC03	ARFT14	ARFT1a	ARFT1a	ARFT1a
	h=2	ARTVFC0a	ARTVFC0a	ARTVFC03	ARTVFC03	ARTVFC03
	h=4	ARFT1b	ARFT1b	ARFT1b	ARFT1b	ARFT1b
4	h=1	ARFT1a	ARFT1a	ARFT14	ARFT14	ARFT14
	h=2	ARTVFC03	ARTVFC03	ARTVFC0a	ARTVFC0a	ARTVFC0a
	h=4	ARTVFC0b	ARTVFC0b	ARTVFC0b	ARTVFC0b	ARFC0b
5	h=1	ARFTP4	ARFTP4	ARFTPb	ARFTPb	ARTVFCP3
	h=2	ARFT1a	ARFT1a	ARFT1a	ARFT1a	ARFT1a
	h=4	ARFC0a	ARTVFC0a	ARTVFC0a	ARFC0b	ARTVFC0b
6	h=1	ARFTPa	ARFT0b	ARFT0b	ARFT0b	ARFTPb
	h=2	ARFT14	ARFT14	ARFT14	ARFT14	ARFC0b
	h=4	ARFC0b	ARFC0b	ARFC0b	ARFC0a	ARFC0a
7	h=1	ARFC14	ARFTPb	ARTVFC0b	ARTVFC0b	ARTVFC0b
	h=2	ARFTPb	ARFT0b	ARFC0b	ARFC0b	ARFC0a
	h=4	ARFC04	ARFC0a	ARFC0a	ARTVFC0a	ARFC04
8	h=1	ARFT0b	ARFTPa	ARFTPa	ARTVFCP3	ARFT0b
	h=2	ARFT0b	ARFC0b	ARFC0a	ARFC0a	ARFT14
	h=4	ARTVFC0a	ARFC04	ARFC04	ARFC04	ARTVFC0a
9	h=1	ARFTPb	ARFC14	ARFTP4	ARTVFC0a	ARTVFC0a
	h=2	ANF1b	ARFTPb	ARFT0b	ARFC04	ARFC04
	h=4	ARTVFC1b	ARTVFC1a	ARTVFCPa	ARTVFCPa	ARTVFCPb
10	h=1	ARFCP4	ARTVFC0b	ARTVFC0a	ARFTPa	ARFC0b
	h=2	ARTVFCPb	ARFTPa	ARFC04	ARFT0b	ARFT0b
	h=4	ARTVFC1a	ARTVFC1b	ARTVFCPb	ARTVFCPb	ARTVFCPa

Notes:
See Table1 for definition of models
The loss function is $Loss^h{}_m = \frac{1}{N}\sum_{n=1}^{N}\frac{Loss^h{}_{n,m}}{Loss^h{}_{n,1}}$, $Loss^h{}_{n,m} = \frac{1}{T-h}\sum_{t=1}^{T-h}\left|e_{t+h,n,m}\right|^{\rho}$, where the benchmark model is ARFC04 and e_{t+h} is the h-step ahead forecast error

Table 3a – Number of series for which a forecasting method has lowest msfe

Method	AR	ES	NoChange	ARTV	LSTAR	ANN
h=1	5	0	0	4	3	3
h=2	6	0	1	2	2	4
h=4	6	0	1	2	0	6

Table 3b – Number of series for which a forecasting model is in the top-N

Method		N = 1	N=5	N=10	N=15	N=20
ARFT1b	h=1	3	5	7	10	11
	h=2	1	3	5	6	8
	h=4	2	6	10	10	11
EX1	h=1	0	0	1	1	2
	h=2	0	1	2	3	3
	h=4	0	3	3	3	3
NOCHANGE	h=1	0	0	1	1	2
	h=2	1	1	2	3	3
	h=4	1	3	3	3	3
ARTVFC03	h=1	2	4	6	7	10
	h=2	1	4	4	5	8
	h=4	1	3	5	6	8
LSP063	h=1	0	1	1	1	6
	h=2	1	2	2	2	2
	h=4	0	1	2	3	4
AN0203	h=1	1	2	2	3	4
	h=2	1	3	3	4	4
	h=4	1	1	2	3	4

Notes: See Table 1 for definition of models
The figures report the number of series for which a model is among the N models with the lowest msfe.
The reported models are the best performers in each class for N=1.

Table 4 – Mean and percentiles of relative msfe for selected forecasting models

Forecast		Mean	0.02	0.10	0.25	0.50	0.75	0.90	0.98
ARFC04	h=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	h=2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	h=4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARFT1a	h=1	0.92	0.45	0.77	0.86	0.92	0.97	1.15	1.31
	h=2	0.91	0.31	0.71	0.77	0.86	1.01	1.16	1.77
	h=4	0.77	0.06	0.26	0.33	0.78	0.93	1.24	2.59
ARFT1b	h=1	0.87	0.45	0.74	0.77	0.86	0.94	0.99	1.31
	h=2	0.89	0.31	0.67	0.77	0.86	0.97	1.15	1.59
	h=4	0.81	0.08	0.26	0.33	0.80	1.00	1.24	2.59
EX1	h=1	2.52	0.80	0.86	0.91	1.28	2.25	4.72	14.38
	h=2	3.00	0.35	0.42	1.03	1.48	4.03	9.20	11.52
	h=4	3.80	0.27	0.53	0.67	1.49	6.28	12.72	14.86
NOCHANGE	h=1	2.52	0.80	0.86	0.91	1.28	2.25	4.71	14.38
	h=2	3.00	0.35	0.42	1.03	1.46	4.03	9.20	11.52
	h=4	3.84	0.27	0.53	0.67	1.54	6.28	12.72	14.86
ARTVFC03	h=1	0.89	0.69	0.73	0.77	0.85	0.95	1.11	1.25
	h=2	0.90	0.56	0.60	0.67	0.85	1.05	1.15	1.73
	h=4	1.05	0.24	0.31	0.68	0.99	1.08	2.22	2.75
LSP063	h=1	1.06	0.65	0.81	0.86	0.99	1.21	1.42	1.95
	h=2	1.19	0.59	0.64	0.79	1.02	1.55	2.02	2.62
	h=4	1.45	0.45	0.60	0.85	1.10	1.89	2.83	3.23
AN0203	h=1	1.88	0.72	0.90	0.98	1.32	2.88	4.04	4.20
	h=2	2.41	0.72	0.90	0.95	1.46	3.43	5.25	9.52
	h=4	2.16	0.58	0.61	1.08	1.26	2.84	5.64	6.46

Notes:
The models are the best in Table 2 and those in Table 3b
The benchmark model is ARFC04
See Table 1 for the definition of the models

Table 5 - Ranking of competing models with different loss functions, including pooling

Rank / ρ		1	1.5	2	2.5	3
1	h=1	ARFT1b	ARFT1b	ARFT1b	ARFT1b	ARFT1b
	h=2	ARFT1b	ARFT1b	ARFT1b	C10	C10
	h=4	ARFT1a	ARFT1a	ARFT1a	ARFT1a	ARFT1a
2	h=1	ARFT14	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03
	h=2	ARTVFC0b	ARTVFC0b	ARTVFC0b	ARFT1b	C30
	h=4	ARFT14	ARFT14	ARFT14	ARFT14	ARFT14
3	h=1	ARTVFC03	ARFT14	C20	C20	C20
	h=2	ARTVFC0a	ARTVFC0a	C10	ARTVFC0b	ARFT1b
	h=4	ARFT1b	ARFT1b	ARFT1b	ARFT1b	ARFT1b
4	h=1	ARFT1a	ARFT1a	ARFT1a	ARFT1a	ARFT1a
	h=2	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03	C11
	h=4	C10	C10	C10	C10	C10
5	h=1	C20	C20	ARFT14	C21	C21
	h=2	ARFT1a	ARFT1a	ARTVFC0a	ARTVFC0a	C31
	h=4	C20	C20	ARTVFC0b	C20	C20
6	h=1	C21	C21	C21	ARFT14	C10
	h=2	ARFT14	C10	ARFT1a	C30	ARTVFC0b
	h=4	C30	ARTVFC0b	C20	ARTVFC0b	ARFC0b
7	h=1	ARFTP4	C10	C10	C10	m1
	h=2	C10	ARFT14	C11	C11	ARTVFC03
	h=4	ARTVFC0b	C30	ARTVFC0a	ARFC0b	ARTVFC0b
8	h=1	C10	C11	C11	C11	C11
	h=2	C11	C20	C20	C31	m3
	h=4	ARFC0a	ARTVFC0a	ARFC0b	ARFC0a	ARFC0a
9	h=1	ARFTPa	m2	m1	m1	ARFT14
	h=2	C31	C11	C30	ARFT1a	ARTVFC0a
	h=4	ARFC0b	ARFC0b	ARFC0a	ARTVFC0a	ARFC04
10	h=1	m2	m1	m2	m2	m2
	h=2	C30	C30	C31	C20	C20
	h=4	ARFC04	ARFC0a	C30	ARFC04	ARTVFC0a

Notes:
See Table1 for definition of models
The loss function is $Loss^h{}_m = \frac{1}{N}\sum_{n=1}^{N}\frac{Loss^h{}_{n,m}}{Loss^h{}_{n,1}}$, $\quad Loss^h{}_{n,m} = \frac{1}{T-h}\sum_{t=1}^{T-h}\left|e_{t+h,n,m}\right|^{\rho}$, where the benchmark model is ARFC04 and e_{t+h} is the h-step ahead forecast error

Table 6a – Number of series for which a forecasting method has lowest msfe

Method	AR	ES	NoChange	ARTV	LSTAR	ANN	C	M	P
H=1	2	0	0	4	3	3	1	0	2
H=2	6	0	1	2	2	4	0	0	0
H=4	6	0	1	2	0	5	0	0	1

Table 6b – Number of series for which a forecasting model is in the top-N

Method		N = 1	N=5	N=10	N=15	N=20
ARFT1a	h=1	0	2	6	8	8
	h=2	0	1	3	8	9
	h=4	3	6	9	9	11
ARFT1b	h=1	2	4	7	9	10
	h=2	1	3	5	6	7
	h=4	0	7	9	10	11
EX1	h=1	0	0	1	1	2
	h=2	0	1	2	3	3
	h=4	0	3	3	3	3
NOCHANGE	h=1	0	0	1	1	2
	h=2	1	1	2	3	3
	h=4	1	2	3	3	3
ARTVFC03	h=1	2	4	6	6	7
	h=2	1	3	4	5	8
	h=4	1	3	5	6	6
LSP063	h=1	1	1	1	1	5
	h=2	1	2	2	2	2
	h=4	0	1	1	3	3
AN0203	h=1	1	2	2	3	3
	h=2	1	3	3	4	4
	h=4	1	1	1	1	3
C31	h=1	0	1	1	1	2
	h=2	0	1	2	5	5
	h=4	1	1	2	2	2
m3	h=1	0	0	1	1	3
	h=2	0	1	2	4	6
	h=4	0	0	0	1	2
PA	h=1	1	3	7	9	10
	h=2	0	1	2	2	3
	h=4	0	2	2	3	5

Notes: See Table 1 for definition of models
The figures report the number of series for which a model is among the N models with the lowest msfe.
The reported models are the best performers in each class for N=1.

Table 7 – Mean and percentiles of relative msfe for selected forecasting models

Forecast		Mean	0.02	0.10	0.25	0.50	0.75	0.90	0.98
ARFC04	h=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	h=2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	h=4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARFT1a	h=1	0.92	0.45	0.77	0.86	0.92	0.97	1.15	1.31
	h=2	0.91	0.31	0.71	0.77	0.86	1.01	1.16	1.77
	h=4	0.77	0.06	0.26	0.33	0.78	0.93	1.24	2.59
ARFT1b	h=1	0.87	0.45	0.74	0.77	0.86	0.94	0.99	1.31
	h=2	0.89	0.31	0.67	0.77	0.86	0.97	1.15	1.59
	h=4	0.81	0.08	0.26	0.33	0.80	1.00	1.24	2.59
EX1	h=1	2.52	0.80	0.86	0.91	1.28	2.25	4.72	14.38
	h=2	3.00	0.35	0.42	1.03	1.48	4.03	9.20	11.52
	h=4	3.80	0.27	0.53	0.67	1.49	6.28	12.72	14.86
NOCHANGE	h=1	2.52	0.80	0.86	0.91	1.28	2.25	4.71	14.38
	h=2	3.00	0.35	0.42	1.03	1.46	4.03	9.20	11.52
	h=4	3.84	0.27	0.53	0.67	1.54	6.28	12.72	14.86
ARTVFC03	h=1	0.89	0.69	0.73	0.77	0.85	0.95	1.11	1.25
	h=2	0.90	0.56	0.60	0.67	0.85	1.05	1.15	1.73
	h=4	1.05	0.24	0.31	0.68	0.99	1.08	2.22	2.75
LSP063	h=1	1.06	0.65	0.81	0.86	0.99	1.21	1.42	1.95
	h=2	1.19	0.59	0.64	0.79	1.02	1.55	2.02	2.62
	h=4	1.45	0.45	0.60	0.85	1.10	1.89	2.83	3.23
AN0203	h=1	1.88	0.72	0.90	0.98	1.32	2.88	4.04	4.20
	h=2	2.41	0.72	0.90	0.95	1.46	3.43	5.25	9.52
	h=4	2.16	0.58	0.61	1.08	1.26	2.84	5.64	6.46
C31	h=1	0.97	0.74	0.80	0.84	0.95	1.07	1.22	1.22
	h=2	0.92	0.62	0.66	0.77	0.86	1.03	1.27	1.31
	h=4	1.09	0.50	0.63	0.73	1.10	1.31	1.52	2.08
m3	h=1	0.96	0.76	0.80	0.86	0.91	1.11	1.12	1.31
	h=2	0.93	0.58	0.73	0.76	0.81	1.09	1.32	1.35
	h=4	1.17	0.55	0.57	0.76	1.10	1.51	1.65	2.36
PA	h=1	1.00	0.65	0.67	0.76	0.87	1.07	1.39	2.11
	h=2	1.26	0.63	0.79	0.96	1.03	1.40	1.56	3.19
	h=4	1.65	0.50	0.82	0.86	1.09	1.49	1.88	8.78

Notes:
The models are the best in Table 5 and those in Table 6b
The benchmark model is ARFC04
See Table 1 for the definition of the models

Table 8 - Best models for each variable with different loss functions

Variable/ρ		1	1.5	2	2.5	3
YER	h=1	LSP103	LSP103	PA	ARFT0a	ARFT0a
	h=2	C11	C20	ARFC1b	ARFCPb	ARFTPb
	h=4	ANFPa	ANFPa	C31	C11	C11
PCR	h=1	ANF0b	ANF0b	C30	C30	C30
	h=2	LSF1a	LSF1a	LSF1a	LS0063	LS0063
	h=4	C30	ARFC04	ARFT1a	ARFT1a	ARFT1a
GCR	h=1	NOCHANGE	AN0203	AN0203	AN0203	AN0203
	h=2	AN0213	ARFT1b	ARFT1b	ARFT1b	ARFT1b
	h=4	AN0213	AN0203	AN0203	AN0203	AN0203
ITR	h=1	ANF0b	LSP063	LSP063	ANF0b	ANF0b
	h=2	ANP213	ANP213	ANP213	ANP213	ANP213
	h=4	ARFT04	ARFT04	NOCHANGE	EX1	EX1
SCR	h=1	ANP223	LSP103	LS0103	LS0103	LS0103
	h=2	ANP213	ANP213	ARFCPb	ARFCPb	ARFCPb
	h=4	ARFCPb	ARFC0b	ARFC0b	ARFCPb	ARFC0b
MTR	h=1	ARFTP4	ARFTP4	P2	P2	P2
	h=2	ARFTP4	ARFCP4	ARFTP4	ARFTP4	ARFTP4
	h=4	ANP223	AN1223	ANP223	ANP223	ANP223
XTR	h=1	ANP203	ANP203	ANP203	AN1203	ARTVFC0a
	h=2	LSP063	LS1063	LSP063	LSP063	LSP063
	h=4	ARTVFC0a	ARTVFC0a	ARTVFC0b	ARTVFC0b	ARTVFC0b
HICP	h=1	ARTVFC1b	ARTVFC1b	ARTVFCPb	ARTVFCPa	ARTVFC1b
	h=2	ARTVFC0b	ARTVFC0b	ARTVFC0b	LSF1a	LSF1a
	h=4	ANF0a	ANF0a	AN0223	AN0223	AN0223
YED	h=1	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03
	h=2	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03
	h=4	ARFT14	ARTVFC03	ARTVFC03	ARTVFC03	ARTVFC03
ULC	h=1	ARTVFC03	ARTVFC03	ARFT1b	ARFT1b	ARFT1b
	h=2	ARTVFC03	ANF1b	ANF1b	ANFPb	ANF1b
	h=4	ARFT1a	ARTVFC03	ARFT14	ARFT14	ARTVFC03
URX	h=1	ARFT1b	ARFT1b	ARTVFC03	ARTVFC03	ARTVFC03
	h=2	EX1	NOCHANGE	NOCHANGE	NOCHANGE	NOCHANGE
	h=4	NOCHANGE	PA	ANFPb	ANF1b	PA
LTN	h=1	ARTVFCPb	P3	ARTVFC1b	ARTVFC13	ARTVFC13
	h=2	ARFT14	ARFT14	ANF1b	ANF1a	ANFPb
	h=4	ARFT14	ARFT14	ARFT14	ARFT14	ARFT14
STN	h=1	ARFTPb	ARFTPb	ARFT1b	ARFT1b	ARFT1a
	h=2	ARFT14	ARFT14	ARFT14	ARFT14	ARFT14
	h=4	ARFT1a	ARFT14	ARFT1a	ARFT1a	ARFT14
EER	h=1	m2	AN0223	AN0223	AN0223	AN0223
	h=2	ARFT0b	m3	AN0203	LS0063	LS0063
	h=4	AN0213	AN0213	AN0213	AN0213	AN0213
TBR	h=1	ARFT14	ARFC1b	LSF1b	LSF1b	LSFPb
	h=2	ARFT1b	ARFCPb	ARFTPb	ARFT1b	ARFT1b
	h=4	ARTVFC1b	ARFT1b	ARFT1a	ARFT1b	ARFT1a

Notes:
See Table1 for the definition of the models

The loss function is $L^h{}_{n,m} = \dfrac{Loss^h{}_{n,m}}{Loss^h{}_{n,1}}, \quad Loss^h{}_{n,m} = \dfrac{1}{T-h}\sum_{t=1}^{T-h}\left|e_{t+h,n,m}\right|^{\rho}$, where the benchmark model is ARFC04 and e_{t+h} is the h-step ahead forecast error

Table 9 –Stability Tests

Variable	NY		KP1		KP2		BP		QLR		MLR		ALR	
yer	0.91		0.71		0.1		5.07		10.97		5.58		3.73	
pcr	0.63		0.73		0.1		0.68		9.33		4.52		3.12	
gcr	1.11		1.06		0.48	**	4.15		14.33		8.86	**	4.93	
itr	0.48		0.62		0.07		3.59		7.36		3.21		2	
scr	1.5	**	0.5		0.04		5.99		25.65	***	12.88	***	9.79	***
mtr	0.58		0.52		0.06		1.46		5.51		3.4		1.92	
xtr	1.92	***	0.58		0.06		1.48		22.15	**	13.73	***	8.57	***
hicp	1.11		1.12		0.32		18.96	***	23.89	***	12.22	***	9.7	***
yed	0.59		0.96		0.12		1.35		14.85		4.86		4.98	
ulc	1.28	*	0.95		0.12		0.84		78.38	***	12.09	***	34.83	***
urx	0.91		0.88		0.22		5.2		11.69		6.01		3.99	
ltn	1.23		1.26	*	0.39	*	0.96		21.56	**	9.21	**	7.62	**
stn	0.44		0.95		0.17		1.15		10.48		2.74		2.35	
eer	1.29	*	0.95		0.21		0.79		39.95	***	6.94		15.88	***
tbr	0.61		0.63		0.06		1.11		12.02		3.34		3.1	

Notes:
The model is an AR4 with constant for the first differenced variables
NY is Nyblom's (1989) test
KP1 and KP2 are the Ploeberger and Kramer's (1992) supremum and mean square tests
BP is Breusch and Pagan's (1979) Lagrange multiplier test
QLR is Quandt's (1960) supremum F-test
MLR is Andrews and Ploeberger's (1994) mean F-test
ALR is Andrews and Ploeberger's (1994) average exponential F-test
*, **, and *** indicate significance at, respectively, 10%, 5% and 1% level

Table 10a - Best models for each variable with different loss functions, BDH dataset

Variable/ρ		1	1.5	2	2.5	3
yerbdh	h=1	ANFPa	ANF1a	ANFPa	ANFPa	ANFPa
	h=3	ANF0b	ANF0b	ANF0b	ANF0b	ANF0b
	h=6	ANF0a	ANF0a	ANF0a	ANF0a	ANF0a
yedbdh	h=1	LSF0a	LSF0a	LSF0a	LSF0a	LSF0a
	h=3	ANF0a	ANF0a	ANF0a	ANF0a	ANF0a
	h=6	ANF0a	ANF0a	ANF0a	ANF0a	ANF0a
m3bdh	h=1	P1	AN0223	ANF0b	ANF0b	ANF0b
	h=3	LSF0b	LSF0b	LSF0b	LSF0b	LSF0b
	h=6	P1	ANF0a	P3	ANF0a	P1

Notes:
See Table1 for the definition of the models
The loss function is $L^h{}_{n,m} = \dfrac{Loss^h{}_{n,m}}{Loss^h{}_{n,1}}$, $Loss^h{}_{n,m} = \dfrac{1}{T-h}\sum_{t=1}^{T-h} \left| e_{t+h,n,m} \right|^{\rho}$, where the benchmark model is ARFC04 and e_{t+h} is the h-step ahead forecast error

Table 10b –Stability Tests, BDH dataset

Variable	NY	KP1	KP2	BP		QLR		MLR		ALR	
yerbdh	0.74	0.65	0.09	4.36		10.44		4.6		3.37	
yedbdh	0.5	0.73	0.11	6.52		7.11		3.73		2.3	
m3bdh	0.96	0.89	0.17	8.39	*	19.5	**	9.57	**	7.2	**

Notes:
The model is an AR4 with constant for the first differenced variables
NY is Nyblom’s (1989) test
KP1 and KP2 are the Ploeberger and Kramer’s (1992) supremum and mean square tests
BP is Breusch and Pagan’s (1979) Lagrange multiplier test
QLR is Quandt’s (1960) supremum F-test
MLR is Andrews and Ploeberger’s (1994) mean F-test
ALR is Andrews and Ploeberger’s (1994) average exponential F-test
*, **, and *** indicate significance at, respectively, 10%, 5% and 1% level

Table 11a - Best models for each variable with different loss functions, MSW dataset

Variable/ρ		1	1.5	2	2.5	3
cpimsw	h=1	P3	PA	P3	AN0223	AN0223
	h=3	AN0203	ANF0a	ANF0a	ANF0a	ANF0a
	h=6	AN0223	AN0223	AN0223	AN0223	AN0223
ipmsw	h=1	EXP	ANP223	LSFPa	LSF1a	LSP103
	h=3	C20	C20	C20	ARTVFC1b	ARTVFCPb
	h=6	C10	C10	C20	ARFT0b	ARFT04

Notes:
See Table1 for the definition of the models
The loss function is $L^h{}_{n,m} = \dfrac{Loss^h{}_{n,m}}{Loss^h{}_{n,1}}, \quad Loss^h{}_{n,m} = \dfrac{1}{T-h}\sum_{t=1}^{T-h}\left(e_{t+h,n,m}\right)^{\rho}$, where the benchmark model is ARFC04 and e_{t+h} is the h-step ahead forecast error

Table 11b –Stability Tests, MSW dataset

Variable	NY		KP1	KP2	BP	QLR		MLR		ALR	
cpimsw	0.98		0.82	0.16	7.44	14.86		8.05	*	4.93	
ipmsw	1.33	*	0.8	0.14	4.65	17.25	*	8.16	*	6.31	**

Notes:
The model is an AR4 with constant for the first differenced variables
NY is Nyblom's (1989) test
KP1 and KP2 are the Ploeberger and Kramer's (1992) supremum and mean square tests
BP is Breusch and Pagan's (1979) Lagrange multiplier test
QLR is Quandt's (1960) supremum F-test
MLR is Andrews and Ploeberger's (1994) mean F-test
ALR is Andrews and Ploeberger's (1994) average exponential F-test
*, **, and *** indicate significance at, respectively, 10%, 5% and 1% level