Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Forecasting 21 (2005) 755–774
international journal of forecasting
www.elsevier.com/locate/ijforecast

Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination

Timo Teräsvirta[a,*], Dick van Dijk[b], Marcelo C. Medeiros[c]

[a]*Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden*
[b]*Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands*
[c]*Department of Economics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225-Gávea, 22453-900 Rio de Janeiro, RJ, Brazil*

Abstract

In this paper, we examine the forecast accuracy of linear autoregressive, smooth transition autoregressive (STAR), and neural network (NN) time series models for 47 monthly macroeconomic variables of the G7 economies. Unlike previous studies that typically consider multiple but fixed model specifications, we use a single but dynamic specification for each model class. The point forecast results indicate that the STAR model generally outperforms linear autoregressive models. It also improves upon several fixed STAR models, demonstrating that careful specification of nonlinear time series models is of crucial importance. The results for neural network models are mixed in the sense that at long forecast horizons, an NN model obtained using Bayesian regularization produces more accurate forecasts than a corresponding model specified using the specific-to-general approach. Reasons for this outcome are discussed.

JEL classification: C22; C53
Keywords: Forecast combination; Forecast evaluation; Neural network model; Nonlinear modelling; Nonlinear forecasting

1. Introduction

In recent years, numerous forecasting competitions between linear and nonlinear models for macroeconomic time series have been held. Comparisons based on a large number of variables have been carried out, and the results on forecast accuracy have generally not been particularly favourable to nonlinear models.

In a paper with impressive depth and a wealth of results, Stock and Watson (1999), henceforth SW, addressed the following four issues, among many others. First, do nonlinear time series models produce forecasts that improve upon linear models in real-

* Corresponding author.
E-mail addresses: Timo.Terasvirta@hhs.se (T. Teräsvirta), djvandijk@few.eur.nl (D. van Dijk), mcm@econ.puc-rio.br (M.C. Medeiros)

0169-2070/$ - see front matter © 2005 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2005.04.010

time? Second, if they do, are the benefits greatest for relatively tightly parameterized models or for more nonparametric approaches? Third, if forecasts from different models are combined, does the combination forecast outperform its components? Finally, are the gains from using nonlinear models and combination forecasts over simple linear autoregressive models large enough to justify their use?[1]

In this paper, we re-examine these four issues. The reason for this, and the motivation for this paper, is the following. SW used two nonlinear models to generate their forecasts: a "tightly parameterized" model and a "more nonparametric" one. The former model was the (logistic) smooth transition autoregressive ((L)STAR) model, (see Bacon & Walts, 1971; Chan & Tong, 1986; and Teräsvirta, 1994) and the latter the autoregressive single hidden layer feedforward neural network (AR-NN) model; see Fine (1999) for a general overview of neural network models. SW applied these models to 215 monthly US macroeconomic time series. They considered three forecast horizons, 1, 6 and 12 months ahead, constructing a different model for each horizon. Furthermore, since they were interested in real-time forecasting, the models were re-estimated each time another observation was added to the information set. Repeating this procedure some 300 times for each of the series (as the (longest possible) forecasting period was January 1972 to December 1996) amounted to estimating a remarkably large number of both linear and nonlinear models.

Carrying out these computations obviously required some streamlining of procedures. Thus, SW chose to employ a large number of different specifications of STAR and AR-NN models, keeping these specifications fixed over time and only re-estimating the parameters each period. This simplification was necessary in view of the large number of time series and forecasts. But then, it can be argued that building nonlinear models requires a large amount of care. As an example, consider the STAR model. First, when the data-generating process is a linear AR model, some of the parameters of the STAR model are not identified. This results in inconsistent parameter estimates, in which case the STAR model is bound to lose any forecast comparison against an appropriate linear AR model. Hence, it is essential to first test linearity before considering a STAR model at all. Second, the transition variable of the STAR model is typically unknown and has to be determined from data. Fixing it in advance may lead to a badly specified model and, again, to forecasts inferior to those from a simple linear model.

Similar arguments can be made for the AR-NN model. The ones SW used contained a linear component, that is, they nested a linear autoregressive model. This is reasonable when NN models are fitted to macroeconomic time series because the linear component can in that case be expected to explain a large share of the variation in the series. But then, if the data-generating process is linear, the nonlinear "hidden units" of the AR-NN model are redundant, and the model will most likely lose forecast comparisons against a linear AR model. Testing linearity is therefore important in this case as well. Furthermore, if the number of hidden units in the AR-NN model is too large, in the sense that some of the units do not contribute to explaining the variation in the time series, convergence problems and implausible parameter estimates may occur. This calls for a careful modelling strategy for AR-NN models as well.

An important part of our re-examination concerns the potential benefits of careful specification of STAR as well as AR-NN models. Specifically, instead of examining the forecasting performance of multiple but fixed specifications of STAR and AR-NN models, we consider a single but dynamic specification of these nonlinear models. For this purpose, model building is carried out "manually" as follows. Linearity is tested for every series and a STAR or AR-NN model is considered only if linearity is rejected. The nonlinear models are then specified using available data-based techniques that will be described in some detail below. This would be a remarkable effort if, to approximate a real-time forecasting situation as closely as possible, it were done sequentially every time another observation is added to the in-sample period. In order to keep the computational burden manageable, the models are respecified only once every 12 months. Besides, we

[1] An advantage of this simulation approach is that forecast densities are obtained directly as a by product. These densities can in turn be used for constructing interval forecasts. It is sometimes argued that the strength of nonlinear models in macroeconomic forecasting lies in such interval and density forecasts; see for example Lundbergh and Teräsvirta (2002) and Siliverstovs and van Dijk (2003). Nevertheless, since useful methods for comparing density forecasts from different models are not as yet available, neither interval nor density forecasts are considered in this study.

shall consider fewer time series than SW did. Even with these restrictions, the human effort involved is still quite considerable. As described below, our data set consists of 47 monthly time series, and for most of these the forecasting period covers 20 years. The scarcity of resources has also forced us to ignore an important part of model building, namely detailed in-sample evaluation of estimated models before applying them to out-of-sample forecasting. This omission may have had an adverse effect on our results.

As noted before, SW defined a different model for each forecast horizon. This approach has the advantage that the complications involved in computing multiple-period ahead forecasts from nonlinear models are avoided. In the case of nonlinear models such as STAR and AR-NN models, it may, however, be difficult to find a useful model for longer horizons. Another part of our re-examination thus consists of asking what happens if we specify a nonlinear model for one-period ahead forecasts only and obtain the forecasts for longer horizons numerically, by simulation or bootstrap.[2] Can such forecasts compete with ones from a linear AR model?

Finally, following SW, we shall also consider combinations of point forecasts. The difference between our study and SW is that the number of forecasts to be combined here will be considerably smaller. This is due to the fact that we generate fewer forecasts for the same variable and time horizon than SW did. One of the most remarkable results in SW was that a combination of a large number of forecasts from nonlinear models performed much better than any individual nonlinear forecast. In this study, we only consider a small number of models and thus do not particularly focus on this issue here; see Granger and Jeon (2004) for an extensive discussion of this topic.

Another difference between our study and SW is that they applied a rolling window in estimating the parameters, whereas we use expanding windows. This means that we use all observations available at each forecast origin to estimate the parameters. An important reason for doing this is that neural network models do not work very well in small samples, and we would like to give them a decent chance to perform well. Consequently, the possibility of structural breaks in series we are going to forecast is de-emphasized in our approach.

The plan of the paper is as follows. Section 2 contains a brief review of previous relevant studies on forecasting with neural network and STAR models. These two models are presented in Section 3 with a short discussion of their specification procedures. The issues involved in forecasting with nonlinear models are discussed in Section 4, including forecast combination. The recursive procedure employed to mimic real-time forecasting is presented in Section 5. The data set is described in Section 6. The empirical results are presented and analyzed in Section 7 and, finally, conclusions can be found in Section 8.

2. Previous studies

There exists a vast literature on comparing time series forecasts from neural network and linear models; Zhang, Patuwo, and Hu (1998) provide a recent survey. Many applications are to other than macroeconomic variables, and the results are mixed. In addition to SW, recent articles that examine macroeconomic forecasting with linear and AR-NN models include Swanson and White (1995, 1997a, 1997b), Tkacz (2001), Marcellino (2005), Rech (2002) and Heravi, Osborn, and Birchenhall (2004). The approach taken by Swanson and White in their articles is quite close in spirit to ours, in the sense that their idea was to select either a linear or an AR-NN model and to choose the size of the NN model using a model selection criterion such as BIC; see Rissanen (1978) and Schwarz (1978). Rech (2002) compared forecasts obtained from several neural network models specified and estimated using different methods and algorithms.

The general conclusion from the papers cited above appears to be that, in general, there is not much to gain from using an AR-NN model instead of the simple linear autoregressive model, at least as far as point forecasts are concerned. Marcellino (2005) is to some extent an exception to this rule. The data set in his study consisted of 480 monthly macroeconomic series

[2] An advantage of this simulation approach is that forecast densities are obtained directly as a byproduct. These densities can in turn be used for constructing interval forecasts. It is sometimes argued that the strength of nonlinear models in macroeconomic forecasting lies in such interval and density forecasts; see for example Lundbergh and Teräsvirta (2002) and Siliverstovs and van Dijk (2003). Nevertheless, since useful methods for comparing density forecasts from different models are not as yet available, neither interval nor density forecasts are considered in this study.

from the 11 countries that originally formed the European Monetary Union. While a linear AR model was the best overall choice in terms of point forecast accuracy, there was a reasonably large number of time series that were predicted most accurately by AR-NN models when the criterion for comparison was the root-mean-squared forecast error.

The number of studies examining the forecasting performance of STAR models relative to linear or other nonlinear models is appreciably smaller.[3] Teräsvirta and Anderson (1992) considered forecasting the volume of industrial production with STAR models. Even here, the results were mixed when the root-mean-squared forecast error was used as a criterion for comparison with linear models. Sarantis (1999) forecast real exchange rates using linear and STAR models and found that there was not much to choose between them. STAR models did, however, produce more accurate point forecasts than Markov-switching models. Similarly, Boero and Marrocu (2002) found that STAR models did not perform better than linear AR models in forecasting nominal exchange rates, although Kilian and Taylor (2003) did find considerable improvements in forecast accuracy from using STAR models for such series, in particular for longer horizons. The results in SW did not suggest that forecasts from individual LSTAR models are more accurate than forecasts from linear models. The findings of Marcellino (2005) were similar to his findings concerning AR-NN models: a relatively large fraction of the series were most accurately forecast by LSTAR models, but for many other series this model clearly underperformed.

3. The models

In this section, we present the LSTAR and AR-NN models and the modelling techniques used in this study. Throughout, we denote by $y_{t,h}$ the h-month (percent) change between times $t–h$ and t of the macroeconomic time series of interest, and $y_{t,1} \equiv y_t$.

3.1. The smooth transition autoregressive model

3.1.1. Definition

The smooth transition autoregressive (STAR) model is defined as follows:

$$y_t = \phi' w_t + \theta' w_t G(y_{t-d,h};\gamma,c) + \varepsilon_t, \tag{1}$$

where $w_t = (1, y_{t-1}, \ldots, y_{t-p})'$ consists of an intercept and p lags of y_t, $\phi = (\phi_0, \phi_1, \ldots, \phi_p)'$ and $\theta = (\theta_0, \theta_1, \ldots, \theta_p)'$ are parameter vectors, and $\varepsilon_t \sim \mathrm{IID}(0, \sigma^2)$. Note that the model in Eq. (1) can be rewritten as

$$y_t = \{\phi + \theta G(y_{t-d,h};\gamma,c)\}' w_t + \varepsilon_t,$$

which shows that the STAR model can be interpreted as a linear model with stochastically time-varying coefficients $\phi + \theta G(y_{t-d,h}; \gamma, c)$.

In general, the transition function $G(y_{t-d,h}; \gamma, c)$ is a bounded function of $y_{t-d,h}$, continuous everywhere in the parameter space for any value of $y_{t-d,h}$. In the present study, we employ the logistic transition function, which has the general form

$$G(y_{t-d,h};\gamma,c) = \left(1 + \exp\left\{-\gamma \prod_{k=1}^{K} (y_{t-d,h} - c_k)\right\}\right)^{-1}, \quad \gamma > 0,\ c_1 \le \ldots \le c_K, \tag{2}$$

with $d \ge 1$, and where the restrictions on the slope parameter and on the location parameters $c = (c_1, \ldots, c_K)'$ are identifying restrictions. Eqs. (1) and (2) jointly define the LSTAR model; see Teräsvirta (1994) for more discussion.

The most common choices of K in the logistic transition function (2) are $K = 1$ and $K = 2$. For $K = 1$, the parameters $\phi + \theta G(y_{t-d,h}; \gamma, c)$ change monotonically from ϕ to $\phi + \theta$ as a function of $y_{t-d,h}$. When $\gamma \to 0$, the LSTAR model becomes a linear AR model with coefficient $\phi + \theta/2$, while the model becomes a two-regime TAR model with c_1 as the threshold value when $\gamma \to 1$.

For $K = 2$, the parameters in Eq. (1) change symmetrically around the mid-point $(c_1 + c_2)/2$ where this logistic function attains its minimum value. The mini-

[3] It should be mentioned though that there is a sizeable literature on the forecasting performance of the threshold autoregressive (TAR) model which, in its simplest form (a single threshold), is nested in the logistic STAR model considered here; see Clements and Krolzig (1998), Clements and Smith (1999), and Siliverstovs and van Dijk (2003), among many others.

mum lies between zero and 1/2, reaching zero when $\gamma \to \infty$ and equaling 1/2 when $c_1 = c_2$ and $\gamma < \infty$. The parameter controls the slope and c_1 and c_2 determine the locations of the transitions.

LSTAR models are capable of generating asymmetric realizations, which makes them an interesting tool for modelling macroeconomic time series, exhibiting, for example, changes in their dynamic properties over the business cycle. Building upon this idea, in our application K in Eq. (2) is set equal to 1 (the same choice is made by SW) and the transition variable $y_{t-d,h}$ is taken to be a lagged annual difference, that is, $h = 12$. As a consequence, for most series in our data set, the resulting regimes associated with the extreme values of the logistic transition function $G(y_{t-d,h}; \gamma, c)$ correspond quite closely with business cycle expansions and recessions. A lagged first difference would be too volatile a transition variable for that purpose. A similar choice was made in Skalin and Teräsvirta (2002) for modelling business cycle asymmetries in quarterly unemployment rate series.

3.1.2. Building STAR models

In building STAR models, we shall follow the modelling strategy presented in Teräsvirta (1998), see also van Dijk, Teräsvirta, and Franses (2002) and Lundbergh and Teräsvirta (2002). As already indicated, building STAR models has to be initiated by testing linearity. The LSTAR model reduces to a linear AR model when either $\theta = 0$ or $\gamma = 0$ in Eqs. (1) and (2). Testing the linearity hypothesis is not straightforward, however, due to the presence of unidentified nuisance parameters that invalidate the standard asymptotic inference. In the STAR context, it is customary to circumvent this problem by approximating the alternative model using a Taylor series expansion of the transition function, as discussed in Luukkonen, Saikkonen, and Teräsvirta (1988). Linearity is tested for a number of "candidate" transition variables $y_{t-d,h}$, $d \in D = \{1, 2, \ldots, d_{\max}\}$ where we set $d_{\max} = 6$, using a significance level of 0.05 for each individual test. The test results are at the same time used to select the delay parameter d, which is taken to be the value for which the p-value of the linearity test is smallest; see Teräsvirta (1994, 1998) for details. Because of our choice to set K equal to 1 a priori, the data-based choice between $K = 1$ and $K = 2$, which normally is part of the specification procedure, need not be made.

The lag structure of the LSTAR model can in principle be specified by starting with a "large" model and removing redundant lags, that is (sequentially) imposing zero restrictions on parameters. During this reduction process, the estimated models can be evaluated by various misspecification tests as discussed in Teräsvirta (1998) to monitor their adequacy. Preliminary experiments indicated, however, that doing so often impairs forecasts, and as a result, this reduction is not used in this paper. Hence, we restrict ourselves to "full" LSTAR models containing all lags up to a certain order p in both the linear and nonlinear parts of the model (as represented by the parameter vectors ϕ and θ, respectively), where p is determined using BIC, allowing for a maximum lag order of 12.

As a whole, the modelling strategy requires a substantial amount of human resources and consequently, as mentioned in the introduction, the STAR model is respecified only once every year. However, the parameters are re-estimated every month.

3.2. The autoregressive artificial neural network model

3.2.1. Definition

The autoregressive single hidden layer feedforward neural network (AR-NN) model used in our work has the form

$$y_t = \beta_0' w_t + \sum_{j=1}^{q} \beta_j G\left(\gamma_j' w_t\right) + \varepsilon_t, \tag{3}$$

where $w_t = (1, y_{t-1}, \ldots, y_{t-p})'$ as before, and β_j, $j = 1, \ldots, q$, are parameters, called "connection strengths" in the neural network literature. Furthermore, the function $G(\cdot)$ is a bounded function called a "hidden unit" or "squashing function" and γ_j, $j = 1, \ldots, q$, are parameter vectors. Our squashing function is the logistic function $G(z) = 1/(1 + \exp(-z))$, comparable to Eq. (2) with $K = 1$. The errors ε_t are assumed to be IID $(0, \sigma^2)$. We include the "linear unit" $\beta_0' w_t$ in Eq. (3) despite the fact that many neural network users assume $\beta_0 = 0$. A theoretical argument used to motivate the use of (AR-)NN models is that they are universal approximators. Suppose that $y_t = H(w_t)$ for some nonlinear function $H(\cdot)$, that is, there exists a functional relationship between y_t and the variables in w_t. Then, under mild regularity conditions on H, there exists a

positive integer $q \leq q_0 < \infty$ such that for arbitrary $\delta > 0$, $|H(w_t) - \sum_{j=1}^{q} \beta_j G(\gamma_j' w_t)| < \delta$ for all w_t. This is an important result because q_0 is finite, so any unknown function $H(\cdot)$ can be approximated arbitrarily accurately by a linear combination of a finite number of hidden units $G(\gamma_j' w_t)$. This universal approximator property of Eq. (3) has been discussed in several papers including Cybenko (1989), Funahashi (1989), Hornik, Stinchombe, and White (1989), andWhite (1990). In principle, Eq. (3) offers a very flexible parametrization for describing the dynamic structure of y_t.

3.2.2. Building AR-NN models using statistical inference

Building AR-NN models involves two crucial choices. First, one has to select the input variables, w_t, for the model. In the univariate case considered here, this is equivalent to selecting the relevant lags of y_t. Second, one has to choose the number of hidden units, q, to be included in the model. Broadly speaking, there exist two alternative ways of building AR-NN models. On the one hand, one may begin with a small model and gradually increase its size. This is sometimes called a "bottom-up" approach or "growing the network" and is applied, for example, in Swanson and White (1995, 1997a, 1997b). On the other hand, it is also possible to have a large model as starting-point and "prune" it, which means (sequentially) removing hidden units and variables. In this paper, we apply both approaches and shall first describe a bottom-up approach based on the use of statistical inference, originally suggested in Medeiros, Teräsvirta, and Rech (2005).

The first step of the bottom-up inference-based strategy is to select the input variables. This is done by applying another universal approximator, a general polynomial. For example, approximating the right-hand side of Eq. (3) by a third-order polynomial yields

$$y_t = \mu_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{p} \sum_{i=j}^{p} \alpha_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p} \sum_{j=i}^{p} \sum_{k=j}^{p} \alpha_{ijk} y_{t-i} y_{t-j} y_{t-k} + \varepsilon_t^*. \quad (4)$$

An appropriate model selection criterion such as BIC is used to sort out the redundant combinations of variables and thus select the relevant lags of y_t, as described in Rech, Teräsvirta, and Tschernig (2001). The automated selection technique by Krolzig and Hendry (2001) may also be used for this purpose.

The second step consists of selecting the number of hidden units. Linearity is tested first, which entails an identification problem similar to the one encountered in STAR models. This is circumvented by using Eq. (4) and the neural network linearity test of Teräsvirta, Lin, and Granger (1993). If linearity is rejected, a model with a single hidden unit ($q = 1$ in Eq. (3)) is estimated using conditional maximum likelihood. Next, this model is tested against an AR-NN model with $q \geq 2$ hidden units as described in Medeiros et al. (2005) and, if rejected, an AR-NN model with two hidden units is estimated. This procedure is continued until the first non-rejection of the null hypothesis. We favour parsimonious models and therefore follow the suggestion of Medeiros et al. (2005) to let the significance levels in the testing sequence form a decreasing sequence of positive real numbers. More specifically, the significance level is halved at each stage, while we set the significance level of the first (linearity) test equal to 0.05.

3.2.3. Building AR-NN models using Bayesian regularization

There exist many methods for pruning a network, see for example Fine (1999, Chapter 6) for an informative account. In this paper, we apply a technique called "Bayesian regularization," as described in MacKay (1992a). The aim of Bayesian regularization is twofold. First, it is intended to facilitate maximum likelihood estimation by penalizing the log-likelihood in case some of the parameter estimates become excessively large. Second, the method is used to find a parsimonious model within a possibly very large model. In order to describe the former aim in more detail, suppose that the estimation problem is "ill-posed" in the sense that the likelihood function is very flat in several directions of the parameter space. This is not uncommon in large neural network models, and it makes numerical maximization of the likelihood difficult. Besides, the maximum likelihood value may be strongly dependent on a small number of data points. An appropriate prior distribution on the parameters acts as a "regularizer" that imposes smoothness and makes estimation easier. For example, the

prior distribution may be defined such that it shrinks the parameters or some linear combinations of them towards zero. Information in the time series is used to find the "optimal" amount of shrinkage. Furthermore, a set of smaller models nested within the large original model is defined. The algorithm allows to choose one of these sub-models and thus reduce the size of the neural network. This requires determining prior probabilities for the models in the set and finding the one with the highest posterior probability.

Bayesian regularization can be applied to feedforward neural networks of type (3), as discussed in MacKay (1992b). In this context, the set of eligible AR-NN models does not usually contain models with a linear unit, and we adhere to that practice here. In our case, the largest model has nine hidden units ($q=9$ in Eq. (3)), and the maximum lag p equals six. We apply the Levenberg–Marquardt optimization algorithm in conjunction with Bayesian regularization as proposed in Foresee and Hagan (1997).

As already mentioned, in the approach based on statistical inference (discussed in Section 3.2.2) parsimony is achieved by starting from a small model and growing the network by applying successively tightening tests for remaining nonlinearity. Bayesian regularization also has parsimony as a guiding principle, but it is achieved from the opposite direction by pruning a large network. In what follows and in all tables, a neural network model obtained this way is called the NN model, whereas a neural network model built as explained in Section 3.2.2 is called the AR-NN model.

4. Forecasting with STAR and neural network models

As pointed out in the introduction, we obtain forecasts for different horizons for a given variable from the same (one-step ahead) model. This means that the forecasts from nonlinear models have to be generated numerically as discussed in Granger and Teräsvirta (1993). Let

$$y_t = f(y_{t-1},..,y_{t-p};\theta) + \varepsilon_t, \tag{5}$$

be a nonlinear model with an additive error term, where θ is the parameter vector and $\varepsilon_t \sim \mathrm{IID}(0, \sigma^2)$. The STAR model (1)–(2) and the AR-NN model (3) considered here are special cases of Eq. (5). The one-step ahead point forecast for y_{t+1} equals

$$\hat{y}_{t+1|t} = f(y_t,..,y_{t-p+1};\hat{\theta}_t),$$

where $\hat{\theta}_t$ indicates that parameter estimates are obtained using observations up to time period t. For the two-step ahead forecast of y_{t+2} one obtains

$$\hat{y}_{t+2|t} = \int_{-\infty}^{\infty} f(\hat{y}_{t+1|t} + \varepsilon_{t+1}, y_t,..,y_{t-p+2};\hat{\theta}_t)\mathrm{d}\varepsilon_{t+1}. \tag{6}$$

For longer horizons, obtaining the point forecast would even require solving a multidimensional integral. Numerical integration in Eq. (6) can be avoided either by approximating the integral by simulation or by bootstrapping the residuals. The latter alternative requires the assumption that the errors of model (5) be independent.

In this paper, we adopt the bootstrap approach (see Lundbergh and Teräsvirta (2002) for another application of this method in the context of STAR models). In particular, we simulate N paths for y_{t+1}, y_{t+2}, …, $y_{t+h_{\max}}$, where we set $N=500$ and $h_{\max}=12$, and obtain the h-period ahead point forecast as the average of these paths. For example, the two-step ahead point forecast is computed as

$$\begin{aligned}\hat{y}_{t+2|t} &= \frac{1}{N}\sum_{i=1}^{N}\hat{y}_{t+2|t}(i) \\ &= \frac{1}{N}\sum_{i=1}^{N} f\left(\hat{y}_{t+1|t} + \hat{\varepsilon}_i, y_t,..,y_{t-p+2};\hat{\theta}\right),\end{aligned}$$

where $\hat{\varepsilon}_i$ are resampled residuals from the model estimated using observations up to time period t. In addition, in this paper, we do not consider h-step ahead forecasts of the 1-month growth rate by $\hat{y}_{t+h|t}$ but focus instead on the economically more interesting forecasts of the h-month growth rate, denoted by $\hat{y}_{t+h,h|t} = \sum_{j=1}^{h}\hat{y}_{t+j|t}$.

Note that by building a separate model for each forecast horizon as in SW and Marcellino (2005), numerical forecast generation is avoided. In that case point forecasts for the h-month growth rate are of the form

$$\hat{y}_{t+h,h|t} = f_h(y_t,..,y_{t-p+1};\hat{\theta}_h), \quad h \geq 1,$$

where $f_h(y_t, \ldots, y_{t-p+1}; \hat{\theta}_h)$ is the estimated conditional mean of $y_{t+h,h}$ at time t according to the model constructed for forecasting h periods ahead.

We also consider combining forecasts from different models, where we limit ourselves to combinations of pairs of models and exclude the combination of forecasts from the NN and AR-NN models. Following Granger and Bates (1969), the composite point forecast based on models M_i and M_j is given by

$$\hat{y}^{(i,j)}_{t+h,h|t} = (1-\lambda_t)\hat{y}^{(i)}_{t+h,h|t} + \lambda_t \hat{y}^{(i)}_{t+h,h|t}, \quad (7)$$

where λ_t is the weight of the h-month forecast by $\hat{y}^{(j)}_{t+h,h|t}$ from model M_j, where the restriction $0 \le \lambda_t \le 1$ may or may not be imposed. The weights can be time-varying and based on previous forecasting performance of M_i and M_j, but they may also be fixed. In fact, a general conclusion from the literature on forecast combination, also reached by SW, appears to be that equal weighting, that is $\lambda_t \equiv 1/2$, does an adequate job in the sense that more refined weighting schemes generally do not lead to further improvements in forecast accuracy. We follow this approach here. In combination forecasts that include the AR model as one of the components, the equal weighting scheme actually favours the linear model. In both STAR and AR-NN approaches, forecasts are obtained from a linear model when linearity is not rejected. The implicit weight of the linear model in combination forecasts is thus greater than the weight indicated by λ_t.

5. Recursive specification, estimation, and forecasting

Specification, estimation, and forecasting are carried out recursively using an expanding window of observations. For most of the series considered in this paper, the first window starts in January 1960 and ends in December 1980, whereas the last window (also starting in January 1960) ends in December 1999. However, for a few series, the starting-date for the windows and the ending date for the last window are slightly different. The general rule is that all windows begin from the first observation and the last window is closed 12 months before the final observation. As already mentioned, all models are respecified only once a year, but parameters are re-estimated each month. For every window we compute point forecasts $\{\hat{y}_{t=h,h|t}\}_{t=R}^{R+P-1}$ of the h-month growth rate $y_{t+h,h}$ of all variables, where $h=1, \ldots, 12$, R is the point where the first forecast is made, and P is the number of windows. This procedure gives us P forecasts for all horizons; for most series in our data set $P=228$.

Neural network models have a tendency to overfit in the sense that the specification procedure may lead to a large number of hidden units and poor out-of-sample performance. Furthermore, as will be pointed out in Section 7.3, estimated AR-NN models may sometimes be explosive although the time series to be modelled appear stationary. For these reasons, we impose an "insanity filter" on the forecasts; see also Swanson and White (1995). If a forecast deviates more than plus/minus two standard deviations from the average of the observed h-month differences, it is replaced by the arithmetic mean of $y_{t,h}$ computed with the available observations up until t. "Insanity" is thus replaced by "ignorance." SW apply a similar technique and call it trimming the forecast.

6. Data

We consider the following monthly macroeconomic variables for each of the G7 countries: volume of industrial production (IP), consumer price index inflation (CPI), narrow money (M1), short-term (3-month) interest rate (STIR), volume of exports (VEX), volume of imports (VIMP), and unemployment rate (UR). The unemployment rates for France and Italy are excluded because sufficiently long monthly series are not available, such that the data set consists of 47 monthly time series. Most series start in January 1960 and are available up to December 2000.

The series are seasonally adjusted with the exception of the short-term interest rate and inflation. With the exception of the NN models estimated with Bayesian regularization, seasonality in these series is modelled by including monthly dummy variables, which are restricted to enter linearly in all models. In the NN model seasonality is modelled by including the first 12 lags as input variables. For all series except the interest rates and unemployment rates, models are specified for monthly growth rates y_t, obtained by first differencing the logarithm of the levels. For interest rates, plain first differences are used. For unemployment rates, models are specified for the levels of the series, which is effectively done by including a lagged level term as an additional variable

in the model for the 1-month change. Most series have been adjusted to remove the influence of outliers. Details, including the data sources and the types of adjustments made, can be obtained from the authors upon request.

7. Results

Before we turn to the empirical results of the forecasting exercise, we briefly discuss the results of the linearity tests. These are summarized in Table 1, showing the fraction of times linearity is rejected with tests performed once a year when the models are respecified.

The results for the two models (or linearity tests) are not identical, the differences being most pronounced for the IP series. There are at least three reasons for this. First, the linear models that form the null hypotheses are not the same for the STAR and AR-NN alternatives. In the STAR case, the linear models contain all lags up to the order p selected by BIC. In the AR-NN case, the variables to be included in the AR-NN model are selected first by means of the technique described in Section 3.2.2. Second, the alternative against which linearity is tested is not the same either. Finally, a "rejection" against the STAR model is the result of carrying out the test against a number of $d_{\max}$ alternatives in which the transition variable of the model is different; see Section 3.1.2.

Table 1
Linearity test results

	IP	CPI	M1	STIR	VEX	VIMP	UR
STAR							
Canada	0.95	0.25	1	1	0.05	0	0.90
France	1	1	0.77	1	0.55	0	–
Germany	0.90	1	0	1	0	0.60	0.95
Italy	0	1	0.22	1	1	1	–
Japan	0.15	1	0.85	1	0.35	0	0.95
UK	0	0.35	0.95	1	0.25	0.20	0
USA	1	1	0.80	1	0.50	1	1
AR-NN							
Canada	0	0.95	0.47	1	0.11	1	0.74
France	0	1	0.35	0.84	0	0	–
Germany	0	1	0	1	0	1	1
Italy	0	1	0.25	0.95	1	1	–
Japan	0.53	0.95	0	0.37	0.16	0.11	0.84
UK	0	1	0.58	1	0.32	1	1
USA	1	0.95	1	1	1	1	0.32

The table contains rejection frequencies of the linearity hypothesis against STAR (upper panel) and AR-NN (lower panel) models, with the respective linearity tests performed once a year when the models are respecified. A dash indicates that the series is not available.

Linearity is rejected somewhat more frequently against LSTAR than against AR-NN models. The short-term interest rate, inflation and the unemployment rate series appear to be most systematically nonlinear when linearity is tested against STAR. In the AR-NN case, inflation, interest rates, and imports are the "most nonlinear" variables. Also note that there are country/series combinations for which linearity is never rejected.

The fact that linearity is always rejected does not, however, imply stability of the STAR and AR-NN specifications over time. As an example, the AR-NN model for the German unemployment rate contains two hidden units. This number first declines to one for a period of time, then fluctuates between two and four before it drops to one again towards the end of the observation period. There are other examples, however, such as the Canadian VIMP, in which the number of hidden units remains unchanged, in this case one, over the whole period.

7.1. Comparing point forecasts using the root-mean-squared forecast error

The point forecasts are evaluated by using the root-mean-squared forecast error (RMSFE). We also computed the mean absolute forecast errors (MAFE), but because they do not seem to contain information not already available in RMSFE values, they are not reported here. Table 2 reports the ratios of the RMSFE for a given forecast horizon, $h=1$, 3, 6, and 12 months, relative to the RMSFE of the linear AR(p) model with p selected by BIC, which we use as benchmark. It also contains the ranks of the four forecasting methods.

We assess the significance of observed differences in MSFE between models by applying the pairwise Diebold–Mariano test of equal forecast accuracy, using the modified form of Harvey, Leybourne, and Newbold (1997), and the pairwise forecast encompassing test developed in Harvey, Leybourne, and Newbold (1998). It should be noted that the standard asymptotic theory for the Diebold–Mariano and fore-

Table 2
Point forecast evaluation: RMSFE ratios

		IP				CPI				M1				STIR			
		1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
Canada	AR	[0.012] (4)	[0.018] (3)	[0.029] (3)	[0.042] (3)	[0.004] (1)	[0.004] (1)	[0.004] (1)	[0.005] (2)	[0.014] (3)	[0.023] (2)	[0.030] (2)	[0.044] (1)	[0.776] (2)	[1.511] (2)	[2.144] (1)	[2.765] (1)
	NN	0.994 (3)	0.990 (2)	0.989 (2)	0.941 (2)	1.119 (4)	1.144 (3)	1.174 (3)	1.069 (3)	1.002 (4)	1.032 (4)	1.078 (3)	1.152 (3)	1.022 (2)	1.108 (4)	1.092 (4)	1.178 (4)
	AR-NN	0.983 (1)	1.034 (4)	1.088 (4)	1.236 (4)	1.001 (3)	1.167 (4)	1.332 (4)	1.251 (4)	1.000 (2)	1.008 (3)	1.081 (4)	1.166 (4)	0.930 (1)	0.964 (1)	0.995 (1)	1.049 (2)
	STAR	0.989 (2)	0.938 (1)	0.890 (1)	0.871 (1)	1.000 (1)	1.000 (1)	1.000 (1)	1.000 (2)	0.992 (1)	0.966 (1)	0.997 (1)	1.025 (2)	1.078 (4)	1.019 (3)	1.006 (2)	1.083 (3)
France	AR	[0.011] (3)	[0.014] (3)	[0.020] (3)	[0.031] (3)	[0.003] (1)	[0.003] (2)	[0.003] (2)	[0.004] (1)	[0.007] (1)	[0.013] (2)	[0.020] (2)	[0.034] (2)	[0.565] (2)	[1.092] (1)	[1.566] (1)	[1.830] (1)
	NN	0.992 (2)	0.971 (1)	0.972 (2)	0.961 (1)	1.004 (2)	0.990 (1)	0.933 (1)	1.040 (2)	1.134 (4)	1.073 (3)	1.130 (3)	1.182 (3)	1.001 (3)	1.099 (4)	1.145 (4)	1.233 (3)
	AR-NN	1.009 (4)	1.031 (4)	1.051 (4)	1.105 (4)	1.107 (4)	1.620 (4)	1.936 (4)	1.760 (4)	1.087 (3)	1.189 (4)	1.349 (4)	1.493 (4)	0.903 (1)	1.016 (3)	1.073 (3)	1.152 (2)
	STAR	0.983 (1)	0.985 (2)	0.964 (1)	0.971 (2)	1.023 (3)	1.034 (3)	1.061 (3)	1.061 (3)	1.084 (2)	0.910 (1)	0.835 (1)	0.763 (1)	1.017 (3)	1.011 (2)	1.039 (2)	1.240 (4)
Germany	AR	[0.015] (4)	[0.019] (3)	[0.027] (3)	[0.042] (4)	[0.003] (2)	[0.004] (1)	[0.004] (1)	[0.005] (1)	[0.008] (2)	[0.015] (2)	[0.022] (2)	[0.029] (1)	[0.371] (1)	[0.739] (1)	[1.174] (1)	[1.606] (1)
	NN	0.985 (3)	0.991 (2)	0.970 (2)	0.913 (2)	1.155 (4)	1.101 (3)	1.309 (3)	1.278 (4)	1.006 (4)	1.000 (4)	0.993 (1)	1.021 (3)	1.005 (2)	1.067 (3)	1.050 (2)	1.068 (2)
	AR-NN	0.970 (1)	1.007 (4)	1.007 (4)	0.998 (3)	0.995 (1)	1.128 (4)	1.337 (4)	1.110 (3)	1.001 (3)	0.998 (1)	1.022 (4)	1.146 (4)	1.035 (4)	1.083 (4)	1.073 (3)	1.097 (3)
	STAR	0.971 (2)	0.902 (1)	0.832 (1)	0.783 (1)	1.000 (2)	1.000 (1)	1.000 (1)	1.000 (1)	1.000 (2)	1.000 (2)	1.000 (2)	1.000 (1)	1.013 (3)	1.042 (2)	1.094 (4)	1.186 (4)
Italy	AR	[0.020] (2)	[0.023] (2)	[0.031] (2)	[0.047] (2)	[0.003] (2)	[0.004] (4)	[0.003] (3)	[0.004] (4)	[0.007] (2)	[0.014] (2)	[0.023] (2)	[0.039] (2)	[0.597] (1)	[1.151] (1)	[1.617] (1)	[2.171] (1)
	NN	0.995 (1)	0.992 (1)	0.962 (1)	0.893 (1)	1.119 (4)	0.958 (3)	1.075 (4)	0.890 (1)	1.065 (3)	1.109 (4)	1.130 (3)	1.193 (3)	1.092 (3)	1.270 (3)	1.350 (3)	1.435 (3)
	AR-NN	1.003 (4)	1.013 (4)	1.018 (4)	1.040 (4)	1.073 (3)	0.952 (2)	0.968 (1)	0.927 (2)	1.081 (4)	1.107 (3)	1.137 (4)	1.233 (4)	1.490 (4)	1.489 (4)	1.519 (4)	1.468 (4)
	STAR	1.000 (2)	1.000 (2)	1.000 (2)	1.000 (2)	0.956 (1)	0.919 (1)	0.995 (2)	0.986 (3)	0.986 (1)	0.974 (1)	0.927 (1)	0.887 (1)	1.070 (3)	1.121 (2)	1.207 (2)	1.220 (2)
Japan	AR	[0.015] (3)	[0.019] (2)	[0.031] (2)	[0.045] (2)	[0.005] (1)	[0.006] (2)	[0.005] (1)	[0.007] (1)	[0.009] (2)	[0.018] (2)	[0.028] (2)	[0.041] (2)	[0.355] (4)	[0.751] (2)	[1.083] (2)	[1.461] (2)
	NN	0.992 (2)	0.972 (1)	0.973 (1)	0.889 (1)	1.160 (3)	1.185 (3)	1.467 (3)	1.551 (4)	1.010 (3)	1.021 (3)	1.037 (3)	1.087 (3)	0.928 (1)	1.049 (4)	1.121 (4)	1.076 (4)
	AR-NN	0.989 (1)	1.022 (4)	1.068 (4)	1.210 (4)	1.289 (4)	1.605 (4)	1.580 (4)	1.408 (3)	1.056 (4)	1.098 (4)	1.148 (4)	1.394 (4)	0.977 (2)	1.015 (3)	1.012 (3)	1.031 (3)
	STAR	1.000 (4)	1.000 (3)	1.002 (3)	1.000 (3)	1.005 (2)	0.979 (1)	1.014 (2)	1.018 (2)	0.984 (1)	0.894 (1)	0.838 (1)	0.870 (1)	0.978 (3)	0.971 (1)	0.941 (1)	0.855 (1)
UK	AR	[0.011] (1)	[0.016] (2)	[0.024] (3)	[0.031] (2)	[0.004] (1)	[0.005] (1)	[0.004] (2)	[0.005] (2)	[0.005] (3)	[0.009] (2)	[0.015] (2)	[0.028] (2)	[0.526] (1)	[1.092] (1)	[1.577] (1)	[2.208] (1)
	NN	1.020 (4)	0.998 (1)	0.995 (1)	0.981 (1)	1.325 (4)	1.257 (3)	1.585 (3)	1.539 (3)	0.994 (2)	1.031 (3)	1.012 (3)	1.018 (3)	1.009 (2)	1.054 (4)	1.150 (3)	1.199 (4)
	AR-NN	1.006 (3)	1.002 (4)	0.997 (2)	1.012 (4)	1.315 (3)	1.997 (4)	2.282 (4)	2.165 (4)	1.075 (4)	1.396 (4)	1.284 (4)	1.227 (4)	1.022 (3)	1.000 (2)	1.014 (2)	1.015 (2)
	STAR	1.000 (1)	1.000 (2)	1.000 (3)	1.000 (2)	1.066 (2)	1.012 (2)	1.003 (1)	1.026 (1)	0.951 (1)	0.860 (1)	0.692 (1)	0.604 (1)	1.023 (4)	1.047 (3)	1.085 (3)	1.101 (3)
USA	AR	[0.007] (4)	[0.015] (3)	[0.023] (2)	[0.034] (3)	[0.003] (3)	[0.004] (2)	[0.004] (2)	[0.004] (2)	[0.006] (3)	[0.012] (3)	[0.020] (2)	[0.035] (3)	[0.705] (2)	[1.323] (4)	[1.686] (3)	[2.320] (4)
	NN	0.992 (2)	1.002 (4)	1.001 (3)	0.990 (2)	0.920 (1)	0.924 (1)	0.966 (1)	0.907 (1)	0.992 (1)	1.000 (2)	1.004 (3)	0.976 (2)	1.006 (3)	0.988 (3)	0.968 (3)	0.891 (3)
	AR-NN	0.995 (3)	0.997 (2)	1.016 (4)	1.080 (4)	0.987 (2)	1.052 (4)	1.194 (4)	1.149 (4)	1.028 (4)	1.126 (4)	1.168 (4)	1.216 (4)	0.966 (1)	1.008 (4)	0.961 (2)	0.856 (1)
	STAR	0.954 (1)	0.941 (1)	0.919 (1)	0.902 (1)	1.000 (3)	1.000 (2)	1.000 (2)	1.000 (2)	0.995 (2)	0.975 (1)	0.949 (1)	0.902 (1)	1.090 (4)	0.974 (1)	0.859 (1)	0.859 (2)

Table entries are ratios of root-mean-squared forecast errors and ranks (in parentheses) of the models by variable, country, and forecast horizon, where the linear AR model is the baseline model. For the AR model, entries in square brackets are RMSFE values.

Table 2 (*continued*)

		VEX				VIMP				UR			
		1	3	6	12	1	3	6	12	1	3	6	12
Canada	AR	[0.040] (1)	[0.048] (2)	[0.060] (2)	[0.090] (2)	[0.043] (3)	[0.051] (2)	[0.069] (2)	[0.108] (3)	[0.250] (3)	[0.488] (3)	[0.785] (3)	[1.181] (3)
	NN	1.036 (4)	1.060 (4)	1.042 (4)	1.045 (4)	0.973 (2)	0.992 (1)	0.978 (1)	0.958 (1)	0.979 (2)	0.978 (1)	0.976 (2)	0.958 (2)
	AR-NN	1.001 (2)	0.988 (1)	1.005 (3)	1.009 (3)	0.931 (1)	1.008 (4)	1.001 (4)	0.995 (2)	0.953 (1)	0.980 (2)	1.080 (4)	1.229 (4)
	STAR	1.004 (3)	0.995 (2)	0.995 (1)	0.995 (2)	1.000 (3)	1.000 (2)	1.000 (2)	1.000 (3)	1.009 (3)	0.982 (3)	0.964 (1)	0.920 (1)
France	AR	[0.040] (3)	[0.042] (2)	[0.050] (1)	[0.070] (1)	[0.033] (2)	[0.040] (2)	[0.055] (2)	[0.088] (1)	–	–	–	–
	NN	0.992 (2)	1.011 (3)	1.105 (3)	1.230 (3)	0.992 (1)	1.000 (1)	0.989 (1)	1.003 (3)	–	–	–	–
	AR-NN	1.012 (4)	1.029 (4)	1.147 (4)	1.307 (4)	1.014 (4)	1.041 (4)	1.053 (4)	1.084 (4)	–	–	–	–
	STAR	0.969 (1)	0.984 (1)	1.076 (2)	1.201 (2)	1.000 (2)	1.000 (2)	1.000 (2)	1.000 (1)	–	–	–	–
Germany	AR	[0.038] (2)	[0.043] (1)	[0.057] (1)	[0.084] (1)	[0.046] (3)	[0.048] (3)	[0.065] (3)	[0.101] (3)	[0.106] (3)	[0.257] (2)	[0.438] (2)	[0.625] (3)
	NN	1.025 (4)	1.028 (4)	1.028 (4)	1.040 (4)	1.016 (3)	0.987 (2)	0.968 (2)	0.895 (2)	0.973 (1)	0.977 (1)	0.970 (1)	0.991 (2)
	AR-NN	0.999 (1)	1.007 (3)	1.004 (3)	1.005 (3)	1.017 (4)	1.043 (4)	1.036 (4)	1.023 (4)	0.989 (2)	1.032 (4)	1.199 (4)	1.492 (4)
	STAR	1.000 (2)	1.000 (1)	1.000 (1)	1.000 (1)	0.979 (1)	0.941 (1)	0.889 (1)	0.833 (1)	1.019 (4)	1.015 (3)	1.002 (3)	0.962 (1)
Italy	AR	[0.087] (2)	[0.084] (2)	[0.097] (1)	[0.137] (1)	[0.093] (2)	[0.092] (2)	[0.111] (2)	[0.159] (2)	–	–	–	–
	NN	1.012 (3)	1.056 (4)	1.064 (4)	1.029 (3)	1.007 (3)	1.034 (3)	1.067 (4)	1.054 (4)	–	–	–	–
	AR-NN	0.999 (1)	0.997 (1)	1.006 (2)	1.016 (2)	0.993 (1)	0.995 (1)	1.004 (3)	1.011 (3)	–	–	–	–
	STAR	1.018 (4)	1.033 (3)	1.052 (3)	1.048 (4)	1.032 (4)	0.994 (1)	0.981 (1)	0.972 (1)	–	–	–	–
Japan	AR	[0.038] (1)	[0.044] (3)	[0.067] (3)	[0.114] (2)	[0.066] (3)	[0.069] (2)	[0.104] (2)	[0.159] (2)	[0.091] (1)	[0.126] (2)	[0.170] (2)	[0.277] (2)
	NN	1.011 (3)	0.986 (1)	0.966 (1)	0.934 (1)	0.999 (2)	0.997 (1)	0.973 (1)	0.858 (1)	1.000 (2)	0.977 (1)	0.981 (1)	0.974 (1)
	AR-NN	1.046 (4)	1.107 (4)	1.140 (2)	1.175 (4)	0.992 (1)	1.042 (4)	1.062 (4)	1.116 (4)	1.007 (3)	1.203 (3)	1.297 (3)	1.305 (4)
	STAR	1.003 (2)	0.997 (2)	0.994 (2)	1.016 (3)	1.000 (3)	1.000 (2)	1.000 (2)	1.000 (2)	1.124 (4)	1.384 (4)	1.380 (4)	1.220 (3)
UK	AR	[0.043] (1)	[0.047] (1)	[0.062] (1)	[0.092] (3)	[0.040] (1)	[0.052] (2)	[0.073] (3)	[0.101] (3)	[0.116] (3)	[0.291] (2)	[0.501] (2)	[0.703] (1)
	NN	1.008 (3)	1.027 (3)	1.005 (3)	0.997 (2)	1.001 (2)	1.007 (4)	0.997 (2)	0.949 (1)	0.929 (1)	0.922 (1)	0.971 (1)	1.162 (3)
	AR-NN	1.013 (4)	1.062 (4)	1.009 (4)	1.026 (4)	1.025 (4)	1.012 (4)	1.008 (4)	1.000 (2)	0.986 (2)	1.116 (4)	1.278 (4)	1.744 (4)
	STAR	1.003 (2)	1.004 (2)	0.997 (1)	0.973 (1)	1.007 (4)	1.006 (3)	1.012 (4)	1.015 (4)	1.000 (3)	1.000 (2)	1.000 (2)	1.000 (1)
USA	AR	[0.035] (1)	[0.042] (1)	[0.051] (1)	[0.077] (1)	[0.053] (2)	[0.058] (3)	[0.073] (3)	[0.100] (3)	[0.176] (3)	[0.362] (4)	[0.598] (4)	[0.911] (2)
	NN	1.052 (4)	1.060 (3)	1.196 (4)	1.274 (4)	1.019 (3)	0.972 (1)	0.965 (1)	0.903 (1)	1.002 (4)	0.997 (3)	0.997 (3)	1.011 (3)
	AR-NN	1.019 (2)	1.079 (4)	1.179 (3)	1.264 (3)	0.980 (1)	0.980 (2)	0.979 (2)	0.971 (2)	0.985 (2)	0.934 (2)	0.942 (2)	1.020 (4)
	STAR	1.039 (3)	1.029 (2)	1.107 (2)	1.141 (2)	1.032 (4)	1.048 (4)	1.070 (4)	1.092 (4)	0.968 (1)	0.902 (1)	0.891 (1)	0.922 (1)

casting encompassing test statistics are invalid whenever the two models involved are nested, see Clark and McCracken (2001), among others. At first sight, in our case this appears to exclude comparisons between the AR, LSTAR and AR-NN models. However, the linear AR model and the linear components in the LSTAR and AR-NN models do not generally contain the same lags, as they are selected using different techniques. We thus maintain that these models are approximations of the same unknown data-generating process and include them in our comparisons. This argument is not valid, however, in cases where linearity is never rejected against STAR; see Table 1, and these cases are removed from comparisons. The NN model does not contain a linear unit, so it can be tested against all the other models without problem. Table 3 contains results of pairwise model comparisons in terms of MSFE, using the Diebold–Mariano test. The entries represent the number of times the model indicated by row has smaller MSFE than the model indicated by column at the 5% significance level. Forecast encompassing results are shown in Table 4.

Several interesting conclusions emerge from these tables. Results in Table 2 suggest, as expected, that no model or method dominates the others, and the model

Table 3
Point forecast evaluation: testing equal forecast accuracy

		$h=1$				$h=3$				$h=6$				$h=12$			
		AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR
IP	AR	–	1	0	0	–	0	2	0	–	0	2	0	–	0	5	0
	NN	1	–	0	0	2	–	2	1	1	–	3	0	3	–	6	2
	AR-NN	0	0	–	0	0	0	–	0	0	0	–	0	0	0	–	0
	STAR	2	2	2	–	3	3	4	–	3	3	5	–	1	0	5	–
Inflation	AR	–	3	4	2	–	2	5	1	–	3	6	1	–	2	3	2
	NN	1	–	1	1	1	–	4	1	1	–	3	1	1	–	4	1
	AR-NN	0	2	–	0	0	0	–	0	0	0	–	0	0	0	–	0
	STAR	0	4	4	–	1	3	5	–	0	3	6	–	0	2	3	–
M1	AR	–	3	4	1	–	3	4	0	–	3	5	0	–	3	6	0
	NN	0	–	3	0	0	–	4	0	0	–	4	0	0	–	5	0
	AR-NN	0	0	–	0	0	0	–	0	0	0	–	0	0	0	–	0
	STAR	0	1	3	–	4	4	5	–	3	4	5	–	2	4	6	–
STIR	AR	–	1	1	2	–	4	1	1	–	4	1	1	–	4	2	2
	NN	1	–	1	1	0	–	1	0	0	–	0	0	1	–	0	0
	AR-NN	2	2	–	2	0	1	–	1	0	2	–	0	1	1	–	0
	STAR	1	0	1	–	1	1	2	–	2	2	3	–	2	1	2	–
VEX	AR	–	3	1	1	–	4	4	1	–	4	3	1	–	4	4	1
	NN	0	–	1	0	0	–	1	0	1	–	2	0	1	–	2	1
	AR-NN	0	1	–	0	0	2	–	1	0	2	–	0	0	1	–	0
	STAR	1	2	2	–	1	4	5	–	0	3	3	–	0	3	4	–
VIMP	AR	–	0	2	2	–	0	2	1	–	0	2	1	–	0	2	1
	NN	0	–	1	0	0	–	2	1	1	–	2	2	5	–	6	3
	AR-NN	2	1	–	3	0	0	–	1	1	0	–	1	0	0	–	1
	STAR	1	1	2	–	1	1	2	–	2	2	4	–	2	1	4	–
UR	AR	–	0	0	2	–	0	1	1	–	0	3	1	–	0	4	1
	NN	1	–	1	3	1	–	2	1	1	–	3	1	1	–	4	1
	AR-NN	0	0	–	1	0	0	–	0	0	0	–	0	0	0	–	0
	STAR	0	0	0	–	1	1	0	–	1	1	2	–	1	1	3	–
All	AR	–	11	12	10	–	13	19	5	–	14	22	5	–	13	26	7
	NN	4	–	8	5	4	–	16	4	5	–	17	4	12	–	27	8
	AR-NN	4	6	–	6	0	3	–	3	1	4	–	1	1	2	–	1
	STAR	5	10	14	–	12	17	23	–	11	18	28	–	8	12	27	–

Table entries represent the number of times the model indicated by row has significantly smaller MSFE than the model indicated by column according to the pairwise modified Diebold–Mariano test, using a nominal significance level 0.05.

performing best is not the same across countries, variables and forecast horizons. This holds in particular for the forecasting performance of the nonlinear STAR and (AR-)NN models relative to the linear AR model. For example, AR models clearly render the most accurate forecasts for the interest rate series. Incidentally, linearity was systematically rejected for these series. But then, with few exceptions, the linear forecasts are consistently beaten by at least one of the nonlinear models for IP, M1, VIMP and unemployment rates. Hence, for some variables, our results support the conventional wisdom that linear time series models are robust forecasting devices, while for others, it seems that there is considerable scope for forecast improvement by using nonlinear models. It should be mentioned though that there does not exist a single country/variable combination such that all three nonlinear models generate more accurate forecasts than the linear AR model; German IP being the example that comes closest. From Table 3, it is seen that in terms of the MSFE, the linear AR model is rejected 25 times against the NN model, 6 times against the AR-NN specification and 36 times against the LSTAR model. The relative performance of the AR-NN model deteriorates with the forecast horizon.

On the whole, the LSTAR model appears to perform slightly better in terms of raw MSFE

Table 4
Point forecast evaluation: testing forecast encompassing

		$h=1$				$h=3$				$h=6$				$h=12$			
		AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR	AR	NN	AR-NN	STAR
IP	AR	–	5	3	4	–	3	0	4	–	3	0	4	–	5	0	3
	NN	1	–	4	5	0	–	0	3	0	–	0	3	0	–	0	2
	AR-NN	3	3	–	6	4	4	–	6	4	4	–	6	6	7	–	6
	STAR	0	2	3	–	0	3	0	–	0	2	0	–	0	2	0	–
Inflation	AR	–	4	3	1	–	7	4	3	–	6	2	0	–	3	4	0
	NN	6	–	6	6	6	–	4	6	6	–	5	5	5	–	4	4
	AR-NN	6	6	–	7	6	7	–	7	6	7	–	7	6	7	–	6
	STAR	3	5	3	–	1	7	4	–	1	6	3	–	2	4	5	–
M1	AR	–	3	2	5	–	1	1	5	–	1	1	4	–	0	0	4
	NN	5	–	2	7	4	–	0	6	4	–	0	5	3	–	0	5
	AR-NN	6	4	–	6	6	4	–	6	6	5	–	6	6	7	–	7
	STAR	2	2	1	–	0	1	1	–	0	0	0	–	0	0	0	–
STIR	AR	–	2	6	1	–	1	3	2	–	2	1	2	–	1	1	2
	NN	5	–	6	2	6	–	7	6	7	–	7	5	6	–	7	3
	AR-NN	3	4	–	4	5	3	–	6	5	3	–	5	5	3	–	3
	STAR	5	6	6	–	2	2	3	–	2	2	0	–	2	2	1	–
VEX	AR	–	1	0	1	–	1	0	2	–	1	0	1	–	1	0	0
	NN	6	–	4	6	6	–	3	6	5	–	4	6	4	–	2	3
	AR-NN	4	2	–	3	5	4	–	5	6	2	–	6	4	2	–	4
	STAR	2	1	2	–	2	1	1	–	3	1	0	–	1	1	0	–
VIMP	AR	–	2	4	1	–	3	1	1	–	3	1	2	–	5	0	2
	NN	3	–	4	3	1	–	2	3	0	–	1	2	0	–	0	2
	AR-NN	3	3	–	3	4	4	–	3	4	3	–	4	2	6	–	4
	STAR	2	3	4	–	1	2	1	–	1	2	1	–	1	3	1	–
UR	AR	–	3	5	1	–	4	1	1	–	3	0	1	–	2	0	1
	NN	0	–	4	1	0	–	1	1	0	–	0	1	1	–	0	2
	AR-NN	3	3	–	3	3	3	–	3	4	4	–	4	4	4	–	5
	STAR	3	4	4	–	2	3	2	–	1	2	0	–	1	1	0	–
All	AR	–	20	23	14	–	20	10	18	–	19	5	14	–	17	5	12
	NN	26	–	30	30	23	–	17	31	22	–	17	27	19	–	13	21
	AR-NN	28	25	–	32	33	29	–	36	35	28	–	38	33	36	–	35
	STAR	17	23	23	–	8	19	12	–	8	15	4	–	7	13	7	–

Table entries represent the number of times the model indicated by row does not forecast encompass the model indicated by column according to the pairwise modified Diebold–Mariano test, using a nominal significance level 0.05.

values than the neural network models, especially for IP and M1, and for series from Canada, Germany and the US. The relative performance of the LSTAR model often improves considerably with the forecast horizon; see German IP and VIMP, and French and UK M1 for examples. SW in their study found that the individual NN models performed better than the individual LSTAR models, whereas the situation is rather the opposite here. A probable cause for this is that SW used a separate model for each forecast horizon, whereas we employ the same model for all horizons. Because the NN model is a flexible functional form, it suffers less from an omission of the shortest lags in the forecasting model than the tightly parameterized LSTAR model. As a whole, the results in Table 3 indicate that the LSTAR model shows better relative performance than the NN model when the linear AR model is the benchmark. The DM test rejects the linear AR model more often when the alternative is the LSTAR model than it does when the AR model is tested against the NN model.

Regarding the two neural network-based methods, the NN models obtained by pruning a large model clearly perform better on average than the bottom-up procedure employed for the AR-NN model, except for the shortest forecast horizon $h=1$, in particular for IP and unemployment rates. Reasons for this will be discussed in Section 7.3. Sometimes both neural network models produce inferior forecasts: inflation for Japan and UK or even Germany are examples. In forecasting inflation, however, the relative differences between models can be large whereas the absolute ones remain small: note the small RMSFE values for the AR model in brackets.

7.2. Value of careful specification: STAR models

The results in the preceding subsection suggest that the LSTAR model is often among the best ones when it comes to forecasting, and it may be argued that on average, it even performs better than the linear AR model. But then, it would also be useful to know whether the modelling strategy applied to building LSTAR models is an important factor for this result. This is a relevant question because it is possible to just choose an LSTAR model without any model selection and use it for forecasting, as SW did. As already mentioned, this may not always be a very good idea because of the identification problem present when the data-generating process is linear. In fact, results in Table 2 already hint at the possibility that linearity tests provide a useful insurance against outright bad forecasts. There are several series for which linearity is never rejected; see the VIMP series of Canada, France and Japan, for example. Consequently, the LSTAR model rarely fails badly. The reader is reminded of the fact that no in-sample evaluation of LSTAR models by misspecification tests is carried out before forecasting, but the quality of forecasts suggests that serious specification failures have been rare.

In order to investigate the value of careful model building, we define three additional LSTAR specifications. First, an LSTAR model is specified without testing linearity first. This means that the linearity tests are performed as described in Section 3.1.2, but only to select the value of the delay parameter d. The maximum lag p is again determined with BIC. In the second and third LSTAR models, we not only skip linearity testing but in addition fix the lag order p in Eq. (1) at 1 and 6, respectively, and set the delay $d=1$. The whole forecasting exercise is repeated using each of these three model specifications, and the results are compared to the ones obtained using the LSTAR model of previous sections as the benchmark. Table 5 contains the relevant RMSFE ratios, where the RMSFE of the forecasts from the original LSTAR model is the denominator.

The results show, somewhat surprisingly, that testing linearity does not seem to matter very much, but carefully selecting the lag order and delay parameter in the LSTAR model does. The overall RMSFE ratios for the first specification are only marginally above one for forecast horizons of 1, 3 and 6 months, and marginally below one for the 12-month horizon. On the other hand, fixed LSTAR models perform less well, which indicates that selecting the delay parameter d and the lag order p is important. The only exception is the short-term interest rate, which appears to be best predicted by a fixed first-order LSTAR model. A closer scrutiny of the results reveals, however, that the gains originate from just two series: the ones for Canada and the US. Fixed models, both of order one and six, can also fail quite badly, as the results of forecasting the unemployment rate series indicate.

Table 5
Forecasting performance of fixed STAR specifications

	IP				CPI				M1				STIR				VEX				VIMP				UR				All			
	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
STAR	0.999	0.998	0.993	0.990	1.022	1.029	1.034	0.999	1.001	0.990	0.980	0.970	1.002	1.029	1.034	0.999	1.001	0.996	1.001	0.996	0.995	1.008	1.013	1.008	1.003	1.003	1.002	1.013	1.003	1.004	1.004	0.996
$p=1$	1.008	1.009	1.011	1.012	1.077	1.041	1.010	0.985	1.009	0.993	0.987	0.976	0.978	0.981	0.973	0.985	1.016	1.038	1.052	1.038	1.016	1.025	1.021	1.018	1.041	1.123	1.168	1.213	1.020	1.026	1.026	1.025
$p=6$	1.001	1.003	1.008	1.015	0.969	0.973	0.995	1.014	1.038	1.033	1.021	1.011	1.032	1.013	0.994	0.999	1.013	1.026	1.018	1.022	1.014	1.039	1.042	1.045	1.020	1.039	1.095	1.183	1.012	1.017	1.022	1.035

Table entries represent ratios of RMSFE of the fixed STAR specifications, where the "fully specified" STAR model is the baseline model. STAR refers to the specification where the delay parameter d is selected by means of the linearity tests and the lag order p is set according to BIC. $p=1$ and $p=6$ refer to the STAR model; the lag order p in Eq. (1) is fixed at 1 and 6, respectively, and the delay d is set equal to 1.

In order to shed more light on the issue of testing linearity, we repeated the analysis by applying significance levels 0.10 and 0.20 in the linearity tests. For the 10% level, the results were very similar to the ones in Table 6. When the significance level was increased to 0.20, however, the models built conditionally on the results of linearity tests performed somewhat better than the three others. It thus appears that it may be advantageous to reject linearity and select the LSTAR model more often than is done when the 5% significance level for each individual test for $d \in D$ is applied.

7.3. Value of careful specification: AR-NN models

Our previous results suggest that the AR-NN models do not perform as well as the other nonlinear models, including the NN models specified using Bayesian regularization. This could be due either to the different specification strategies employed or to the treatment of the linear unit. Recall that the AR-NN model includes a linear unit, while the Bayesian regularization approach omits it. In order to shed light on these issues, we repeat the forecasting exercise for four alternative NN model specifications. First, we use Bayesian regularization to estimate an NN model with a linear unit included, denoted as NN-L in the following. In addition, we use conditional maximum likelihood to estimate three fixed NN model specifications without a linear unit. The first one, called the S(mall)-NN model, contains three hidden units and just the first lag of the series as the input variable. The second one, the L(arge)-NN model, also has 3 hidden units but uses (the first) six lags as inputs; and, finally, the third one, called the XL(arge)-NN specification, has 10 hidden units and six lags as input variables. Table 6 shows the ratios of the RMSFE of the these models with respect to the "fully specified" AR-NN models.

This table gives rise to three clear-cut conclusions. First, inclusion of a linear hidden unit improves the forecasting performance. This follows from the observation that the NN-L model renders more accurate forecasts than the NN model, while the same generally holds for the AR-NN models compared with the fixed specifications S-, L- and XL-NN. Second, specification (in the sense of selecting the input variables and the number of hidden units) is an important factor. The

Table 6
Forecasting performance of neural network specifications

	IP				CPI				M1				STIR				VEX				VIMP				UR				All			
	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
S-NN	1.024	1.003	1.004	1.013	1.024	0.964	1.005	1.109	1.009	0.985	0.995	1.040	0.968	1.007	1.139	1.386	0.998	0.997	1.031	1.089	1.025	0.961	1.035	1.148	0.999	0.980	1.007	1.031	1.007	0.985	1.031	1.116
L-NN	1.059	1.014	1.006	0.995	1.064	0.978	0.942	0.951	1.037	0.992	0.955	0.991	1.000	1.021	1.054	1.091	1.005	0.936	0.951	0.951	1.010	0.926	0.944	0.960	0.999	0.999	1.000	1.018	1.025	0.980	0.979	0.994
XL-NN	1.136	1.053	1.035	1.001	1.185	1.097	1.054	1.028	1.144	1.104	1.067	1.030	1.082	1.149	1.218	1.291	1.096	1.033	1.008	1.008	1.148	0.997	0.993	0.960	1.113	1.063	1.098	1.089	1.129	1.071	1.068	1.058
NN-L	1.030	0.989	0.940	0.903	0.974	0.900	0.864	0.862	0.963	0.925	0.894	0.864	0.928	0.951	0.965	0.951	0.939	0.882	0.863	0.791	0.934	0.865	0.870	0.834	0.995	0.968	0.956	0.939	0.966	0.926	0.908	0.878
NN	1.049	1.024	0.972	0.929	1.000	0.913	0.869	0.865	1.020	0.980	0.951	0.928	0.964	0.990	1.010	0.971	0.966	0.901	0.909	0.852	0.979	0.887	0.909	0.897	1.004	0.971	0.960	0.925	0.997	0.952	0.940	0.910

Table entries represent ratios of RMSFE of fixed NN specifications estimated with Bayesian regularization, where the "fully specified" AR-NN model is the baseline model. S-NN, L-NN and XL-NN contain three, three and ten hidden units, respectively, and one, six and six lags as input variable(s), respectively. NN-L and NN are neural network models estimated with Bayesian regularization with and without a linear unit, respectively.

three fixed model specifications do not fare well. In fact, the XL-NN specification is the worst performing model on average, while the S-NN model performs badly, in particular at long forecast horizons. Third, NN models specified with Bayesian regularization outperform the AR-NN models, as the RMSFE ratios for both the NN and NN-L models are always less than one. The problem with the estimated AR-NN models turns out to be that they are frequently explosive or close to nonstationary. This is only possible when the model contains an autoregressive linear unit; see Trapletti, Leisch, and Hornik (2000). This is less of a problem in the NN-L model because the parameters of the linear unit are also shrunk towards zero. We conclude that in-sample evaluation of NN models is important. In this study, it has not been possible to evaluate every AR-NN model before using it for forecasting and, due to this omission, explosive AR-NN models have been used. Damage control has only occurred in the form of the insanity filter, which apparently does not work sufficiently well. A possible way of avoiding explosive models could be to shrink the parameters of the linear unit towards zero in the estimation as is done in the context of Bayesian regularization.

7.4. Combining point forecasts

As indicated in Section 4, our aim is also to consider the accuracy of combination forecasts. We consider the three possible combinations of pairs involving a linear AR model and two combinations that only involve nonlinear models. The latter ones are labelled NN+STAR and AR-NN+STAR.

Table 7 contains results for the RMSFE comparisons. The benchmark is again the linear AR model, which means that the entries in the table are ratios of the RMSFE of the combination forecast and the corresponding RMSFE of the linear AR model. A general conclusion is that combining often improves forecast accuracy, unless one of the models generates strongly inferior forecasts. In that case, the forecasts are more accurate than the ones from the inferior model but still less accurate than the linear AR benchmark. In this study, the inferior model is most often the AR-NN model.

Sometimes, the combination of two nonlinear models produces very good results. A case in point is the

Table 7
Performance of combination forecasts

		IP				CPI				M1				STIR				VEX				VIMP				UR			
		1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
Canada	AR+NN	0.989	0.960	0.936	0.926	1.000	1.000	1.000	1.000	0.990	0.980	0.996	1.010	1.026	0.987	0.971	1.030	1.001	0.996	0.996	0.995	1.000	1.000	1.000	1.000	0.998	0.982	0.973	0.948
	AR+AR-NN	0.982	1.001	1.029	1.111	0.962	1.011	1.067	1.022	0.993	0.999	1.035	1.080	0.956	0.971	0.987	1.016	0.998	0.992	0.999	1.000	0.957	1.001	0.997	0.996	0.962	0.969	1.014	1.096
	AR+STAR	0.996	0.994	0.993	0.969	1.000	0.999	0.995	0.935	0.994	1.011	1.034	1.072	0.998	1.042	1.035	1.082	1.014	1.027	1.019	1.022	0.982	0.993	0.987	0.978	0.985	0.983	0.983	0.975
	NN+STAR	0.985	0.954	0.929	0.893	1.000	0.999	0.995	0.935	0.994	0.994	1.033	1.084	1.030	1.041	1.021	1.119	1.016	1.024	1.016	1.017	0.982	0.993	0.987	0.978	0.986	0.969	0.960	0.927
	AR-NN+STAR	0.975	0.968	0.970	1.040	0.962	1.011	1.067	1.022	0.992	0.983	1.035	1.093	0.982	0.961	0.964	1.050	1.000	0.989	0.995	0.994	0.957	1.001	0.997	0.996	0.962	0.956	0.991	1.053
France	AR+NN	0.988	0.988	0.977	0.981	1.009	1.014	1.027	1.029	1.017	0.936	0.888	0.837	0.983	0.985	1.002	1.104	0.978	0.984	1.026	1.094	1.000	1.000	1.000	1.000	–	–	–	–
	AR+AR-NN	1.001	1.012	1.021	1.051	1.036	1.163	1.283	1.221	1.023	1.078	1.154	1.236	0.931	0.987	1.017	1.062	0.996	1.002	1.058	1.146	1.005	1.017	1.023	1.040	–	–	–	–
	AR+STAR	0.993	0.983	0.983	0.979	0.951	0.954	0.933	0.998	1.031	1.029	1.058	1.084	0.992	1.043	1.066	1.111	0.994	1.000	1.044	1.110	0.991	0.997	0.993	1.000	–	–	–	–
	NN+STAR	0.981	0.973	0.962	0.962	0.959	0.968	0.961	1.028	1.049	0.960	0.940	0.913	0.979	1.026	1.068	1.220	0.975	0.993	1.087	1.213	0.991	0.997	0.993	1.000	–	–	–	–
	AR-NN+STAR	0.994	1.005	1.004	1.034	1.044	1.177	1.302	1.247	1.038	1.001	1.021	1.052	0.937	0.987	1.033	1.179	0.982	1.000	1.106	1.250	1.005	1.017	1.023	1.040	–	–	–	–
Germany	AR+NN	0.980	0.943	0.907	0.883	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.005	1.016	1.040	1.082	1.000	1.000	1.000	1.000	0.987	0.967	0.941	0.914	1.007	1.005	0.997	0.967
	AR+AR-NN	0.979	1.000	1.001	0.998	0.944	0.976	1.046	0.965	0.995	0.992	1.004	1.070	0.970	1.013	1.020	1.037	0.998	1.003	1.001	1.002	1.003	1.013	1.012	1.009	0.961	0.945	1.024	1.177
	AR+STAR	0.991	0.993	0.982	0.955	0.983	0.963	1.043	1.036	1.002	0.998	0.995	1.009	0.958	0.996	0.984	0.977	1.003	1.006	1.008	1.018	1.001	0.986	0.976	0.944	0.979	0.982	0.977	0.982
	NN+STAR	0.973	0.939	0.894	0.840	0.983	0.963	1.043	1.036	1.002	0.998	0.995	1.009	0.962	1.009	1.023	1.066	1.003	1.006	1.008	1.018	0.990	0.957	0.922	0.860	0.986	0.985	0.970	0.945
	AR-NN+STAR	0.962	0.943	0.907	0.879	0.944	0.976	1.046	0.965	0.995	0.992	1.004	1.070	0.974	1.027	1.059	1.129	0.998	1.003	1.001	1.002	0.990	0.978	0.952	0.923	0.966	0.947	1.015	1.139
Italy	AR+NN	1.000	1.000	1.000	1.000	0.967	0.947	0.991	0.988	0.987	0.978	0.951	0.924	1.025	1.043	1.081	1.087	1.002	1.010	1.018	1.017	1.013	0.996	0.990	0.985	–	–	–	–
	AR+AR-NN	1.001	1.006	1.008	1.020	1.015	0.954	0.968	0.950	1.005	0.990	0.992	1.045	1.139	1.164	1.177	1.189	0.998	0.998	1.003	1.007	0.996	0.997	1.002	1.005	–	–	–	–
	AR+STAR	0.991	0.991	0.975	0.944	1.010	0.933	0.982	0.926	0.999	0.991	0.985	1.017	1.022	1.111	1.152	1.191	0.990	1.011	1.016	1.007	0.988	1.000	0.991	0.985	–	–	–	–
	NN+STAR	0.991	0.991	0.975	0.944	0.988	0.890	0.974	0.916	0.994	0.981	0.950	0.960	1.048	1.155	1.235	1.285	0.998	1.028	1.039	1.028	0.999	0.995	0.978	0.970	–	–	–	–
	AR-NN+STAR	1.001	1.006	1.008	1.020	0.987	0.906	0.962	0.942	1.000	0.979	0.955	0.985	1.168	1.208	1.253	1.272	1.003	1.009	1.021	1.025	1.009	0.993	0.992	0.991	–	–	–	–
Japan	AR+NN	1.000	1.000	1.001	1.000	0.997	0.981	0.997	1.003	0.979	0.930	0.894	0.888	0.987	0.982	0.964	0.918	1.001	0.996	0.994	1.006	1.000	1.000	1.000	1.000	1.041	1.162	1.165	1.090
	AR+AR-NN	0.983	1.003	1.022	1.093	1.080	1.103	1.113	1.033	1.020	1.035	1.061	1.188	0.952	0.995	1.000	1.010	1.015	1.039	1.059	1.084	0.993	1.011	1.023	1.054	0.982	1.056	1.096	1.107
	AR+STAR	0.994	0.984	0.984	0.941	0.984	1.011	1.084	1.139	1.004	1.010	1.018	1.043	0.954	1.014	1.049	1.027	1.004	0.991	0.982	0.966	0.998	0.993	0.980	0.924	0.996	0.986	0.987	0.982
	NN+STAR	0.994	0.984	0.985	0.941	0.982	0.995	1.081	1.139	0.983	0.938	0.910	0.926	0.943	1.000	1.021	0.953	1.005	0.987	0.976	0.973	0.998	0.993	0.980	0.924	1.037	1.148	1.152	1.075
	AR-NN+STAR	0.983	1.003	1.023	1.093	1.085	1.095	1.105	1.035	0.997	0.962	0.953	1.069	0.940	0.979	0.966	0.928	1.018	1.040	1.059	1.093	0.993	1.011	1.023	1.054	1.020	1.206	1.245	1.192
UK	AR+NN	1.000	1.000	1.000	1.000	1.022	0.995	0.999	1.006	0.953	0.890	0.787	0.727	1.009	1.020	1.038	1.045	0.999	1.001	0.997	0.984	0.999	1.000	1.005	1.007	1.000	1.000	1.000	1.000
	AR+AR-NN	1.002	1.001	0.998	1.006	1.085	1.344	1.418	1.274	1.025	1.132	1.094	1.096	0.994	0.983	0.997	1.005	1.005	1.021	0.998	1.007	1.009	1.004	1.003	0.999	0.959	0.986	1.054	1.276
	AR+STAR	1.007	0.997	0.997	0.989	1.071	1.021	1.148	1.137	0.992	1.009	1.000	1.004	0.993	1.013	1.062	1.093	1.001	1.011	1.002	0.998	0.998	1.001	0.996	0.972	0.938	0.932	0.954	1.037
	NN+STAR	1.007	0.997	0.997	0.989	1.095	1.024	1.152	1.141	0.953	0.904	0.797	0.746	1.003	1.035	1.102	1.139	1.002	1.013	0.999	0.982	0.997	1.001	1.001	0.979	0.938	0.932	0.954	1.037
	AR-NN+STAR	1.002	1.001	0.998	1.006	1.113	1.360	1.421	1.274	0.983	1.029	0.892	0.842	1.001	1.001	1.033	1.049	1.005	1.023	0.996	0.992	1.011	1.006	1.009	1.006	0.959	0.986	1.054	1.276
USA	AR+NN	0.969	0.962	0.952	0.945	1.000	1.000	1.000	1.000	0.992	0.977	0.956	0.922	1.026	0.970	0.910	0.919	1.013	1.008	1.046	1.066	1.011	1.021	1.032	1.044	0.972	0.938	0.934	0.951
	AR+AR-NN	0.995	0.995	1.005	1.038	0.978	0.992	1.041	1.022	1.005	1.032	1.064	1.098	0.960	0.979	0.956	0.922	0.997	1.022	1.078	1.128	0.988	0.988	0.988	0.984	0.983	0.952	0.951	0.984
	AR+STAR	0.995	0.999	0.999	0.994	0.943	0.947	0.969	0.942	0.992	0.998	1.000	0.986	0.983	0.965	0.942	0.908	1.017	1.019	1.085	1.132	0.997	0.971	0.965	0.941	0.999	0.997	0.998	1.004
	NN+STAR	0.964	0.961	0.951	0.939	0.943	0.947	0.969	0.942	0.986	0.978	0.962	0.917	1.016	0.950	0.868	0.836	1.036	1.036	1.143	1.201	1.011	0.989	0.993	0.984	0.973	0.936	0.932	0.956
	AR-NN+STAR	0.967	0.960	0.959	0.985	0.978	0.992	1.041	1.022	1.001	1.016	1.032	1.028	0.993	0.963	0.886	0.844	1.015	1.040	1.133	1.196	1.000	1.010	1.021	1.029	0.963	0.900	0.895	0.942

Table entries are ratios of root-mean-squared forecast errors of the models by variable, country, and forecast horizon, where the linear AR model is the baseline model.

Canadian CPI. Combining forecasts from the NN and LSTAR model leads to remarkably good forecasts for longer horizons, even though the forecasts from the NN model are not particularly accurate; see Table 2. Combinations in which the linear AR model is one of the components are conservative, in the sense that they further emphasize the linear model. For example, combining the linear and the LSTAR model often lead to forecasts that are slightly more accurate than the forecasts from the linear model. This is due to the fact that some of the STAR forecasts may be "linear" in the sense that they arise from a linear model. This happens when linearity is not rejected against STAR, so that the model actually used for forecasting is the linear AR model.

8. Conclusions

In this paper, we consider the forecast accuracy of a linear AR model and three different nonlinear models, the LSTAR model and two neural network models. A general result that emerges is that in order to obtain acceptable results with nonlinear models, modelling has to be carried out with care. When it comes to neural network models, there seems to be a risk of explosive models, and thus for implausible forecasts at long forecasting horizons. Controls have to be applied to detect them and to replace them by simple rule-of-thumb forecasts.

The first question posed in the introduction was whether nonlinear models produce real-time forecasts that improve upon linear models. The answer seems to be mixed. It appears that LSTAR models generate forecasts that are to some extent more accurate than forecasts from linear models. The same holds for NN models specified with Bayesian regularization, but not for the AR-NN models. The answer to the second question thus appears to be that tightly parameterized models, here represented by the LSTAR family, have an edge over more nonparametric approaches such as neural network models.

Furthermore, combining forecasts improves the accuracy of point forecasts. This answer to the third question in the introduction is not without reservations, but by and large our results seem favourable to the idea. It should be noticed, however, that gains from pooling forecasts may be more substantial if the number of forecasts is large, which is not the case here. Finally, there is no unique answer to the final question concerning the amount of gain in forecast accuracy from nonlinear models. Whether or not these gains are worthwhile depends on how large the costs of careful nonlinear model specification are estimated to be compared to the improvement in forecast accuracy achieved by these models.

Our results are not fully comparable with the results in SW and Marcellino (2005). As already mentioned, these authors built a separate model for each forecast horizon, whereas in this study, the same model has been used for generating forecasts for all horizons. Whether or not this is an important difference is worthy of investigation, but this is left for future research. At any rate, our results indicate that building nonlinear models with care has a positive effect on forecast accuracy. This is true for LSTAR models and it should also be true for neural network models. It appears, however, that the possibility of obtaining explosive models when applying the modelling technique presented in Medeiros et al. (2005) for AR-NN models has to be accounted for. For example, in estimating the final AR-NN model, the coefficients of the linear unit may be shrunk towards zero, which has not been attempted in this paper. It is obvious that building these models requires more individual care than it has been possible to exercise in our simulated forecasting experiment.

The results in this paper are based on the implicit assumption that all models have constant parameters during the estimation period. Evaluation of models should include testing parameter constancy. This requirement is difficult to satisfy in a study with a large number of time series and models, and therefore, no evaluation tests have been carried out here, although such tests do exist. It may therefore be possible to build better models than the ones used for forecasting in this study. Applying a rolling window in modelling may also mitigate the effects of parameter change should this be a problem. But then, if the number of series to be predicted is large, it may be that the forecaster cannot devote sufficient resources to building the required forecasting models. The results of this study at least indicate that when one considers choosing a forecasting model from a large family of models, careful specification (selecting a member of this family) may substantially improve the precision of forecasts.

Acknowledgments

This research has been supported by Jan Wallander's and Tom Hedelius's Foundation, Project No. J02-35. The first version of this paper was prepared for the First International Institute of Forecasters' Workshop "Nonlinearities, Business Cycles and Forecasting," Madrid, December 2003. Material from the paper has been presented at the workshop in honour of Clive W.J. Granger entitled "Predictive Methodology and Application in Economics and Finance," La Jolla, CA, January 2004, the conference "Recent Advances in Time Series Analysis," Protaras, Cyprus, June 2004, the 24th International Symposium on Forecasting, Sydney, July 2004, and seminars at Stockholm School of Economics and Magyar Nemzeti Bank, Budapest. Comments from the participants of these occasions, Alfonso Novales and Mark Watson in particular, are gratefully acknowledged. Any errors and shortcomings in this paper remain our own responsibility.

References

Bacon, D. W., & Watts, D. G. (1971). Estimating the transition between two intersecting straight lines. *Biometrika*, *58*, 525–534.

Boero, G., & Marrocu, E. (2002). The performance of non-linear exchange rate models: A forecast comparison. *Journal of Forecasting*, *21*, 513–542.

Chan, K. S., & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, *7*, 178–190.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*, 85–110.

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

Clements, M. P., & Smith, J. (1999). A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics*, *14*, 123–141.

Cybenko, G. (1989). Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, *2*, 303–314.

Fine, T. L. (1999). *Feedforward neural network methodology*. Berlin: Springer-Verlag.

Foresee, F. D., & Hagan, M. T. (1997). Gauss–Newton approximation to Bayesian regularization. *IEEE International Conference on Neural Networks*, *vol. 3* (pp. 1930–1935). New York: IEEE.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Granger, C. W. J., & Bates, J. (1969). The combination of forecasts. *Operation Research Quarterly*, *20*, 451–468.

Granger, C. W. J., & Jeon, Y. (2004). Thick modeling. *Economic Modelling*, *21*, 323–343.

Granger, C. W. J., & Teräsvirta, T. (1993). *Modelling nonlinear economic relationships*. Oxford: Oxford University Press.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*, 281–291.

Harvey, D., Leybourne, S., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, *16*, 254–259.

Heravi, S., Osborn, D. R., & Birchenhall, C. R. (2004). Linear versus neural network forecasts for European industrial production series. *International Journal of Forecasting*, *20*, 435–446.

Hornik, K., Stinchombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Kilian, L., & Taylor, M. (2003). Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, *60*, 85–107.

Krolzig, H.-M., & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*, 831–866.

Lundbergh, S., & Teräsvirta, T. (2002). Forecasting with smooth transition autoregressive models. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 485–509). Oxford: Blackwell.

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, *75*, 491–499.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation*, *4*, 415–447.

MacKay, D. J. C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Computation*, *4*, 448–472.

Marcellino, M. (2005). Instability and non-linearity in the EMU. In C. Milas, P. Rothman, & D. van Dijk (Eds.), *Nonlinear Time Series Analysis of Business Cycles*. Amsterdam: Elsevier.

Medeiros, M. C., Teräsvirta, T., & Rech, G. (2005). Building neural network models for time series: A statistical approach. *Journal of Forecasting,* forthcoming.

Rech, G. (2002). Forecasting with artificial neural network models. *SSE/EFI Working Paper Series in Economics and Finance*, *vol. 491*. Stockholm School of Economics.

Rech, G., Teräsvirta, T., & Tschernig, R. (2001). A simple variable selection technique for nonlinear models. *Communications in Statistics. Theory and Methods*, *30*, 1227–1241.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Sarantis, N. (1999). Modelling non-linearities in real effective exchange rates. *Journal of International Money and Finance*, *18*, 27–45.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *4*, 461–464.

Siliverstovs, B., & van Dijk, D. (2003). Forecasting industrial production with linear, nonlinear, and structural change models. *Econometric Institute Report* EI 2003-16, Erasmus University Rotterdam.

Skalin, J., & Teräsvirta, T. (2002). Modeling asymmetries and moving equilibria in unemployment rates. *Macroeconomic Dynamics*, *6*, 202–241.

Stock, J. H., & Watson, M. W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. F. Engle, & H. White (Eds.), *Cointegration, causality and forecasting. A festschrift in honour of Clive W.J. Granger* (pp. 1–44). Oxford: Oxford University Press.

Swanson, N. R., & White, H. (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics*, *13*, 265–275.

Swanson, N. R., & White, H. (1997a). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting*, *13*, 439–461.

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

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, *89*, 208–218.

Teräsvirta, T. (1998). Modeling economic relationships with smooth transition regressions. In A. Ullah, & D. E. Giles (Eds.), *Handbook of applied economic statistics* (pp. 507–552). New York: Dekker.

Teräsvirta, T., & Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, *7*, S119–S136.

Teräsvirta, T., Lin, C.-F., & Granger, C. W. J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, *14*, 309–323.

Tkacz, G. (2001). Neural network forecasting of Canadian GDP growth. *International Journal of Forecasting*, *17*, 57–69.

Trapletti, A., Leisch, F., & Hornik, K. (2000). Stationary and integrated autoregressive neural network processes. *Neural Computation*, *12*, 2427–2450.

van Dijk, D., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models—a survey of recent developments. *Econometric Reviews*, *21*, 1–47.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*, 535–550.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*, 35–62.

Timo Teräsvirta is Professor of Econometrics at the Stockholm School of Economics, Sweden. His main research interest is time series econometrics, nonlinear models and modelling in particular. He is a co-author of a book on nonlinear econometrics and has published a number of articles in international journals.

Dick van Dijk is Associate Professor at the Econometric Institute, Erasmus University Rotterdam, and Associate Director of the Erasmus Research Institute of Management (ERIM). His research interests include nonlinear time series analysis, business cycle analysis, and financial econometrics.

Marcelo C. Medeiros is assistant professor at the Department of Economics of the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). His main research interest is nonlinear time series econometrics and the link between machine learning and econometric theory.