SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2011-27

Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques

Anders Bredahl Kock and Timo Teräsvirta

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

FORECASTING MACROECONOMIC VARIABLES USING NEURAL NETWORK MODELS AND THREE AUTOMATED MODEL SELECTION TECHNIQUES

ANDERS BREDAHL KOCK AND TIMO TERÄSVIRTA
AARHUS UNIVERSITY AND CREATES

Abstract. In this paper we consider the forecasting performance of a well-defined class of flexible models, the so-called single hidden-layer feedforward neural network models. A major aim of our study is to find out whether they, due to their flexibility, are as useful tools in economic forecasting as some previous studies have indicated. When forecasting with neural network models one faces several problems, all of which influence the accuracy of the forecasts. First, neural networks are often hard to estimate due to their highly nonlinear structure. In fact, their parameters are not even globally identified. Recently, White (2006) presented a solution that amounts to converting the specification and nonlinear estimation problem into a linear model selection and estimation problem. He called this procedure the QuickNet and we shall compare its performance to two other procedures which are built on the linearisation idea: the Marginal Bridge Estimator and Autometrics. Second, one must decide whether forecasting should be carried out recursively or directly. Comparisons of these two methodss exist for linear models and here these comparisons are extended to neural networks.

Finally, a nonlinear model such as the neural network model is not appropriate if the data is generated by a linear mechanism. Hence, it might be appropriate to test the null of linearity prior to building a nonlinear model. We investigate whether this kind of pretesting improves the forecast accuracy compared to the case where this is not done.

Keywords: artificial neural network, forecast comparison, model selection, nonlinear autoregressive model, nonlinear time series, root mean square forecast error, Wilcoxon's signed-rank test

JEL Classification Codes: C22, C45, C52, C53

Date: August 24, 2011.
Financial support from CREATES, funded by the Danish National Research Foundation, is gratefully acknowledged. Part of this work was carried out when the first author was visiting the Department of Economics at the University of California, Berkeley, and the second author the Department of Economics at the European University Institute, Florence. We are thankful for the kind hospitality of these institutions during our visits. Material from this paper has been presented at the workshop in Econometric Aspects of Price Transmission Analysis, Georg-August University of Göttingen, August 2010, the 19th Symposium of the Society of Nonlinear Dynamics and Econometrics, Washington DC, March 2011, the 31st International Annual Symposium in Forecasting, Prague, June 2011, and seminars at Banque de France and the European University Institute, Florence. We thank participants of these occasions for their comments. The authors are solely responsible for any errors and shortcomings in this work. email: `akock@creates.au.dk` and `tterasvirta@econ.au.dk`.

1. Introduction

Artificial Neural Networks (ANN) have been quite popular in many areas of science for describing various phenomena and forecasting them. They have also been used in forecasting macroeconomic time series and financial series, see Kuan and Liu (1995) for a successful example on exchange rate forecasting, and Zhang et al. (1998) and Rech (2002) for more mixed results. The main argument in their favour is that ANNs are universal approximators, which means that they are capable of approximating arbitrarily accurately functions satisfying only mild regularity conditions. The ANN models thus have a strong nonparametric flavour. One may therefore expect them to be a versatile tool in economic forecasting and adapt quickly to rapidly changing forecasting situations. Recently, Ahmed et al. (2010) conducted an extensive forecasting study comprising more than 1000 economic time series from the M3 competition Makridakis and Hibon (2000), and a large number of what they called machine learning tools. They concluded that the ANN model that we are going to consider, the single hidden-layer feedforward ANN model or multi-layer perceptron with one hidden layer, was one of the best or even the best performer in their study. A single hidden-layer ANN model is already a universal approximator; see Cybenko (1989) and Hornik et al. (1989).

A major problem in the application of ANN models is the specification and estimation of these models. A large number of modelling strategies have been developed for the purpose. It is possible to begin with a small model and increase its size ("specific-to-general", "bottom up", or "growing the network"). Conversely, one can specify a network with a large number of variables and hidden units or "neurons" and then reduce its size ("general-to-specific", "top down" or "pruning the network"). Since the ANN model is nonlinear in parameters, its parameters have to be estimated numerically, which may be a demanding task if the number of parameters in the model is large. Recently, White (2006) devised a clever strategy for modelling ANNs that converts the specification and ensuing nonlinear estimation problem into a linear model selection problem. This greatly simplifies the estimation stage and alleviates the computational effort. It is therefore of interest to investigate how well this strategy performs in macroeconomic forecasting. A natural benchmark in that case is a linear autoregressive model.

Quite often, application of White's strategy leads to a situation in which the number of variables in the set of candidate variables exceeds the number of observations. The strategy handles these cases without problems, because it essentially works from specific to general and then back again. We shall also consider a one-way variant from specific to general in this study. One may want to set a maximum limit for variables to be included in the model to control its size.

There exist other modelling strategies that can also be applied to selecting the variables. In fact, White (2006) encouraged comparisons between his method and other alternatives, and here we shall follow his suggestion. In this work, we consider two additional specification techniques. One is Autometrics by Doornik (2009), see also Krolzig and Hendry (2001) and Hendry and Krolzig (2005), and the other one is the Marginal Bridge Estimator

(MBE), see Huang et al. (2008). The former is designed for econometric modelling, whereas the latter one has its origins in statistics. Autometrics works from general to specific, and the same may be said about MBE. We shall compare the performance of these three methods when applying White's idea of converting the specification and estimation problem into a linear model selection problem and selecting hidden units for our ANN models. That is one of the main objectives of this paper.

The focus in this study is on multiperiod forecasting. There are two ways of generating multiperiod forecasts. One consists of building a single model and generating the forecasts for more than one period ahead recursively. The other one, called direct forecasting, implies that a separate model is built for each forecasting horizons, and no recursions are involved. For discussion, see for example Teräsvirta (2006), Teräsvirta et al. (2010, Chapter 14), or Kock and Teräsvirta (2011). In nonlinear forecasting, the latter method appears to be more common, see for example Stock and Watson (1999) and Marcellino (2002), whereas Teräsvirta et al. (2005) constitutes an example of the former alternative. A systematic comparison of the performance of the two methods exists, see Marcellino et al. (2006), but it is restricted to linear autoregressive models. Our aim is to extend these comparisons to nonlinear ANN models.

Nonlinear models can sometimes generate obviously insane forecasts. One way of alleviating this problem is to use insanity filters as in Swanson and White (1995, 1997a,b) who discuss this issue. We will compare two filters to the unfiltered forecasts and see how they impact on the forecasting performance of the neural networks.

In this work the ANN models are augmented by including lags of the variable to be forecast linearly in them. As a result, the augmented models nest a linear autoregressive model. It is well known that if the data-generating process is linear, the augmented ANN model is not even locally identified; see for example Lee et al. (1993), Teräsvirta et al. (1993) or Teräsvirta et al. (2010, Chapter 5) for discussion. A general discussion of identification problems in ANN models can be found in Hwang and Ding (1997). It may then be advisable to first test linearity of each series under consideration before applying any ANN modelling strategy to it. But then, it may also be argued that linearity tests are unnecessary, because the set of candidate variables can be (and in our case is) defined to include both linear lags and hidden units. The modelling technique can then choose among all of them and find the combination that is superior to the others. We shall compare these two arguments. This is done by carrying out pretesting and only fitting an ANN model to the series if linearity is rejected. Forecasts are generated from models specified this way and compared with forecasts from the ANN models obtained using White's method and the three automatic modelling techniques.

The main criterion of comparing forecasts is the Root Mean Square Forecast Error (RMSFE), which implies a quadratic loss function. Other alternatives are possible, but the RMSFE is commonly used and thus even applied here. We rank the methods, which makes some comparisons possible. Furthermore, we also carry out Wilcoxon signed rank tests but principally for

descriptive purposes, so the tests are not used as an ex post model selection criterion; see Costantini and Kunst (2011) for a discussion.

It might be desirable to compare White's method with modelling strategies which are not based on linearising the problem but in which statistical methods such as hypothesis testing and nonlinear maximum likelihood estimation are applied. Examples of these include Swanson and White (1995, 1997a,b), Anders and Korn (1999) and Medeiros et al. (2006). These approaches do, however, require plenty of human resources, unless the number of time series under consideration and forecasts generated from them are small. This is because nonlinear iterative estimation cannot be automated and the algorithms left to their own devices. Each estimation needs a non-negligible amount of tender loving care, and when the number of time series to be considered is large, ANN model building and forecasting tend to require a substantial amount of resources.

In this paper we investigate the forecasting performance of the above techniques. We first conduct a small simulation study to see how well these techniques perform when the data are generated by a known nonlinear model. The economic data sets consist of the monthly unemployment and consumer price index series from the 1960's until 2009.

The plan of the paper is as follows. The neural network model is presented in Section 2 and estimation techniques in Section 3. The recursive and direct forecasting methods are discussed in Section 4 and the results are summarized in Section 5, while Section 6 concludes.

2. The model

We begin by briefly introducing the Artificial Neural Network (ANN) model and reviewing some of its properties. The techniques for specifying the structure of the model and estimating the parameters will be considered in the next section. Our model is the so-called single-hidden-layer feedforward autoregressive neural network model or single-hidden-layer perceptron

$$y_t = \beta_0'\mathbf{z}_t + \sum_{j=1}^{q} \beta_j (1 + \exp\{\gamma_j'\mathbf{z}_t\})^{-1} + \varepsilon_t \tag{1}$$

where $\mathbf{z}_t = (1, y_{t-1}, ..., y_{t-p})'$, $\beta_0 = (\beta_{00}, \beta_{01}, ..., \beta_{0p})'$, $\gamma_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{jp})$ and $\varepsilon_t \sim \text{iid}\mathcal{N}(0, \sigma^2)$. The weak stationarity condition of (1) is the same as that of the corresponding linear AR(p) model. The ANN model is a so-called universal approximator in the following sense. Suppose there is a functional relationship between y and $\mathbf{z}$: $y = H(\mathbf{z})$. Then under appropriate regularity conditions for any $\delta > 0$ there exists a positive integer $q < \infty$ such that $\left\| H(\mathbf{z}) - \sum_{j=1}^{q} \beta_j (1 + \exp\{\gamma_j'\mathbf{z}\})^{-1} \right\| < \delta$ where $\|\cdot\|$ is an appropriate norm. This indicates that (1) is a very flexible functional form and thus in principle capable of satisfactorily approximating various nonlinear processes.

Before forecasting with the model (1), the number of logistic functions or hidden units q has to be specified and its parameters estimated. Various specification techniques have been proposed in the literature. One possibility is to begin with a large model (large q) and reduce the size of the model, that is, to prune the network. Another possibility is to begin with a small

model and add hidden units, which is called 'growing the network'. Either way, one also has to estimate the parameters of the model which, given that it is heavily nonlinear, may be numerically demanding, in particular when q is large. For discussion, see for example Fine (1999, Chapter 6), Goffe et al. (1994), or Simon (1999).

Nevertheless, if the parameter vectors γ_j, $j = 1, ..., q$, are known, the model is linear in parameters. This opens up the possibility to combine specification and estimation into a single linear model selection problem. White (2006) suggested this technique for specifying and estimating artificial neural network models. The linear model selection problem encountered is the one of choosing a subset of variables from the set

$$S = \{y_{t-i}, i = 1, ..., p; (1 + \exp\{\gamma_j' \mathbf{z}_t\})^{-1}, j = 1, ..., M\} \tag{2}$$

where M is large. Since the quality of the estimates depends on the size of S, the number of variables in a typical macroeconomic application is likely to exceed the number of observations. Model selection techniques that can handle such a situation are discussed in the next section.

The neural network model (1) is not the only possible universal approximator for this application. White (2006) mentions ridgelets, Candès (1998, 2003), as an alternative. Polynomials would probably in this context not be the best possible class of universal approximators. The fit of the estimated polynomials often deteriorates at both ends of the series they describe, which is not a desirable feature in forecasting economic variables such as growth rates. Another universal approximator, the Fourier Flexible Form (FFF), is discussed in Gallant (1984). In applying the FFF, the problem of constructing the variables would have two aspects. One would have to choose the linear combinations $\gamma_j' \mathbf{z}_t$, but one would also have to decide the number of frequencies in the sum of trigonometric components. We settle for the ANN model, because it is, alongside the polynomials, probably the most commonly used universal approximator, and because QuickNet was originally designed to solve the specification and estimation problem for this model.

3. Modeling with three automatic model selection algorithms

We consider three model selection algorithms that apply to our modelling problem, in which the number of variables exceeds the number of observations. They are Autometrics, constructed by Doornik (2009), Marginal Bridge Estimator (MBE), see Huang et al. (2008), and QuickNet, White (2006). Autometrics is built on the principle of moving from general to specific, which means beginning with a large model and gradually reducing its size. QuickNet may be characterised as a specific-to-general-to specific procedure, although we shall also report results on a simplified specific-to-general version. The starting-point of MBE also involves all variables, but the process of selecting the final model is very different from Autometrics. We shall now describe these three techniques in more detail, beginning with Autometrics.

3.1. **Autometrics.**

Modelling begins with a linear model called the General Unrestricted Model (GUM). When the number of variables is less than the number of observations the GUM contains all candidate variables. The

model is subjected to significance tests. If all variables have statistically significant coefficient estimates, the GUM is the final model. Otherwise, because there is no unique way of going from general to specific, the algorithm searches simpler models using different search paths. It does that by removing variables with insignificant coefficients. When the model cannot be reduced any more, it is subjected to diagnostic tests. If it passes the tests, it is called a terminal model. Since there are many search paths, there will in general be several terminal models as well.

After reaching this stage, Autometrics forms the union of the terminal models and tests the terminal models against it. The union of the models that pass the tests form a new GUM. The general-to-specific testing procedure is then repeated and a new set of terminal models obtained. If all models in this set are rejected against the new union model, the union will be the final model. Otherwise, modelling restarts with yet another GUM and continues until a final model has been reached.

In our case, the number of variables exceeds the number of observations. We follow Hendry and Krolzig (2005) and divide the variables into subsets, each of which contains fewer variables than observations. This implies that at the outset there exists more than one GUM. Each of these GUMs now forms a starting-point for Autometrics and the algorithm yields a set of terminal models for each GUM. The terminal models derived from all subsets of variables or all GUMs are merged to form a single union model. If the number of variables in this model is less than the number of observations, which happens in our application, model selection proceeds from this union model as described above.

Autometrics is partly a black box. The user can, however, affect the outcomes by selecting a number of settings, such as the significance level of the tests the algorithm relies on.

3.2. **Marginal Bridge estimator.** MBE is designed for situations often occurring in statistical and genomic applications in which there is a large number of candidate variables but only a small subset of these may belong to the model. Following Huang et al. (2008), consider first the Bridge estimator (BE). This is a shrinkage estimator for a linear regression model

$$y_i = \alpha + \beta' \mathbf{x}_i + \varepsilon_i, \ i = 1, ..., n \tag{3}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip_n})'$ is a $p_n \times 1$ observation vector (p_n may increase in n but $p_n < n$,) and $\alpha = 0$ without loss of generality. Furthermore, $\varepsilon_i \sim \text{iid}(0, \sigma^2)$. BE estimates β by minimizing

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - \beta' \mathbf{x}_i\right)^2 + \lambda_n \sum_{k=1}^{p_n} |\beta_k|^\gamma \tag{4}$$

where $\gamma > 0$ and $\lambda_n > 0$ determines the size of the penalty. Let the true parameter vector be $\beta_0 = (\beta_{10}', \beta_{20}')'$ with β_{10} having no zero entries, $\beta_{20} = \mathbf{0}$, and let $\hat{\beta}_n = (\hat{\beta}_{1n}', \hat{\beta}_{2n}')'$ be the corresponding estimator from (4). BE minimizes the OLS objective function plus a penalty for parameters different from zero. Hence, it shrinks esimates towards zero . Huang et al. (2008)

showed that under regularity conditions parameters are i) estimated consistently ($\hat{\beta}_n \to \beta_n$ in probability), ii) the truly zero parameters are set to zero $\left(P(\hat{\beta}_{2n} = 0) \to 1\right)$ and iii) the asymptotic distribution of the estimators of nonzero parameters is the same as if only these had been included in the model. This means that the parameters of the nonzero coefficients are estimated (asymptotically) as efficiently as if only the relevant variables had been included in the model from the outset.

For BE to possess this property one needs $p_n < n$. When this condition no longer holds, MBE is applicable. The idea is to run a series of 'mini' or 'marginal' regressions, with a penalty on parameters that differ from zero. The function to be minimized equals

$$Q_n(\beta) = \sum_{k=1}^{p_n} \sum_{i=1}^{n} (y_i - \beta_k x_{ik})^2 + \lambda_n \sum_{k=1}^{p_n} |\beta_k|^\gamma \tag{5}$$

Let $\widetilde{\beta}_n = (\widetilde{\beta}'_{1n}, \widetilde{\beta}'_{2n})'$ be the estimator of β_0 from (5). Under regularity conditions and $0 < \gamma < 1$, (a) the estimator $\widetilde{\beta}_{2n} = \mathbf{0}$ with probability converging to one, and (b) $P(\widetilde{\beta}_{1nk} \neq 0, \widetilde{\beta}_{1nk} \in \widetilde{\beta}_{1n}) \to 1$, as $n \to \infty$. Property (a) is similar to ii) for the BE. According to (b), the elements of $\widetilde{\beta}_{1n}$ converge to nonzero values. Thus, (a) and (b) jointly can be expected to efficiently separate the relevant variables from the rest.

Of the conditions underlying the above result the so-called partial orthogonality condition is problematic in a time series context. It states that the correlation between the relevant and irrelevant variables is not allowed to be too high. This condition can be violated if the explanatory variables are lags and functions of lags of the dependent variable as in our case. However, as we shall see in Section 5, MBE works quite well even in our context.

3.3. **QuickNet.** QuickNet (QN) resembles an earlier modelling device called RETINA, see Perez-Amaral et al. (2003). The idea of RETINA is to find the explanatory variables that in absolute terms are most strongly correlated with y_t. The most correlated variable is selected first, and the following ones one by one thereafter. QuickNet differs from RETINA in that the set of candidate variables is different, as is the model selection criterion used for final selection. QuickNet works as follows. First, the set of candidate variables S, see (2), is constructed. The variables have to be such that they show sufficient variation in the sample and are not perfectly linearly correlated; see White (2006) for details. This set of candidate variables is also used when Autometrics and MBE are applied. Once this has been done, a predetermined number of variables, $\overline{q}$, are added to the model from the set S, according to the rule that selects the variable with the strongest (positive or negative) correlation with the residuals of the previously estimated model. Then a model selection criterion is applied to choose a subset of the $\overline{q}$ variables. We used 10-fold cross validation as suggested by Hastie et al. (2009).

We also experiment with a simplified unidirectional version of this method. The variables are selected one at a time as before, but the significance of the the added variable is tested at each step. Parsimony is appreciated, so the significance level of the tests is decreased as the number of variables the

model increases. Adding variables is terminated at the first non-rejection of the null hypothesis, so this is a pure specific-to-general strategy. In the empirical section, we apply this method such that the significance level of the first test in the sequence equals 0.2. Beginning with this value, the significance level is then halved at each step. In reporting results in Section 5, this method is called QN-SG. To compare the forecasts of the neural network models to genuinely nonparametric ones, direct Nadaraya-Watson kernel regression forecasts (NP) are generated. Finally, no change forecasts (NC), which forecast that the variable of interest takes the same value at any future point in time as it does at the time of forecasting, are computed and compared with the others.

4. Forecasting

4.1. **Two ways of generating multiperiod forecasts.** There are two main ways of creating multiperiod forecasts. One can either generate the forecasts recursively, or one may apply direct forecasting. In the former case, one and the same model is used for all forecast horizons. Direct forecasting implies that a separate model is built for each forecast horizon. In the empirical section of the paper we shall compare results from these two approaches. A brief discussion of these two techniques follows next.

4.1.1. *Recursive forecasts.* In order to illuminate recursive forecasting, consider the model (1) with $p = q = 1$. These restrictions are for notational simplicity only. Assuming the information set $\mathcal{F}_{T-1} = \{y_{T-j}, j \geq 1\}$ is independent of future error terms, the one-period-ahead forecast made at time T equals

$$y_{T+1|T} = E\left(y_{T+1}|\mathcal{F}_T\right) = \beta_{00} + \beta_{01}y_T + \beta_1(1 + \exp\{\gamma_0 + \gamma_1 y_T\})^{-1}.$$

The corresponding conditional mean $y_{T+2|T}$, that is, the two-period forecast, becomes

$$\begin{aligned} y_{T+2|T} &= E\left(\beta_{00} + \beta_{01}y_{T+1} + \beta_1(1 + \exp(\gamma_0 + \gamma_1 y_{T+1}))^{-1} + \varepsilon_{T+2}|\mathcal{F}_T\right) \\ &= \beta_{00} + \beta_{01}y_{T+1|T} + \beta_1 E\left(1 + \exp(\gamma_0 + \gamma_1(y_{T+1|T} + \varepsilon_{T+1}))^{-1}|\mathcal{F}_T\right) \\ &= \beta_{00} + \beta_{01}y_{T+1|T} + \beta_1 \int_{-\infty}^{\infty} (1 + \exp(\gamma_0 + \gamma_1(y_{T+1|T} + z)))^{-1}\phi(z)dz \end{aligned} \tag{6}$$

where $\phi(z)$ is the density of the $\mathcal{N}(0, \sigma^2)$ random variable. The integral in (6) can be computed by numerical integration. Note that it becomes a multiple integral when the forecast horizon $h > 2$. It is therefore better to calculate its value by simulation or by bootstrapping the residuals of the model, because this remains a computationally feasible method even when $h > 2$. Some authors bypass this complication altogether by setting $\varepsilon_{T+1} = 0$ in the logistic function, and as a result their forecasts are biased estimates of the conditional mean.

In this work we apply the bootstrap. It has the advantage over simulation that unconditional heteroskedasticity of unknown form is allowed in the error process. More discussion about recursive forecasting can be found in Teräsvirta (2006), Kock and Teräsvirta (2011) or Teräsvirta et al. (2010, Chapter 14) among others.

4.1.2. *Direct forecasts.* In direct forecasting, the conditional mean estimate arises from a different model for each time horizon. Given the information set $\mathcal{F}_T$, the forecast for $T+h$ made at T equals

$$y^{\mathrm{D}}_{T+h|T} = g_h(y_T, y_{T-1}, ..., y_{T-p+1})$$

where g_h is a function of y_T and its lags. In our case, model selection is made using the three aforementioned techniques, but there is a 'gap' in the model in that $y_{T+h-1}, ..., y_{T+1}$ do not enter the equation. The advantage of the direct method lies in its computational simplicity: no recursions are needed. But then, a separate model has to be specified for each forecast horizon.

4.1.3. *Forecasts based on differences and forecast errors.* The forecasts based on differences are obtained in the following way. When forecasting recursively first differences $\Delta y_t = y_t - y_{t-1}$ are being modelled and forecast. The p lags of the left hand side variable are thus $\Delta y_{t-1}, ..., \Delta y_{t-p}$ To get an h-periods-ahead forecast, which is of y_{T+h}, the first-difference forecasts have to be cumulated[1]:

$$E\left(y_{T+h}|\mathcal{F}_T\right) = \sum_{j=1}^{h} E\left(\Delta y_{T+j}|\mathcal{F}_T\right) + y_T. \tag{7}$$

The corresponding forecast error is $e_{T+h|T} = y_{T+h} - E\left(y_{T+h}|\mathcal{F}_T\right)$.

In direct h-periods-ahead forecasting, the variable to be modeled is $\Delta_h y_t = y_t - y_{t-h}$. The p lags of the left-hand side variable are thus $\Delta_h y_{t-h}, ..., \Delta_h y_{t-h-p+1}$ and the corresponding forecast of y_{T+h} is $E\left(\Delta_h y_{T+h}|\mathcal{F}_T\right) + y_T$. The estimated model yields direct estimates of the conditional mean.

The measure of performance in this work is the root mean square forecast error (RMSFE). It is calculated for each time series from out-of-sample forecasts for the forecasting period beginning at T_0 and ending at $T - h_{\max}$, where T is the last available observation and h_{max} is the maximum forecast horizon. Thus,

$$\mathrm{RMSFE}_h = \{(T - h_{max} - T_0 + 1)^{-1} \sum_{t=T_0}^{T-h_{\max}} e^2_{t+h|t}\}^{1/2}.$$

4.2. **Insanity Filters.** Nonlinear models may sometimes generate forecasts that are deemed unrealistic in the light of the hitherto observed values of the time series. This has prompted forecasters to introduce precautions in order to avoid excessive forecast errors. The idea is to replace an unrealistic forecast with a more conventional and believable one. It has been applied, among others, by Swanson and White (1995, 1997a,b) who call the procedure the insanity filter, Stock and Watson (1999) and Teräsvirta et al. (2005). We shall make use of two insanity filters. The first one works as follows: If the h-step ahead predicted change exceeds the maximum h-step change observed during the estimation period, the most recently observed value of the variable to be predicted is the forecast. Hence, in the words of Swanson and White (1995) we "replace craziness by ignorance". We shall call this filter the Swanson and White (SW) filter. In the second filter,

[1]The unknown $E\left(\Delta y_{T+j}|\mathcal{F}_T\right)$ are of course replaced by their bootstrapped counterparts.

the extreme predicted change is replaced by a forecast from our benchmark linear autoregressive model: craziness is replaced by linearity.

5. Results

The above techniques are applied to the monthly Consumer Price Index (CPI) and unemployment series for the G7 countries as well as the four Scandinavian countries. Before considering these macroeconomic series a small Monte Carlo experiment is conducted. As mentioned in the introduction, the purpose of this exercise is to see how the three modelling procedures perform under controlled circumstances when the data generating process is known and contained in the linear span of S and thus is possible to select.

5.1. **General methodology and data.** The technique for generating the potential hidden units for the ANN model (1) is described in the Appendix. We have modified the original White (2006) technique somewhat to make it more suitable to our modelling problem. For QuickNet and MBE we used 10-fold cross validation as in Hastie et al. (2009) to determine the number of hidden units to be included. We also used the hv-Cross Validation procedure of Racine (2000) but this did not improve the results, so they are omitted. Following the suggestion of White (2006), the maximum number of variables in the ANN models was set to ten.

The macroeconomic series are obtained from the OECD Main Economic Indicators. Most series begin in the 1960s and end in December 2009 or January 2010. The CPI series were transformed to logarithms before modelling them, and the forecast errors discussed in the paper are errors in forecasting the transformed series.

5.2. **Monte Carlo.** For our simulation study we chose a strongly nonlinear model from Medeiros et al. (2006). These authors took the well-known annual Wolf's sunspot number series and, after transforming the observations using the Box-Cox transformation as in Ghaddar and Tong (1981), fitted an ANN model (1) with two hidden units to the transformed series. The model is:

$$\begin{aligned} y_t = &- 0.17 + 0.85y_{t-1} + 0.14y_{t-2} - 0.31y_{t-3} + 0.08y_{t-7} \\ &+ 12.8G_1(\mathbf{y_{t-1}}) + 2.44G_2(\mathbf{y_{t-1}}) + \epsilon_t \end{aligned} \tag{8}$$

where the two hidden units are

$$G_1(\mathbf{y_{t-1}}) = \left(1 + \exp\left(-0.46\left(0.29y_{t-1} - 0.87y_{t-2} + .40y_{t-7} - 6.68\right)\right)\right)^{-1}$$

and

$$G_2(\mathbf{y_{t-1}}) = \left(1 + \exp\left(-1.17 \times 10^3\left(0.83y_{t-1} - 0.53y_{t-2} - 0.18y_{t-7} + 0.38\right)\right)\right)^{-1}$$

and $\epsilon_t \sim$i.i.d.N(0,1). We generate 500 time series of 600 observations from this model. The set of potential variables consists of G_1, G_2, 1000 other hidden units, and ten lags of y_t. The number of variables thus greatly exceeds

Recursive	Hor.	DGP	AR	QN	MBE	Autom.	QN-SG
	1	1.82	1.456	1.343	1.730	1.105	1.805
NF	2	2.739	1.536	3.282	1.659	1.073	1.568
	5	4.172	1.337	$9 \cdot 10^4$	1.394	4023	1.283
	1	1	1.456	1.513	1.730	1.105	1.855
SW	2	1.001	1.536	1.532	1.658	1.074	1.552
	5	1.001	1.392	1.218	1.395	1.028	1.269
	1	1	1.456	1.322	1.730	1.105	1.776
AR	2	1.001	1.536	1.366	1.658	1.074	1.552
	5	1.001	1.337	1.214	1.395	1.028	1.269

TABLE 1. Average root mean square forecast error ratios for the recursive forecasts of the simulated sunspot series. DGP: Data generating process, AR: Autoregression, QN: QuickNet, MBE: Marginal Bridge Estimator, Autom.: Autometrics, QN-SG: Quick-Net specific to general. NF: No Filter (for the DGP the NF subcolumn contains the actual root mean square forecast error from forecasting with the DGP), SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1.456	1.343	1.730	1.105	1.805	1.546	3.560
NF	2	1.518	9.575	1.549	1.652	1.436	1.332	4.226
	5	1.306	1.353	1.241	1.359	1.293	1.124	3.984
	1	1.456	1.513	1.730	1.105	1.855	1.658	
SW	2	1.518	1.52	1.549	1.532	1.733	1.424	
	5	1.363	1.326	1.241	1.322	1.293	1.124	
	1	1.456	1.322	1.730	1.105	1.776	1.555	
AR	2	1.518	1.35	1.549	1.355	1.444	1.335	
	5	1.306	1.219	1.241	1.246	1.293	1.124	

TABLE 2. Average root mean square forecast error ratios for the direct forecasts of the simulated sunspot series. NP: Non-parametric, NC: No Change forecasts. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

the number of observations. The forecast horizons are one, two, and five years, and the maximum number of variables per each selected model equals ten. We report RMSFE ratios such that the denominator is the RMSFE of forecasts from (8), computed from the 500 replications.

Table 1 contains these ratios for the recursive forecasts. The first three entries in the column named DGP contain the RMSFE for the forecasts from the true model (8). As expected, all RMSFE ratios exceed unity. Autometrics-selected models generate by far the most accurate forecasts of the alternatives to the DGP, indicating that the method works well when there is a true model that can be selected from the set of variables available for the purpose. The other methods lead to models whose forecasts are of more or less the same quality. The forecasts from MBE-selected models do not need filtering but are nevetheless slightly more inaccurate than the other (filtered) ones.

The performance of direct models is reported in Table 2. Models selected by Autometrics no longer generate more accurate forecasts than the other nonlinear models. Every possible direct model is misspecified by definition

Rec	Hor.	DGP	AR	QN	MBE	Autom.	QN-SG
	1	3.35	4.53	4.03	4.92	3.68	4.6
AR	2	5.18	7.28	6.13	7.68	5.77	7.01
	5	5.5	7.28	6.24	7.16	5.49	6.6

Dir	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	4.53	4.03	4.92	3.68	4.6	4.5	6.23
AR	2	7.19	6.65	7.54	6.2	6.93	6.67	10.4
	5	7.1	6.6	6.79	6.47	7.07	6.15	10.4

TABLE 3. Average ranks based on the absolute forecasts errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the DGP forecasts and the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

because the shortest lag (two-year model) or lags (five-year model) of y_t cannot be used, and Autometrics clearly suffers from this. Note the good performance of the nonparametric model forecasting five years ahead. The kernel autoregression seems to make most of the available information, and the forecasts hardly need filtering. In fact, the SW filter has a negative effect on the accuracy of the forecasts from this model. As may be expected, the No Change forecast does not perform well in predicting these strongly cyclical realisations.

We also compare the methods by calculating the average ranks of the absolute forecast errors. Only the results for the AR filtered forecasts are reported since the ranks obtained from the SW filtered ones are similar.

The ranks can be found in Table 3. As can be expected from the RMSFE results, the forecasts from the DGP have the lowest ranks. However, the ranks of the recursive forecasts by Autometrics are not much higher and even as low as the DGP ones at the five year horizon. Of the remaining neural network procedures MBE forecasts have the highest ranks while the No Change forecasts are by far the least accurate overall. This is not surprising due to the cyclical nature of the series to be forecast. The nonparametric forecasts perform about as well as the ANN-based procedures at the shortest horizons and better than them at the five year horizon.

Another robust way of considering the results is to use Wilcoxon's signed-rank test (Wilcoxon (1945)) for comparing forecasts from the DGP with the others. The null hyothesis is that the absolute forecast error of the DGP and that of the other model have the same mean whereas the alternative is that the alternative model has a lower mean absolute forecast error. The tests are carried out separately for each horizon. The results are reported in Table 4. A normal approximation has been used in calculating the p-values. This is appropriate due to the large number of forecasts (500). Small p-values indicate that the alternative model produces more accurate forecasts than the DGP. If the alternative hypothesis is that the forecasts from the DGP have the lowest mean, one simply subtracts the reported p-values from one

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1	1	1	1
AR	3	1	1	1	1	1
	5	1	1	1	0.852	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1	1	1	1	1	1
AR	3	1	1	1	1	1	1	1
	5	1	1	1	1	1	1	1

TABLE 4. p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from the DGP being equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Recursive forecasts. Bottom panel: Direct forecasts.

Recursive	Total	Linear	Nonlinear	DGP units
QN	9.55	0.348	9.21	1.64
MBE	9.22	0.756	8.47	0.77
Autom	11	1.5	9.51	3.47
QN-GS	5.3	0.324	4.98	1.12

TABLE 5. Average number of variables selected for the recursive forecasts of the CPI based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, "Nonlinear" gives the number of hidden units included, and DGP units gives the number of units included from the data generating process.

and obtains the p-values of this test. All tests are based on the AR-filtered forecasts[2].

As can be seen from Table 4, the results in Tables 1 and 2 accord with those from the Wilcoxon test. It is not possible to reject the hypothesis that the absolute forecast errors of the DGP forecasts and those from the alternative model have the same mean if the alternative hypothesis is that the alternative model has a lower mean. If the alternative hypothesis is that the DGP forecast errors have a lower mean, the null of equal means is rejected with a single exception: the recursive five-year forecasts from the Autometrics-selected ANN model.

Table 5 offers some background to the results in Tables 1 and 2. It contains information about the size and variable types in the nonlinear models for recursive forecasting. The average number of variables in every type of model is larger than the size of (8) which is six variables as the intercept is not counted. It is worth noting that Autometrics, while selecting the largest models, picks up elements of the true model more frequently than the other model selection techniques. This is probably the most important factor in explaining its success in forecasting. Moreover, Autometrics on

[2]Alternatively, one could consider the Giacomini-White test (Giacomini and White (2006)) which includes the Diebold-Mariano test (Diebold and Mariano (1995)) as a special case. The Giacomini-White test, however, relies on a rolling window. The Giacomini-White test was also carried out but most often the conclusions were the same as for the Wilcoxon test and so the results are not reported here.

average chooses more linear lags than the other models, although fewer than their number in the true model. The average number of linear lags in the other models is rather small. The specific-to-general QN-SG is clearly more parsimonious than QuickNet, but this result is not invariant to the choice of significance levels in the test sequence. QuickNet-based recursive forecasts are somewhat more accurate than QN-SG ones at one-and five-year horizons.

5.3. **Macroeconomic forecasts.** The CPI and unemployment series are forecast at the 1, 3, 6, and 12-month horizons. The CPI series are transformed into logarithms, and 240 forecasts based on an expanding window are generated for each horizon[3]. Forecasts from models of differenced series are formed as described in Section 4.1.3. The pool of variables contains 600 hidden units with $p = 6$ in (1) and the first six linear lags of the dependent variable.

The models are respecified every six months. This is because of Autometrics is quite slow: otherwise respecification could easily be done every month. Pretesting linearity and letting the nonlinear model selection operate only if the linearity hypothesis was rejected did not on average improve the performance of the nonlinear models. This may be due the fact that linear lags are included into the pool of hidden units which makes it possible to select a linear model anyway.

5.4. **Consumer Price Index.** The RMSFE ratios for recursive CPI forecasts from models of differenced series can be found in Table 6. The denominator in the RMSFE ratio is now the RMSFE of the recursive linear AR forecasts. It is seen that filtering the forecasts is necessary. All four model selection techniques lead to ANN models that generate some very inaccurate forecasts. This is the case already for one-month forecasts and is due to the fact that some models contain very strongly correlated variables. A pair of them typically has large (in absolute value) coefficients with opposite signs. Forecasting with such a model yields inaccurate forecasts and cumulating them in forecasting more than one month ahead makes the situation even worse. This is clearly seen from the table. Furthermore, all ratios exceed one, which means that on average no ANN model, not even after filtering, generates more accurate recursive forecasts than the linear AR model. Models selected by MBE perform slightly better than the other nonlinear models.

These results may be compared with the ones in Table 7. This table contains the RMSFE ratios for direct forecasts from models built using differenced series. Models built using QuickNet and Autometrics still generate a few forecasts that require filtering, whereas MBE-based forecasts do not. After filtering the six- and 12-month forecasts from the ANN models are more accurate than the benchmark ones. This is also the case for forecasts from direct linear AR models. Their RMSFE ratios are comparable to those obtained from models built by MBE which is the best-performing model selection technique. The forecasting performance of the nonparametric model

[3]For some of the shorter data sets the number of forecasts is less than 240, because the first window was set to include at least 200 observations.

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	16.82	1.02	257.9	1.043
NF	3	1	$5 \cdot 10^4$	$2 \cdot 10^6$	$2 \cdot 10^9$	1.052
	6	1	$4 \cdot 10^5$	$1 \cdot 10^6$	$6 \cdot 10^9$	2.411
	12	1	$1 \cdot 10^6$	$1 \cdot 10^6$	$1 \cdot 10^{10}$	$3 \cdot 10^5$
	1	1	1.040	1.020	1.074	1.047
SW	3	1.004	1.033	1.020	1.075	1.061
	6	1.003	1.055	1.020	1.085	1.076
	12	1.011	1.107	1.034	1.172	1.091
	1	1	1.042	1.019	1.072	1.044
AR	3	1	1.025	1.014	1.058	1.052
	6	1	1.036	1.017	1.047	1.071
	12	1	1.066	1.032	1.105	1.088

TABLE 6. Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	16.82	1.02	257.9	1.043	1.148	1.133
NF	3	0.976	2.699	0.9893	2464	1.02	1.074	1.169
	6	0.8123	20.77	0.8239	1869	0.8362	0.9335	1.159
	12	0.7336	3.286	0.7284	20.08	0.7436	0.8203	1.134
	1	1	1.040	1.020	1.074	1.047	1.150	
SW	3	0.976	1.039	0.9893	1.059	1.030	1.081	
	6	0.8123	0.8452	0.8239	0.8987	0.836	0.9335	
	12	0.7336	0.7584	0.7284	0.8355	0.7397	0.8203	
	1	1	1.042	1.019	1.072	1.044	1.147	
AR	3	0.976	1.020	0.9893	1.042	1.019	1.075	
	6	0.8123	0.840	0.8239	0.8819	0.835	0.9335	
	12	0.7336	0.7591	0.7284	0.8371	0.7395	0.8203	

TABLE 7. Average root mean square forecast error ratios for the direct forecasts of the CPI series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

is below average, and the 'no change' forecasts are less accurate than even the corresponding recursive ones.

The RMSFE ratios in Table 8 refer to recursive forecasts from models built on CPI levels. Filtered forecasts are more accurate on average than the corresponding forecasts in Table 6. MBE-based forecasts are the most accurate ones and models built by QN-SG generate the least accurate recursive forecasts: all ratios remain above one. Recursive linear AR models built on levels are somewhat superior to ones built on differences. The RMSFE ratios lie below one for the two longest horizons but are greater than the corresponding ratios for forecasts from models obtained by MBE, QuickNet and Autometrics.

Table 9 contains the RMSFE ratios for direct forecasts from models specified and estimated from the level series. It appears that MBE is the best model-building method when the criterion is the RMSFE. The ratios are even smaller than the ones found in Tables 6–8. Direct models selected by QuickNet also perform better than the recursive ones, whereas the same

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
NF	1	1.011	1.013	0.977	1.062	1.139
	3	1.001	20.39	0.9315	6311	1.195
	6	0.9728	$3 \cdot 10^5$	0.8535	$1 \cdot 10^7$	1.223
	12	0.9372	$3 \cdot 10^6$	0.787	$3 \cdot 10^8$	1.309
SW	1	1.011	1.013	0.977	1.062	1.139
	3	1.001	0.9685	0.9314	1.003	1.184
	6	0.9728	0.896	0.8532	0.9299	1.187
	12	0.9372	0.823	0.7871	0.8489	1.185
AR	1	1.011	1.013	0.977	1.062	1.139
	3	1.001	0.9661	0.9314	1.003	1.181
	6	0.9728	0.8923	0.8532	0.9299	1.187
	12	0.9372	0.8143	0.7871	0.8489	1.167

TABLE 8. Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
NF	1	1.011	1.013	0.977	1.062	1.139	16.77	1.133
	3	0.9661	0.9418	0.9057	0.9761	1.198	8.037	1.169
	6	0.9053	3.401	0.8114	0.9982	1.204	5.072	1.159
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	1.134
SW	1	1.011	1.013	0.977	1.062	1.139	3.783	
	3	0.9661	0.9418	0.9057	0.9761	1.198	5.172	
	6	0.9053	0.8305	0.8114	0.954	1.204	4.907	
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	
AR	1	1.011	1.013	0.977	1.062	1.139	3.675	
	3	0.9661	0.9418	0.9057	0.9761	1.198	5.136	
	6	0.9053	0.8303	0.8114	0.9564	1.204	4.904	
	12	0.7771	0.7205	0.6928	0.9416	1.173	3.119	

TABLE 9. Average root mean square forecast error ratios for the direct forecasts of the CPI series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

cannot be said of models based on Autometrics or QN-SG. In the light of these results, going from specific to general and back again (QuickNet) is a better idea than going from specific to general only (QN-SG), but this finding cannot be generalized. It may be noted that the nonparametric model built on levels generates much less accurate forecasts than the same model estimated from differenced series. Its RMSFE ratios are remarkably larger than any other ratio. Summing up, it seems that direct forecasts are on average more accurate than the recursive ones. Exceptions do exist: compare Autometrics-based six- and 12-month RMSFE ratios in Tables 8 and 9. It should be pointed out that these results are general ones and do not necessarily hold for all 11 countries.

As was the case for the simulation study we also compare the forecast performance of the methods applied by considering their ranks. This is done for all countries and forecast horizons. Furthermore, forecasts from models built on differences and the ones based on levels are included in the same comparison.

Rec Diff	Hor.	AR	QN	MBE	Autom.	QN-SG
AR	1	6.53	6.62	6.57	6.89	6.71
	3	11.8	11.9	11.7	12	12.4
	6	12.9	12.8	12.7	12.8	13.5
	12	14.5	14.7	14.4	15.1	15.1

Dir Diff	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1	6.53	6.62	6.57	6.89	6.71	8.2	7.38
	3	11	11.2	11.2	11.1	11.3	13.8	13.6
	6	8.95	9.25	9.05	9.85	9.37	12.5	14.2
	12	8.9	8.66	8.65	9.44	8.52	11.4	14.4

Rec Level	Hor.	AR	QN	MBE	Autom.	QN-SG
AR	1	6.62	6.32	6.21	6.49	7.68
	3	11.7	10.6	10.7	10.8	14.1
	6	12.1	10.8	10.7	10.5	14.9
	12	12.3	10.2	10.5	9.89	14.9

Dir Level	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1	6.62	6.32	6.21	6.49	7.68	7.97	7.38
	3	10.5	10.3	10.1	10.7	14.3	17.7	13.6
	6	10.5	9.95	9.73	10.8	15	22	14.2
	12	8.91	9.32	8.94	9.89	14.5	21.6	14.4

TABLE 10. Average ranks based on the absolute forecasts errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

The results can be found in Table 10. At the 1-month horizon absolute forecast errors from the ANN procedures have ranks very close to each other, which is in accordance with the findings from Tables 6-9. The nonparametric and No Change forecasts have considerably higher ranks than the other procedures. This is true for the forecasts based on differences as well as the ones based on levels. In particular the high ranks for the nonparametric forecasts are no surprise in the light of the high relative RMSFE in Table 9. In general, the direct methods have the lowest average ranks. This is the case in particular for the forecasts based on the differences of the series. The overall winner at the 12-month horizon is the QN-SG method based on differences while the MBE-based forecasts come in second. This again agrees with the results reported in Tables 6-9.

Similarly to the simulated example, we conduct Wilcoxon's signed-rank test for pairs of absolute forecast error series. The benchmark, the recursive linear AR forecast, is always one of the forecasts in the pair. As already discussed, the null hypothesis of the test is that the means of the absolute forecast errors are equal, and the alternative is that the absolute forecast errors of the 'other model' have the smaller mean of the two. The upper panel of Table 11 contains p-values of the test for recursive forecasts from differenced models. Most of them are close to one, which means that the null hypothesis is rejected in the opposite direction. This accords with the information in Table 6, where all RMSFE ratios were greater than one.

Recursive	Hor.	QN	MBE	Autom.	QN-SG
AR	1	0.998	0.778	1	1
	3	0.963	0.0582	0.982	1
	6	0.949	0.266	0.995	1
	12	1	1	1	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1		0.998	0.778	1	1	1	1
	3	0	0.011	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	0.005	1	1
	6	0	0	0	0	0	0	1
	12	0	0	0	0	0	0	0.968

TABLE 11. p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the CPI series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on differences. Bottom panel: Models estimated directly on differences.

The MBE-based forecasts at horizons up to six months constitute the only exception.

The lower panel contains the p-values for direct forecasts. They are mostly close to zero for long forecasting horizons. The no change forecast is the only exception: all p-values are large. Direct forecasts can thus be deemed superior to recursive ones when the models are built on differenced CPI-series. This strengthens conclusions that emerge from Tables 4 and 5.

Table 12 contains p-values of the same test and null hypothesis when the forecasts are obtained using models built on CPI series in levels. Results in the upper panel show that the null hypothesis is rejected in favour of the recursive linear AR forecasts when compared to the model selected by QN-SG. The other methods generate ANN models that yield more accurate recursive forecasts than the linear AR model (p-values are close to zero) or forecasts for which the null hypothesis is not rejected (QuickNet and Autometrics one-month forecasts). The lower panel shows that QN-SG-based direct models do not perform well either. The same can be said about the nonparametric model and the 'no change' forecasts. Considering all four horizons at once, MBE emerges as the best-performing model selection criterion for direct models when Wilcoxon's test is used as the yardstick.

As in the simulated example, it is interesting to see whether the size of the model and the accuracy of the forecasts from it have to do with each other. Table 13 contains information about the size and composition of models based on differenced series. When forecasting recursively, it is seen from the left panel that QN-SG selects the most parsimonious models which do not, however, yield the most accurate forecasts. MBE selects somewhat less parsimonious models that on average yield the most accurate recursive forecasts. It also chooses the largest fraction of linear lags, although their average number remains below one. Models selected by Autometrics are by far the largest ones. There does not seem to be a clear connection between the model size and forecast accuracy.

Recursive	Hor.	QN	MBE	Autom.	QN-SG
AR	1	0.160	$1 \cdot 10^{-4}$	0.171	1
	3	$1 \cdot 10^{-8}$	0	$4 \cdot 10^{-6}$	1
	6	0	0	0	1
	12	0	0	0	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1	1	0.160	$1 \cdot 10^{-4}$	0.171	1	1	1
	3	0	0	0	$4 \cdot 10^{-8}$	1	1	1
	6	0	0	0	0	1	1	1
	12	0	0	0	0	1	1	0.968

TABLE 12. p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the CPI series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on levels. Bottom panel: Models estimated directly on levels.

Recursive	Total	Linear	Nonlinear
QN	6.35	0.298	6.05
MBE	5.51	0.818	4.69
AM	15.5	0.393	15.1
QN-SG	4.03	0.195	3.83

Direct MBE	Total	Linear	Nonlinear
1 mth	5.51	0.818	4.69
3 mths	5.48	2.45	3.03
6 mths	5.29	3.55	1.74
12 mths	2.69	1.72	0.964

TABLE 13. Left panel: Average number of variables selected for the models generating recursive forecasts of the CPI based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on differences by MBE.

Recursive	Total	Linear	Nonlinear
QN	5.35	1.09	4.27
MBE	7.19	5.64	1.55
AM	19.1	1.34	17.7
QN-SG	1.39	1	0.386

Direct MBE	Total	Linear	Nonlinear
1 mth	7.19	5.64	1.55
3 mths	7.24	5.74	1.49
6 mths	7.42	6	1.42
12 mths	7.21	6	1.21

TABLE 14. Left panel: Average number of variables selected for the models generating recursive forecasts of the CPI based on levels. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on levels by MBE.

The right-hand panel of Table 13 contains the average size and composition of models based on differenced series and selected by MBE for direct forecasting. The average number of variables is halved when one moves from six- to 12-month models, whereas the share of linear lags of the total increases up to six-month models and remains about the same for 12-month ones.

Table 14 contains the same information for models built on levels. All methods now select more linear variables than in the previous case. QN-SG is still the most parsimonious technique, and even QuickNet selects fewer

variables than MBE. As Tables 9 and 12 indicate, forecasts from MBE are still the most accurate ones on average. The use of Autometrics leads to largest models. They perform better than QN-SG-selected models but less well than ones specified using MBE. The right panel of the table shows that MBE select a large number of linear lags for all direct models. In fact, every MBE-model built for the two longest horizons contains all six lags and only a small number of hidden units. A comparison of the RMSFE ratios in Tables 6 and 8 on the one hand and Tables 7 and 9 on the other (indirectly) suggests that direct models based on level data and selected by MBE may be slightly superior to the same type of model, selected by the same technique, but based on differenced series. Whether or not this is due to the larger amount of linear lags in the former models is not clear, however.

5.4.1. *Individual countries.* To shed light on some of the cross-country variation in the results that cannot be seen in the summary tables we now consider results for some individual countries, Italy, Japan, and the US. They are selected because there are interesting differences between them. The remaining country-specific RMSFE are available at `http://econ.au.dk/research/research-centres/creates/research/research-papers/supplementary-downloads/rp-2011-28/`.

Tables 15 and 16 show the RMSFE ratios for the US CPI forecasts based on differences (only the results for the AR-filter are presented). It is seen that it is indeed possible to improve upon the linear AR model even when forecasting recursively, although this is not true for all three methods. In fact, only MBE outperforms the linear autoregression at all horizons, which again indicates it may be superior to QuickNet and Autometrics in forecasting the CPI series. The Wilcoxon tests were also carried out on the individual countries. Based on these, the above findings are significant since at no horizon does one observe a higher p-value than 0.044 when testing the AR forecasts against the recursive MBE ones.

On average MBE selects seven variables of which 3.38 are linear lags. It is more parsimonious and includes a higher number of linear lags than the other procedures.

A comparison of Tables 15 and 16 shows that for the US the the recursive forecasts are less accurate than the corresponding direct ones. The differences in the RMSFE are, however, less pronounced than in Tables 6 and 7. The finding that the direct forecasts are more accurate than the recursive ones is uniform across all countries. All ANN models, independent of the variable selection procedure, work well in direct forecasting. However, for the US they are at the longest horizons outperformed by the nonparametric model and perform less well than they do in general. For the direct forecasts MBE is again the most parsimonious procedure whereas Autometrics on average selects the largest models. MBE-based models also contain the largest number of linear units.

The averaged results for the forecasts based on levels also sometimes hide differences between the individual countries. To illustrate this, Tables 17 and 18 present RMSFE ratios for Italy, Japan, and the US (only the results based on the AR-filter are shown). Table 17 shows that there can be considerable variation in the performance of the variable selection procedures.

US Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1.03	0.9803	1.044	1.011
AR	3	1	0.9811	0.9722	1.003	1.034
	6	1	0.9836	0.9388	0.9769	1.006
	12	1	1.017	0.9484	1.033	1.058

TABLE 15. Average root mean square forecast error ratios for the recursive forecasts of the US CPI series based on differences. AR: Insane forecasts replaced by linear autoregressive ones.

US Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1.03	0.9803	1.044	1.011	1.044	1.221
AR	3	0.9952	0.9742	0.9988	0.9298	0.9403	0.9417	1.179
	6	0.8312	0.8749	0.8332	0.8656	0.8825	0.8098	1.224
	12	0.9483	1.014	0.9145	0.9856	0.9501	0.8423	1.598

TABLE 16. Average root mean square forecast error ratios for the direct forecasts of the US CPI series based on differences. AR: Insane forecasts replaced by linear autoregressive ones.

MBE is the most stable procedure and the only one which has RMSFE ratios below unity for all three countries at all horizons[4], but for each country a different variable selection procedure is dominant. The relative stability of MBE is most likely due to the fact that for all three countries this procedure selects the largest number of linear units. For Italy and the US it includes all six linear units and for Japan it chooses a purely linear model every time (though not the AR(6)). Nevertheless, MBE is outperformed by Autometrics which generally chooses only a small fraction of linear lags. However, MBE includes unusually few linear units (3.6) for Japan, so it may still argued that models with a high number of linear units combined with a few nonlinear ones perform well on average.

A comparison of the results in Table 17 with the ones in Table 18 indicates that that the direct forecasts are superior to their recursive counterparts. This accords with the overall results. Moreover, the nonparametric model generates very inaccurate forecasts for these three countries, which is also in line with the general results. The direct MBE forecasts again have RMSFE ratios below unity. The performance of Autometrics varies quite remarkably. In forecasting 12 months ahead, Autometrics-based forecasts are an excellent choice for Japan, a mediocre one for the US, and are definitely not to be recommended for forecasting the Italian CPI. The situation is the same if recursive forecasts in Table 17 are considered. Autometrics-based forecasts are much better than the recursive AR ones for Japan, except at the one-month horizon, still acceptable for the US, and very inaccurate for Italy.

Results on forecasting the CPI series sugest that forecasts based on levels are superior to their counterparts based on differences. Furthermore, direct

[4]Recall, however, that the benchmark is the AR(6) model forecasted recursively based on differences of the time series. But even then, it still illustrates the rather stable performance of MBE. Its relative RMSFE are actually below one for the recursive level based forecasts for all countries at all horizons except for the UK and Denmark for which the 1-month forecasts have ratios above one.

ITA Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1.002	1.165	0.9263	1.603	1.757
AR	3	1.055	1.188	0.8965	1.629	1.955
	6	1.135	1.252	0.9073	1.677	2.067
	12	1.195	1.264	0.8355	1.715	1.855
JP Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9667	1.02	0.9885	1.069	1.021
AR	3	0.9188	0.9652	0.9612	0.8725	0.9721
	6	0.7846	0.8702	0.8419	0.6776	0.8807
	12	0.6717	0.755	0.7242	0.5016	0.7644
US Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9999	1.008	0.9477	0.9458	1.088
AR	3	0.9961	0.87	0.9152	0.9035	0.9962
	6	0.9877	0.7469	0.8597	0.8102	0.9509
	12	0.9633	0.6898	0.9324	0.8573	1.125

TABLE 17. Average root mean square forecast error ratios for the recursive forecasts of the CPI series based on levels. AR: Insane forecasts replaced by linear autoregressive ones.

ITA Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1.002	1.165	0.9263	1.603	1.757	0.9978	1.689
AR	3	0.9942	1.091	0.875	1.423	1.977	13.64	1.943
	6	0.9135	1.01	0.8358	1.428	1.949	8.529	1.942
	12	0.7215	0.8067	0.7564	1.571	1.772	4.438	1.733
JP Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9667	1.02	0.9885	1.069	1.021	4.768	0.9762
AR	3	0.883	0.925	0.9675	0.8878	0.9711	4.708	0.872
	6	0.7352	0.8424	0.7352	0.7064	0.8715	2.999	0.662
	12	0.5334	0.6727	0.5334	0.5289	0.699	1.591	0.4879
US Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9999	1.008	0.9477	0.9458	1.088	0.9979	1.221
AR	3	0.9994	0.8962	0.9212	0.9275	1.014	1.883	1.179
	6	1.007	0.7585	0.8741	0.9568	0.9696	4.698	1.224
	12	1.032	0.7579	0.9107	0.9179	1.147	4.062	1.598

TABLE 18. Average root mean square forecast error ratios for the direct forecasts of the CPI series based on levels. AR: Insane forecasts replaced by linear autoregressive ones. The NC forecasts are not filtered.

forecasting is preferable to recursive forecasts and MBE is the most stable forecasting procedure. This last observation may be attributed to the high number of linear units MBE includes and which it supplements with a few relevant nonlinear units.

5.5. **Unemployment.** A common feature of results on forecasting unemployment rate series with those on forecasting the CPI is the appearance of some vastly inaccurate forecasts and the consequent need for filtering. This is first seen from Table 19 that contains the RMSFE ratios for recursive forecasts when the models are built on differenced series. For filtered foreacsts, all ratios still lie above one. Models selected by MBE appear to lead to most

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	1	1.124	1.005	$2 \cdot 10^9$	1.045
NF	3	1	97.78	1.001	$7 \cdot 10^9$	1.054
	6	1	3333	1.003	$1 \cdot 10^{10}$	1.051
	12	1	$5 \cdot 10^4$	1.006	$1 \cdot 10^{10}$	1.026
	1	1	1.090	1.006	1.216	1.079
SW	3	1.004	1.081	1.007	1.239	1.073
	6	1.004	1.058	1.008	1.26	1.056
	12	1.001	1.026	1.008	1.221	1.026
	1	1	1.068	1.005	1.161	1.049
AR	3	1	1.07	1.002	1.152	1.058
	6	1	1.05	1.004	1.142	1.051
	12	1	1.021	1.006	1.092	1.025

TABLE 19. Average root mean square forecast error ratios for the recursive forecasts of the unemployment series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	1	1.124	1.005	$2 \cdot 10^9$	1.045	0.9999	1.109
NF	3	0.9979	59.88	1.024	$7 \cdot 10^6$	1.063	1.024	1.167
	6	1.002	1.133	1.031	$2 \cdot 10^9$	1.133	1.046	1.148
	12	1.031	250.1	1.054	$2 \cdot 10^8$	1.197	1.091	1.055
	1	1	1.090	1.006	1.216	1.079	1.013	
SW	3	1.001	1.060	1.028	1.197	1.062	1.022	
	6	1.002	1.104	1.030	1.196	1.116	1.046	
	12	1.030	1.101	1.049	1.223	1.128	1.080	
	1	1	1.068	1.005	1.161	1.049	0.9999	
AR	3	0.9979	1.053	1.025	1.184	1.058	1.023	
	6	1.002	1.101	1.028	1.195	1.113	1.042	
	12	1.031	1.109	1.047	1.215	1.138	1.082	

TABLE 20. Average root mean square forecast error ratios for the direct forecasts of the unemployment series based on differences. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

accurate nonlinear forecasts, and they do not need filtering. Autometrics-selected models are, even after filtering, the most inaccurate ones. Table 20 indicates that on average, direct forecast are not superior to recursive ones. This is true for both linear and nonlinear forecasts. Nonparametric forecasts do not require much filtering but are less accurate than the ones from the MBE-forecasts.

In the case of unemployment series, the models built on levels do not produce forecasts superior to their counterparts from models based on differences. Table 21 contains the RMSFE ratios for recursive forecasts. Again, MBE-selected models seem to generate more accurate forecasts than the others, whereas model selection using Autometrics leads to models with most inaccurate forecasts. The most striking feature of Table 22 is that the nonparametric forecasts, which need no filtering, are nevertheless on average distinctly more inaccurate than forecasts generated by any other model or

Recursive	Hor.	AR	QN	MBE	Autom.	QN-SG
	1	0.9994	$2 \cdot 10^5$	1.007	1.302	1.048
NF	3	1.010	$3 \cdot 10^5$	1.028	$1 \cdot 10^8$	1.076
	6	1.024	$5 \cdot 10^6$	1.041	$8 \cdot 10^8$	1.087
	12	1.016	$9 \cdot 10^6$	1.036	$1 \cdot 10^9$	1.065
	1	0.9994	1.064	1.007	1.148	1.048
SW	3	1.01	1.079	1.027	1.147	1.076
	6	1.018	1.106	1.036	1.135	1.087
	12	1.011	1.085	1.030	1.098	1.061
	1	0.9994	1.067	1.007	1.146	1.048
AR	3	1.010	1.080	1.027	1.150	1.076
	6	1.024	1.108	1.042	1.135	1.087
	12	1.016	1.093	1.034	1.104	1.061

TABLE 21. Average root mean square forecast error ratios for the recursive forecasts of the unemployment series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1	0.9994	$2 \cdot 10^5$	1.007	1.302	1.048	1.514	1.109
NF	3	1.006	8.759	1.033	2760	1.068	1.507	1.167
	6	1.005	1.327	1.045	1.995	1.084	1.499	1.148
	12	1.008	29.72	1.056	3.201	1.040	1.354	1.055
	1	0.9994	1.064	1.007	1.148	1.048	1.514	
SW	3	1.006	1.08	1.033	1.195	1.068	1.503	
	6	0.999	1.138	1.039	1.261	1.084	1.495	
	12	1.005	1.063	1.051	1.202	1.04	1.332	
	1	0.9994	1.067	1.007	1.146	1.048	1.514	
AR	3	1.006	1.072	1.033	1.196	1.068	1.503	
	6	1.005	1.133	1.044	1.257	1.084	1.495	
	12	1.008	1.055	1.054	1.203	1.04	1.333	

TABLE 22. Average root mean square forecast error ratios for the direct forecasts of the unemployment series based on levels. NF: No Filter, SW: Swanson-White filter, AR: Insane forecasts replaced by linear autoregressive ones.

method. It can also be noted that direct linear forecasts from linear AR models built on untransformed series have RMSFE ratios close to one, while no filtering has been necessary. The 'no change' forecasts are somewhat less accurate as the ones generated by QuickNet-selected models but better than Autometrics-ones.

Table 23 contains the average ranks of the models based on the absolute forecasts errors. In this comparison ANN forecasts from Autometrics-selected models have the highest ranks. In accordance with Tables 19-22, MBE is the best performing ANN selection method. However, none of them has a lower average rank than the linear AR model. The performance of the nonparametric forecasts is highly dependent on whether the models are built on differences or levels. In the former case the ranks are much lower than in the latter, which are by far the highest overall. This accords with the RMSFE results in Table 22.

Rec Diff	Hor.	AR	QN	MBE	Autom.	QN-SG
AR	1	6.58	6.89	6.77	7.17	6.84
	3	11.1	11.9	11.2	12.2	12
	6	10.9	11.8	11	12.2	12
	12	11.1	11.7	11.3	12.2	11.9

Dir Diff	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1	6.58	6.89	6.77	7.17	6.84	6.52	5.98
	3	11	11.9	11.6	12.4	11.7	11.6	12.6
	6	10.8	12	11.3	12.3	11.9	11.2	12.9
	12	11.4	12	11.3	12.7	12.1	12.2	12.4

Rec Level	Hor.	AR	QN	MBE	Autom.	QN-SG
AR	1	6.36	6.91	6.45	7.22	6.89
	3	11	11.9	11.3	12.7	12.2
	6	11	12.2	11.3	12.8	12.2
	12	10.8	12.1	11.2	12.7	12.1

Dir Level	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
AR	1	6.36	6.91	6.45	7.22	6.89	9.23	5.98
	3	10.8	11.7	11.2	13.1	11.8	15.5	12.6
	6	10.5	11.9	11	13.6	12	15.7	12.9
	12	10.7	11.6	11.3	13.1	11.3	15	12.4

TABLE 23. Average ranks based on the absolute forecasts errors. For each procedure for which forecasts are carried out recursively as well as directly the forecasts from the two alternatives are identical at the 1-month horizon. Hence, the comparison is only made across the direct forecasts at the 1-month horizon and by construction the ranks are the same for the recursive counterparts.

Table 24 contains p-values of Wilcoxon's signed-rank test of absolute forecast errors for AR-filtered forecasts from models based on differenced unemployment series. The null hypothesis of equal means is mostly rejected (the p-value is close to unity) in favour of the linear recursive AR absolute forecast errors having a smaller mean than the ones from the other model. MBE forecasts three months ahead are the only ANN-exception. For direct linear three- and six-month forecasts the null hypothesis of equal means is not rejected either.

Table 25 contains the same information with the difference that the alternative model is built on levels instead of differences. In this case, the null hypothesis is never rejected for MBE-based direct models, but it turns out that at longest horizons, direct linear AR forecasts have smaller absolute errors than the recursive AR ones (the corresponding p-values in Table 25 are close to zero). The result for 12-month forecasts requires an explanation. In Table 22, the 12-month RMSFE ratio of the direct linear AR forecasts equals 1.008, which does not indicate superiority of these forecasts over the recursive linear ones. Even after filtering, the direct linear 12-month model generates, however, a couple of large absolute forecast errors. This has a considerable effect on the RMSFE but a lesser one on the signed-rank statistic, in which the size of a particular error weighs less than in the RMSFE. The direct 12-month AR forecasts do have a smaller RMSFE ratio than the

Recursive	Hor.	QN	MBE	Autom.	QN-SG
	1	1	0.99	1	1
AR	3	1	0.645	1	1
	6	1	0.962	1	1
	12	1	0.999	1	1

Direct	Hor.	AR	QN	MBE	Autom.	QN-SG	NP	NC
	1		1	0.990	1	1	0.332	1
AR	3	0.226	1	0.999	1	1	1	1
	6	0.367	1	0.999	1	1	0.991	1
	12	1	1	1	1	1	1	1

TABLE 24. p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the unemployment series from the linear AR estimated on differences is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on differences. Bottom panel: Models estimated directly on differences.

Recursive	Hor.	QN	MB	Autom.	QN-SG
	1	1	0.823	1	1
AR	3	1	0.999	1	1
	6	1	0.999	1	1
	12	1	0.723	1	1

Direct	Hor.	AR	QN	MB	Autom.	QN-SG	NP	NC
	1	0.333	1	0.823	1	1	1	1
AR	3	0.204	1	0.764	1	1	1	1
	6	0.003	1	0.404	1	1	1	1
	12	$3 \cdot 10^{-6}$	0.990	0.148	1	0.828	1	1

TABLE 25. p-values of the Wilcoxon signed-rank test for testing the null of the mean of the forecast errors from recursive forecasts of the unemployment series from the linear AR estimated on levels is equal to the mean of the corresponding forecast error from the model in each column of the table. The tests are carried out separately at each horizon and the alternative hypothesis is that the model in the table has a lower mean. Top panel: Models estimated recursively on levels. Bottom panel: Models estimated directly on levels.

other methods in Table 22, which is in accord with the information in Table 25.

Statistics on the size and composition of the ANN models for forecasting based on differenced CPI series can be found in Table 26. When forecasting recursively MBE generates the smallest models and Autometrics the largest ones. QuickNet and QN-SG lie in between. Most of the selected variables are hidden units. The average size of the MBE-based direct models decreases slightly with the forecasting horizon. It appears that there is positive correlation with the size of the model and its forecasting performance. In Table 19 and 20 models selected by MBE have the smallest RMSFE ratios, whereas Autometrics-based models have the largest ones. Models chosen using QuickNet and QN-SG lie in the middle.

Table 27 contains the same statistics for models based on levels. QN-SG now produces the most parsimonious models when forecasting recursively,

Recursive	Total	Linear	Nonlinear
QN	6.62	0.0569	6.56
MBE	3.66	0.115	3.55
AM	13.7	0.172	13.5
QN-SG	6.12	0.0569	6.06

MBE	Total	Linear	Nonlinear
1 mth	3.66	0.115	3.55
3 mths	2.88	0.295	2.59
6 mths	2.58	0.227	2.35
12 mths	2.49	0.0556	2.43

TABLE 26. Left panel: Average number of variables selected for the models gnerating recursive forecasts of the unemployment series based on differences. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on differences by MBE.

Recursive	Total	Linear	Nonlinear
QN	5.02	1	4.02
MBE	6.08	5.23	0.848
AM	13.6	1.42	12.2
QN-SG	2.98	1	1.98

MBE	Total	Linear	Nonlinear
1 mth	6.08	5.23	0.848
3 mths	6.03	5.31	0.723
6 mths	5.89	5.29	0.6
12 mths	5.12	4.34	0.782

TABLE 27. Left panel: Average number of variables selected for the models generating recursive forecasts of the unemployment series based on levels. "Total" indicates total number of variables included, "Linear" indicates the number of linear units included, and "Nonlinear" gives the number of hidden units included. Right panel: Average number of variables selected for the direct forecasts of the CPI based on differences by MBE.

and even QuickNet-based models have a smaller average size than the ones chosen by MBE. The share of linear lags is now appreciably greater than in Table 26, and this is the case for all four procedures. Autometrics selects the largest direct models, whose average size is practically the same as it is in the models for recursive forecasting. It appears that in this case, leaving out lags does not affect the average size of the model.

The correlation between the size of the model and the accuracy of the forecasts is weaker than in the previous case. A look at Table 21 shows that MBE-selected models still have the smallest RMSFE ratios, although they do not have the smallest size. Note, however, that they contain the largest number of linear lags, which may have affected the outcome. The position of Autometrics is unchanged: largest models and largest RMSFE ratios.

All models for direct forecasting contain more linear terms when the models are in levels than when they are in differences. They share this feature with the corresponding models built for forecasting the CPI.

5.5.1. *Individual countries.* The average results for the unemployment series are indicative of the results for the individual countries. No large differences can be found on the country level. However, to illustrate that not all relative RMSFE ratios are close to unity we discuss a few individual country results. The tables with the relative RMSFE for each individual country can be found at `http://econ.au.dk/research/research-centres/creates/research/research-papers/supplementary-downloads/rp-2011-28/`.

For example the direct forecasts on the differences of the Italian unemployment series are around ten percent more precise than the recursive forecasts

from the linear autoregression. Similarly to the average results MBE delivers the most accurate forecasts of the ANN procedures while Autometrics is the most imprecise in particular at the short horizons. In fact these two observations are both very stable across all eleven countries emphasizing the fact that the average results reflect the results for the individual countries rather well.

The direct forecasts on the levels of the German unemployment series are another example of a country for which it is possible to outperform the recursive linear autoregressive model on the differences. These forecasts are also an instance of a series for which the SW-filter produces more accurate forecasts than the AR-filter. MBE is again the most successful nonlinear model.

6. Conclusions

In this paper we consider macroeconomic forecasting with a flexible nonlinear model, the single-hidden layer feedforward neural network model that is a universal approximator. We apply the idea of White (2006) of transforming the specification and estimation problem of this model to a linear model specification problem. This leads to a situation in which the number of candidate variables to choose among vastly exceeds the number of observations. Three modelling techniques, White's QuickNet among them, that can handle this difficulty are compared and the models selected are used for forecasting.

The benchmark in our forecast comparisons is, with one exception, the linear AR model with recursive forecasts. It is turns out to be difficult to improve upon its forecasting precision using recursive forecasting, while the direct method seems to be a more successful approach. It appears that the Marginal Bridge Estimator of Huang et al. (2008) yields the best performing ANN models overall, but the results do vary from one country to the other. Autometrics of Doornik (2009) selects models with excellent forecasting performance when there is a well-fitting model to be discovered but does poorly when no potential model fits the data well. QuickNet selects models whose average forecasting performance lies between that of the two others. Parsimony plays a role since MBE often selects models with the fewest variables of the available alternatives. The purely nonparametric model generates relatively accurate forecasts for inflation series but is much less successful in forecasting unemployment rates. The performance of the models may also vary as a function of the forecasting horizon.

All three techniques often produce models that yield some very erroneous or 'insane' forecasts, which makes filtering them necessary. The two insanity filters considered in this paper perform almost equally well, although the AR filter may have a slight edge over the filter that Swanson and White (1995) introduced. Multicollinearity is the main reason for insane forecasts, and it might be a good idea to develop all three modelling strategies further in order to reduce the probability of the outcomes in which the final model contains very strongly linearly correlated variables.

We find that that testing linearity before variable selection does not help in choosing useful models. It may do so for certain countries and variables

but may lead to weakened forecasting performance in some others. For this reason it cannot be recommended as a part of any of the three modelling strategies under consideration.

Forecasts are generated using both the recursive and the direct method. Overall, direct forecasting is somewhat superior to the recursive technique, but it does not dominate the latter. The results vary from one country and variable to the other. This is true also in comparing the accuracy of recursive and direct forecasts from linear AR models: on average direct forecasts are more accurate than the recusive ones.

When it comes to choosing between models based on first differences of the series and ones specified and estimated using levels it turns out that in forecasting the CPI models built on levels tend to generate more accurate forecasts on average than the corresponding models constructed using first differences. It is not clear why that is the case. In forecasting unemployment rates the outcome is less clear: the models based on levels cannot be viewed as superior to models built on first differences.

A general conclusion is that the ANN model can be useful in macroeconomic forecasting, but that the linear AR model is a serious competitor. In practice, the forecaster may experiment with several models and methods between settling for one, if the final goal is to find a model with the best performance for a given country and variable. Another possibility left for further work would be to combine recursive and direct forcasts obtained with various linear AR and ANN models.

Finally, the purpose of this work has not been to compare the forecasting performance of different nonlinear models. Doing so in a satisfactory fashion would require a vast amount of resources. It would also shift the focus away from our main aim: comparing different modelling techniques for the single-hidden layer ANN model made possible by the work of White (2006), and has not been attempted here.

References

Ahmed, N. K., A. F. Atiya, N. El Gayer, and H. El-Shishiny (2010). An empirical comparison of machine learning tools for time series forecasting. *Econometric Reviews 29, 594-621*.

Anders, U. and O. Korn (1999). Model selection in neural networks. *Neural Networks 12*, 309–323.

Candès, E. J. (1998). *Ridgelets: theory and applications*. PhD thesis, Dept. of Statistics, Stanford University.

Candès, E. J. (2003). Ridgelets: estimating with ridge functions. *The Annals of Statistics 31*, 1561–1599.

Costantini, M. and R. M. Kunst (2011). On the usefulness of the Diebold-Mariano test in the selection of prediction models: some Monte Carlo evidence. Working paper, University of Vienna.

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

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 134–144.

Doornik, J. A. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics*, pp. 88–122. Oxford University Press, Oxford.

Fine, T. L. (1999). *Feedforward neural network methodology*. Springer Verlag, New York.

Gallant, A. R. (1984). The Fourier flexible form. *American Journal of Agricultural Economics 66*, 204–208.

Ghaddar, D. K. and H. Tong (1981). Data transformation and self-exciting threshold autoregression. *Applied Statistics 30*, 238–248.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*, 1545–1578.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*, 65–99.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, New York.

Hendry, D. F. and H. M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal 115*, 32–61.

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

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 587–613.

Hwang, J. T. G. and A. A. Ding (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 748–757.

Kock, A. B. and T. Teräsvirta (2011). Forecasting with nonlinear time series models. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, pp. 61–87. Oxford University Press, Oxford.

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

Kuan, C.-M. and T. Liu (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics 10*, 347–364.

Lee, T.-H., H. White, and C. W. J. Granger (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics 56*, 269–290.

Makridakis, S. and M. Hibon (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Marcellino, M. (2002). Instability and non-linearity in the EMU. Discussion Paper No. 3312, Centre for Economic Policy Research.

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*, 499–526.

Medeiros, M. C., T. Teräsvirta, and G. Rech (2006). Building neural network models for time series: A statistical approach. *Journal of Forecasting 25*,

49–75.

Perez-Amaral, T., G. M. Gallo, and H. White (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics 65*, 821–838.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics 99*, 39–61.

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

Simon, H. (1999). *Neural networks: a comprehensive foundation*. Prentice Hall, Upper Saddle River, NJ.

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

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

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

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

Teräsvirta, T. (1998). Modeling economic relationships with smooth transition regressions. In A. Ullah and D. E. A. Giles (Eds.), *Handbook of Applied Economic Statistics*, pp. 507–552. Dekker, New York.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 413–457. Elsevier, North-Holland.

Teräsvirta, T., C. W. J. Granger, and D. Tjøstheim (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford.

Teräsvirta, T., C. F. Lin, and C. W. J. Granger (1993). Power of the neural network linearity test. *Journal of Time Series Analysis 14*, 209–220.

Teräsvirta, T., D. van Dijk, and M. C. Medeiros (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting 21*, 755–774.

White, H. (2006). Approximate nonlinear forecasting methods. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 459–512. Elsevier, Amsterdam.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 80–83.

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

Appendix A. Creating the pool of hidden units

First we shall consider the procedure of White (2006). It consists of three steps.

(1) Rewrite the argument of the logistic function in (1) as follows:

$$\gamma' \mathbf{z}_t = \gamma_0 + \gamma_1(\gamma_2' \widetilde{\mathbf{z}}_t)$$

using the notation: $\mathbf{z}_t = (1, \widetilde{\mathbf{z}}_t')'$. For convenience, assume that each element of $\widetilde{\mathbf{z}}_t$ has mean zero. The vector γ_2 is the direction vector whose length equals one, and it is selected first. This is done as follows. Let the random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Then set $\gamma_2 = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1/2}$ which is uniformly distributed on the unit sphere $\mathcal{S}^{p-1}$ in $\mathbb{R}^p$.

(2) Given γ_2, choose $\gamma_1 > 0$ such that it is at least of the magnitude of $\widehat{\sigma}_z = \text{std}(\gamma_2' \widetilde{\mathbf{z}}_t)$ with the range spanning modest multiples of $\widehat{\sigma}_z$. Draw γ_1 at random from this range. The scalar γ_1 gives the length of the vector γ_2 and controls the slope of the hidden unit as a function of $\gamma_2' \widetilde{\mathbf{z}}_t$.

(3) Choose γ_0 such that it has mean zero and standard deviation comparable to $\text{std}(\gamma_1(\gamma_2' \widetilde{\mathbf{z}}_t))$. Draw γ_0 at random from this distribution. This scalar controls the location of the hidden unit.

In our experiments, selecting γ_1 as in step (2) above frequently led to values of this parameter that were too small in the sense that the hidden unit did not display sufficient variation in the sample. This had adverse consequences to the forecasts. To avoid them, we constructed a modification with the following structure:

(1) Rewrite the argument of the logistic function in (1) as follows:

$$\gamma' \mathbf{z}_t = \gamma_1 / \hat{\sigma}_z(\gamma_2' \tilde{\mathbf{z}}_t - \gamma_0) \tag{9}$$

Choose γ_2' as described above in step (1) above.

(2) Next obtain γ_0. Consider the values $x_t = \gamma_2' \tilde{\mathbf{z}}_t$, $t = 1, ..., T,$. Let $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of this sequence. Let $d = x_{\max} - x_{\min}$. Now draw γ_0 from a uniform distribution on $[x_{\min} + \delta d, x_{\max} - \delta d]$ for $\delta \in [0, 0.5]$. We choose $\delta = 0.1$. In this way we make sure that the hidden units are not centered at very small or large values of $\gamma_2' \tilde{\mathbf{z}}_t$. As a result of the parameterization (9), demeaning $\widetilde{\mathbf{z}}_t$ is not necessary.

(3) Finally, the slope parameter γ_1 is chosen uniformly at random from the set $\left\{1.25^j : \ j = 0, 1, ..., 20\right\}$. Hence, the smallest possible value of γ_1 is 1 while the largest possible value is 87. The set is deliberately constructed to be denser for small values since the slope of the logistic function changes more for changes in γ_1 when γ_1 is small than when γ_1 is big. For large values of γ_1 changes in γ_1 will not affect the slope of the logistic function much and so it is less important to have a dense grid here.

The decisive difference between the two strategies lies in choosing γ_1. In the strategy of White (2006), γ_1 is not a scale-free parameter. That is, a change of units in $\tilde{\mathbf{z}}_t$ affects the set of possible slopes that can be selected, which is

a disadvantage. In (9), γ_1 is a scale-free parameter due to the division by $\widehat{\sigma}_z$, for discussion, see for example Teräsvirta (1998). This makes it possible to define a reasonable range for this parameter. The minimum value of the scale-free γ_1 is set to unity in order to avoid logistic functions with too little sample variation.

Research Papers
2011

CREATES
Center for Research in Econometric Analysis of Time Series

2011-13: Dennis Kristensen: Nonparametric Detection and Estimation of Structural Change

2011-14: Stefano Grassi and Paolo Santucci de Magistris: When Long Memory Meets the Kalman Filter: A Comparative Study

2011-15: Antonio E. Noriega and Daniel Ventosa-Santaularia: A Simple Test for Spurious Regressions

2011-16: Stefano Grassi and Tommaso Proietti: Characterizing economic trends by Bayesian stochastic model specification search

2011-17: Søren Johansen and Theis Lange: Some econometric results for the Blanchard-Watson bubble model

2011-18: Tom Engsted and Thomas Q. Pedersen: Bias-correction in vector autoregressive models: A simulation study

2011-19: Kim Christensen, Roel Oomen and Mark Podolskij: Fact or friction: Jumps at ultra high frequency

2011-20: Charlotte Christiansen: Predicting Severe Simultaneous Recessions Using Yield Spreads as Leading Indicators

2011-21: Bent Jesper Christensen, Olaf Posch and Michel van der Wel: Estimating Dynamic Equilibrium Models using Macro and Financial Data

2011-22: Antonis Papapantoleon, John Schoenmakers and David Skovmand: Efficient and accurate log-Lévi approximations to Lévi driven LIBOR models

2011-23: Torben G. Andersen, Dobrislav Dobrev and Ernst Schaumburg: A Functional Filtering and Neighborhood Truncation Approach to Integrated Quarticity Estimation

2011-24: Cristina Amado and Timo Teräsvirta: Conditional Correlation Models of Autoregressive Conditional Heteroskedasticity with Nonstationary GARCH Equations

2011-25: Stephen T. Ziliak: Field Experiments in Economics: Comment on an article by Levitt and List

2011-26: Rasmus Tangsgaard Varneskov and Pierre Perron: Combining Long Memory and Level Shifts in Modeling and Forecasting of Persistent Time Series

2011-27: Anders Bredahl Kock and Timo Teräsvirta: Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques