Forecasting Economic Time Series Using Flexible Versus Fixed Specification and Linear Versus Nonlinear Econometric Models

Norman R. Swanson
The Pennsylvania State University, 521 Kern Graduate Bldg.,
Department of Economics, University Park, PA 16802, USA.
phone: (814) 865-2234; fax: (814) 863-4775;
email: nswanson@psu.edu; web: http://packrat.la.psu.edu/econ/nswanson/homens.htm

and

Halbert White
Research Group for Econometric Analysis, Department of Economics,
University of California, San Diego, La Jolla, California 92093, USA.
phone: (619) 534-3502 fax: (619) 534-0877
email: hwhite@ucsd.edu

this revision: March 1997

Forecasting Economic Time Series Using Flexible Versus Fixed Specification and Linear Versus Nonlinear Econometric Models

ABSTRACT

Nine macroeconomic variables are forecast in a real-time scenario using a variety of flexible specification, fixed specification, linear, and nonlinear econometric models. All models are allowed to evolve through time, and our analysis focuses on model selection and performance. In the context of real-time forecasts, flexible specification models (including linear ARX and nonlinear artificial neural network) appear to offer a useful and viable alternative to less flexible fixed specification linear models for a subset of the economic variables which we examine, particularly at forecast horizons greater than 1-step ahead. We speculate that one reason for this result is that the economy is evolving (rather slowly) over time. This feature cannot easily be captured by fixed specification linear models, however, and manifests itself in the form of evolving coefficient estimates. We also provide additional evidence supporting the claim that models which "win" based on one model selection criterion (say a squared error measure) do not necessarily win when an alternative selection criterion is used (say a confusion rate measure), thus highlighting the importance of the particular cost function which is used by forecasters and "end-users" to evaluate their models. A wide variety of different model selection criteria and statistical tests are used to illustrate our findings.

Keywords: Cointegration, Confusion Rate, Linearity, Model Selection, Nonlinearity, Parameter Evolution, Real Time Forecasting.

1. Introduction and Overview

In a recent paper, Swanson and White (1996: SW) examined the usefulness of a class of artificial neural networks (ANNs) for forecasting nine macroeconomic variables. Forecasts based on the ANNs as well as various linear econometric models were compared to professional forecasts from the *Survey of Professional Forecasters* (SPF: see Croushore (1993)). The analysis provided direct evidence with regard to the usefulness of econometric models relative to economic forecasts made available by various professional forecasters in both the private and public sectors. In order to ensure that the same information was available for the construction of both the econometric and professional forecasts, an *ex ante* or *real−time* experiment was conducted. Results indicated that: (i) The in-sample Schwarz Information Criterion (SIC) is not superior to true out-of-sample forecast performance measures for selecting models; (ii) Econometric models frequently outperform the SPF based on a variety of model selection criteria; (iii) Flexible specification linear models appear to be promising, although flexible specification nonlinear (ANN) models are not particularly useful in the context examined.

In this paper we examine the same dataset from a different perspective, and focus on a number of related but different issues. First, we compare 5 different econometric models, each with the other, and do not focus our statistical analysis solely on a comparison of professional forecasts with econometric forecasts. We consider three fixed specification linear models: a random walk, a random walk with drift, and a "best" linear autoregressive model with exogenous variables (ARX). Two flexible specification models are also examined. They are a flexible specification linear model and an flexible specification nonlinear model. The nonlinear model is a single hidden layer "feedforward" artificial neural network. Only the flexible specification models are allowed to "choose" new specifications at each point in time, while all of our models are re-estimated at each point in time using "windows" of data of varying lengths, in order to construct sequences of what we term a "rolling window" based forecasts. In this context we ask the following questions: (1) "Are the flexible specification models more or less susceptible to parameter variability than the fixed specification models?" (2) "Is it useful to allow for so much flexibility in our models, or do fixed specification models perform equally well?" Second, we construct models using levels rather than differenced data. The use of levels data has the advantage that cointegration of unspecified form is allowed for. This is particularly useful if some concern exists that cointegrating spaces, ranks, etc. are poorly estimated (e.g. when small samples of data are used). However, the approach has the obvious drawback that cointegrating restrictions are not explicitly estimated.

Our approach of estimating models in levels poses no problem for the univariate random walk and random walk with drift models. However, given the nonstationary behavior of the variables, the other models are estimated inefficiently. This leads to two questions: (1) "How well do inefficiently estimated multivariate models perform relative to simple random walk models?" (2) "How well do our levels models perform relative to efficiently parameterized cointegrating systems?" These questions are addressed below, and are interesting for a number of reasons. First, parsimonious models are widely known to forecast at least as well as more complex multivariate models for a number of economic time series. Second, cointegrating vectors and the ranks of cointegrating spaces are difficult to estimate precisely given small samples, and this problem may be exaggerated by our use of a real-time forecasting scenario. Third, since we are interested in forecasts, and not in conducting statistical inference on the parameters of our models, estimating models in levels may provide a useful alternative to many more complex estimation strategies, *without* having any effect on any of the statistical inferences which we carry out. There are a number of recent papers which address some of these issues (e.g. Clements and Hendry (1995), Hoffman and Rasche (1996), and Lin and Tsay (1996)), and we attempt to add to the discussion.

Another aspect of our experimental setup is that we directly focus on the relative merits of (i) flexible specification versus fixed specification models and (ii) linear versus nonlinear models. We compare these types of models using a variety of statistical tests (Wilcoxon signed rank tests, market timing tests, and forecast error based tests) as well as by directly comparing the point estimates of a number of different model selection criteria (including forecast mean squared error, mean absolute forecast error deviation, mean absolute percentage forecast error, confusion rate, and mean SIC among others). By doing so we hope to further our understanding of the relative merits of our various different modeling strategies with respect to forecast horizon and the individual characteristics of each variable.

Our results suggest that ANN models appear to offer a useful and viable alternative to less flexible fixed specification linear models for a small subset of the economic variables which we examine, particularly at forecast horizons greater than 1-step ahead. We also provide additional evidence supporting the claim that models which "win" based on one model selection criterion (say a squared error measure) do not necessarily win when an alternative selection criterion is used (say a confusion rate measure), thus highlighting the importance of the particular cost function which is used by forecasters and "end-users" to evaluate their models. Finally, our levels models outperform random walk models and ARIMA models for 7 of the 9 variables which we examine, suggesting that it is advantageous to estimate models in levels (as opposed to differences), particularly when the nonstationary data being examined are cointegrated.

In summary, we adopt a rolling window approach to model selection and forecasting for "choosing" between flexible specification and fixed specification as well as linear and nonlinear econometric models. By doing so, we hope to contribute to the understanding of the relative merits of various econometric modeling strategies. Contributions are also attempted in a number of other related areas. In particular, in the context of constructing economic forecasts we ask how one may choose among the wide variety of different model selection criteria which are available, and what the consequences of such choices may be. We also compare five different econometric models in pairwise fashion in an attempt to determine whether some models tend to outperform other models in a systematic manner. We place particular emphasis on the comparison of simple models such as random walks with more complex models such as artificial neural networks. The rest of paper is organized as follows. Section 2 discusses the data. Sections 3 and 4 outline our fixed specification and flexible specification models, while Section 5 discusses flexible specification versus parameter evolution in the current context. Section 6 outlines estimation strategies, and Section 7 summarizes the model selection criteria which we examine. Our results are gathered in Section 8, and Section 9 concludes.

2. Data

We use the same data as that examined in SW. However, we focus on economic data which is publicly available, and do not use the SPF. For the period 1960:1 to 1993:3 we examine the following nine quarterly U.S. macroeconomic time series (The data are available in hard copy or by email upon request):

U	*Civilian Unemployment Rate* : *SA*, *%*, *Averaged monthly.*
R	*Aaa Corporate Bond Yield* : *Moody´s*, *%*, *Averaged monthly.*
IP	*Industrial Production Index* : *SA*, *index*, 1987=100, *Averaged monthly.*
NGNP	*Gross National Product* : *SA*, *$billions*, *Quarterly.*
Π	*Corporate Profits After Taxes* : *SA*, *$billions*, *Quarterly.*
RGNP	*Gross National Product* : *SA*, *$billions* 1987, *Quarterly.*
PCE	*Personal Consumption Expenditures* : *SA*, *$billions* 1987, *Quarterly.*
ΔBI	*Change in Business Inventories* : *SA*, *$billions* 1987, *Quarterly.*
Net X	*Net Exports of Goods* and *Services* : *SA*, *$billions* 1987, *Quarterly.*

A significant feature of this data set is that all of the data are as first reported in each successive issue of the Survey of Current Business. We call such "first available" data the "unrevised" data. The data are quite different (particularly for the non financial variables) from data which have been periodically revised over time (which we will call revised data), and which are available through various on-line sources such as CITIBASE. What makes this data collection strategy different is that it allows us to formulate and estimate econometric models at time period *t*-1, for instance, using only data which were available prior to period *t*. This allows us to guard against future information creeping into our econometric specifications, and thus our forecasts, through data revisions, definitional changes, benchmark revisions, and seasonal two-sided moving average adjustments, for instance. Specifically, we avoid using data available after period *t-h*, where *h* is the horizon of our forecasts, In this respect we are constructing *ex ante* or *real−time* forecasts, thus addressing the data revision problems pointed out by Fair and Shiller (1990).[1]

[1] Our approach is to use only the "first available" data. Thus, our data are only real-time in the sense that were we to gather them from CITIBASE, say, then the older data would have been revised, seasonally adjusted, etc., many times. This process in turn would potentially allow future information to creep into past observations, thus ensuring that our data were not "real-time". Further, the newer data would be

This particular type of data set is of value when comparing econometric forecasts with professional forecasts (which are necessarily real-time) as in SW. However, real-time forecasting using unrevised data should be of interest for a number of other reasons also, and hence we use such a dataset here. For instance, traders of stocks, bonds and commodities in various business settings need to react very quickly to changing market information. This naturally entails reacting to new economic information as soon as it becomes available. Regardless of whether the traders obtain their "new" information from public sources such as government agencies, or from private forecasting firms, the forecasts they receive are real-time, insofar as very recent information was used to construct the forecasts. As an example, note that many relevant economic variables (such as gross national product, industrial production, exports, imports, inventories, sales, and employment) are subject to revision, and are periodically updated (in some cases these updates continue ad infinitum). Thus, the information which traders react to is at least partially "unrevised" in the sense discussed above. In this context, if we are attempting to provide econometric forecasts which are meant to compete with forecasts gotten from other sources (such as professional forecasting services or, simply, individuals acting on "gut" instinct) it seems reasonable to use the same (at least partially unrevised) information which is available to our competitors. Thus, at least in some cases the use of unrevised data is preferable to revised data. Given these arguments, it follows that constructing our econometric models using only fully revised data, and then later plugging unrevised data into the models to produce forecasts does not make sense, for example. It seems more reasonable to use the same type of data to estimate the model as we later feed into the model to yield economic forecasts.

3. Nonflexible Specification Models

unrevised, while the older data would be revised many times. This adds further complexity to the picture, as one is then left wondering whether one is forecasting unrevised or revised vintages of data. Our approach avoids "mixing up" the data in this way, but suffers from using less information than may actually be available, if a new data set were collected for each variable at each point in time. In summary, we attempt to forecast only first available information (which is only one vintage of data which may be of interest to policy makers, etc.), and ensure that future information has not been allowed to creep into the dataset which we use at each point in time to construct our forecasts. It should perhaps be stressed that in order to construct a truly "real-time" dataset, we would need 45 datasets for each variable. In such cases, users can form models to extract information from past revisions, etc.. Some references which contain related discussions about data revisions, extracting information from multiple vintages of data, and preliminary data error include: Boschen and Grossman (1982), Maravall and Pierce (1983), Mariano and Tanizaki (1994) and Patterson (1995).

Some of the models and model selection criteria discussed in Sections 3, 4, 6 and 7 are similar and/or related to those in SW. For completeness, some parts of the descriptions of the models, etc. are taken from SW. For further details, the reader is referred to SW. The linear models specified in this paper are all special cases of the following autoregressive model with exogenous variables (ARX):

$$y_{t+h-1} = \alpha_0 + \sum_{i=1}^{K1} \beta_i y_{t-i} + \sum_{i=1}^{K2} \gamma_i x_{t-i} + \sum_{i=1}^{K3} \delta_i z_{t-i} + u_{t+h-1}, \tag{1}$$

where y_t is one of the nine macroeconomic variables, h is the horizon of our forecast, and u_{t+h-1} is a scalar innovation. The selection of regressors (x_t and z_t) is discussed below. (ARIMAX version of (1) were also estimated, and results are discussed below). We estimate a total of 22 versions of (1). The first model corresponds to a random walk, where $\alpha_0 = 0$, $\beta_1 = 1$, $\beta_i = 0$, $i=2,..., K1$ and $\gamma_i = \delta_i = 0$, $i=1,...,5$. A second model allows for a nonzero drift term, so that $\alpha_0 \neq 0$ in general. The next five models are AR(K1) processes, where K1=1,...,5 and $\gamma_i = \delta_i = 0$, for all i. The remaining versions of (1) which we examine are ARX models in which: (i) K1=K2=1,...,5 and $\delta_i = 0$, $i=1,...,K3$, (ii) K1=K3=1,...,5 and $\gamma_i = 0$, $i=1,...,K2$, and (iii) K1=K2=K3=1,...,5. This group of models augments the models examined in SW by including a random walk with drift model in both the "pre-sample" model selection stages and the forecasting stages of our experiment. One advantage of this approach is that we can directly compare our flexible specification linear and nonlinear models with random walk models. This is a useful comparison, as random walk models have been seen to perform especially well relative to a wide class of more complex alternatives when forecasting macroeconomic variables. Furthermore, the random walk models explicitly impose unit roots, although they do not allow for cointegration among the data. On the other hand, the ARX models allow for cointegration of unspecified form, as the equations are estimated using levels data (see discussion below). [2] We consider these linear models as special cases of a fairly broad array of forecasting models, while realizing that various other linear models that we don't examine here are also available. Below, the random walk and random walk with drift models are denoted using

[2] It should be noted that we do not consider issues related to seasonality here (for an interesting survey on seasonality see Franses (1995)). Instead we take the approach of using seasonally adjusted data in our analysis, which we believe is not unreasonable, given that much U.S. data are available only as seasonally adjusted series. However, issues related to spurious nonlinearity, for example, arise, as discussed by Ghysels, Granger, and Siklos (1996), among others.

obvious notation, while the other models are referred to by the ordered triplet (K1,K2,K3), and are called LINEAR ARX MODELS.

4. Adaptive Artificial Neural Network Models

We examine the subset of the class of flexible nonlinear models called artificial neural networks commonly known as single hidden layer "feedforward" networks. A primary motivation for the ANN models is an effort by cognitive scientists to emulate the computational structure of the human brain. Progress to date has resulted in the development of network models that are in many respects similar, but potentially richer than the wide variety of flexible functional forms and semi-parametric models familiar in econometrics. [3] In economics, ANN models have been used in a variety of financial applications, for example. For some discussion of these applications see White (1988, 1989), Moody and Utans (1991), Dorsey, Johnson and van Boening (1994), Kuan and White (1994), and Swanson and White (1995). ANNs have also been used to forecast and model macroeconomic data in recent years (for example, see Moody, Levin and Rehfuss (1993), Maasoumi, Khotanzad and Abaye (1994), Häfke and Helmenstein (1994), Soni, Otruba, Häfke and Natter (1995) and Swanson and White (1996)). The ANN regression models considered here have the form:

$$f(w,\ \theta) = \tilde{w}'\kappa + \sum_{j=1}^{q} G(\tilde{w}'\ \pi_j)\ \lambda_j \tag{2}$$

where $\tilde{w} = (1, w')'$ is a (column) vector of explanatory variables,

$$w = (y_{t-1}, \ldots , y_{t-K1}, x_{t-1}, \ldots , x_{t-K2}, z_{t-1}, \ldots , z_{t-K3})',$$

$\theta = (\kappa', \lambda', \pi')'$, $\lambda = (\lambda_1, \ldots, \lambda_q)'$, $\pi = (\pi'_1, \ldots, \pi'_q)'$, q is a given integer, and G is a given nonlinear function, in our case, the logistic cumulative distribution function (c.d.f.) $G(z) = 1/(1 + \exp(-z))$. The variables w correspond to the variables considered in the linear forecasting models described above, and the number of "hidden units", q, is set equal to 5. Clearly, (2) is a generalization of (1), and is equal to (1)

[3] Other nonlinear methods are also often applied in economics. For example, Mulhern and Caprara (1994) examine the usefulness of a nearest neighbor model for forecasting market response. An interesting discussion of the usefulness of nonlinear models is given in Ramsey (1996), while many examples and further references are contained in Brock, Hsieh and LeBaron (1991), Granger (1993), and Granger and Teräsvirta (1993).

when $\lambda_i = 0$, $i=1,...,5$. We will attempt to determine whether inclusion of the nonlinear terms enhances forecasting ability, assuming that overfitting is properly avoided.

A network interpretation of (2) which is also given in SW is as follows. "Input units" send signals $w_0(=1)$, $w_1, \ldots, w_r$ over "connections" that amplify or attenuate the signals by a factor ("weight") π_{ji}, $i=0, \ldots, r$, $j=1, \ldots, q$. The signals arriving at "intermediate" or "hidden" units are first summed (resulting in $\tilde{w}'\pi_j$) and then converted to a "hidden unit activation" $G(\tilde{w}'\pi_j)$ by the operation of the "hidden unit activation function", G. The next layer operates similarly, with hidden activations sent over connections to the "output unit." As before, signals are attenuated or amplified by weights λ_j and summed. In addition, signals are sent directly from input to output over connections with weights κ. A nonlinear activation transformation at the output is also possible, but we avoid this here for simplicity. In network terminology, $f(w, \theta)$ is the "network output activation" of a "hidden layer feedforward network" with "inputs" w and "network weights" θ. The parameters π_j are called "input to hidden unit weights," while the parameters λ_j are called "hidden to output unit weights." The parameters κ are called "input to output unit weights." It should perhaps be noted that Hornik, Stinchcombe and White (1989, 1990) Carroll and Dickinson (1989) and Funahashi (1989), among others, have shown that functions of the form (2) are capable of approximating arbitrary functions of w arbitrarily well given q sufficiently large and a suitable choice of θ. This property is known as "universal approximation", and is an appealing characteristic of ANN models, perhaps in part accounting for the reported success of such models.

Versions of equation (2), with the number of hidden units set to zero are also estimated. These models form our class of flexible specification linear models, and are estimated much the same way as the ANN model (as explained in the next section), except that $\lambda_i = 0$, $i=1,...,5$. In this way we attempt to differentiate between potential gains brought about by the inclusion of the hidden units from gains due to the flexible nature of our ANN models.

5. Model Adaptation versus Parameter Evolution

Each of the models in our experiment is re-estimated 45 times in order to construct a sequence of 45 quarterly 1 and 4-quarter ahead forecasts. In this way, the parameters in all of the models are all allowed to evolve over time. One feature of re-estimation strategies such as this is that if the "window" or sample of data used to estimate the model is allowed to increase over time, then the system may be seen to "evolve" to some final form (where by final form we mean that a model has a fixed specification and relatively constant parameters). Examples of analyses conducted using increasing windows of data include Fair and Shiller (1990), Leitch and Tanner (1991), Pesaran and Timmerman (1994a), Thoma (1994) and Swanson (1996). Here we allow for the underlying relation between the economic variables to be evolving over time, but perhaps not to some final form. In particular, we use fixed window sizes of 40, 68 and 76 quarters of data in our regression estimations. Thus, while the sample changes with each period, it does not increase in size over time. In this scenario, the system may still evolve to some final form as in the increasing windows case (assuming that the window sizes are sufficiently large). However, our approach has the added feature that if the system is not evolving to some final form, then the model is allowed to update by discarding older and less relevant observations. To examine parameter evolution in the various models we track the parameter values of all of the regressors used in each of our models for the entire 45 quarter ex ante forecast sample period. We hope to add thereby to the discussion concerning the evolutionary nature of the parameters and the usefulness of rolling windows in econometric models.

Our use of rolling windows in the context of model (1) does not allow for the econometric specifications of our models to change over time. In fact, it seems reasonable that if the parameter values are changing, then perhaps the "best" econometric specifications with respect to forecasting may also be changing over time. Models which have the characteristic that new specifications are chosen at each point in time are called "flexible". From (2) we have flexible linear as well as nonlinear models. By estimating these flexible models in addition to a wide variety of fixed specification models we hope to find evidence pertaining to the relative merits of the two different classes of models. For example, if the fixed specification models forecast better than the flexible specification models, and exhibit little parameter evolution, then we have direct evidence that the flexible specification models are perhaps too general for the purpose at hand. As one crude measure of model flexibility, we keep a running count of the total number of parameters estimated by the flexible specification models. One shortcoming of this approach is that the number of parameters estimated may remain stable while the model specification is changing. Thus, the variation in the number of parameters estimated is a conservative estimate of the extent to which model flexibility is being taken advantage of. Also, even if the flexible models change each period, there is no guarantee that the "flexible" model forecasts will be superior to those produced by a "fixed specification" model. The same argument holds for the examination of parameter evolution. Thus, it should be stressed that our examination of model flexibility and parameter evolution is not meant as a substitute for our forecast comparison. Rather, it is meant to augment our knowledge concerning the empirical characteristics of the various forecast models examined.

6. Estimation Strategies

The parameters of the linear models are all estimated using standard least squares, with all variables entering into the models in levels. For the relevant cases (which do not include the random walk models) we make the assumption that the variables are cointegrated. [4] In this sense, the single equations that we estimate can be seen to arise from inefficiently estimated vector error correction (VEC) models. In our real-time scenario, this procedure has at least two advantages. First it allows us to avoid re-estimating the rank of any potential cointegration between the variables at each point in time, while still accounting for any cointegration that may be present, albeit in an inefficient way. Second, we need not

[4] This assumption is borne out by the data. In particular, augmented Dickey-Fuller tests were applied to each series, and the null hypothesis that each series was I(1) failed to reject for all 9 variables. We then used Johansen's trace test statistic to estimate the rank of the cointegrating space for each of the nine models, using data up until 1982:2 (which is the last period before the start of our out-of-sample period). These tests were carried out using the lag order suggested for our "best" linear models, and were constructed for each of the five standard deterministic trend specifications commonly used in the context of Johansen's methodology. For 5 of the models, cointegration was found regardless of trend specification, while for 2 of the models cointegration was found for three out of 5 trend specifications. For the other two models (interest rates and net exports), the "best" linear model was the random walk, and so no cointegration test statistics were constructed. This suggests that levels models (or cointegrated vector error correction models) are relevant for our analysis, and that differenced data models may be inappropriate. In order to obtain a clearer picture of the usefulness of differenced data models, we re-did our analysis using ARMA models with differenced data instead of AR models in levels. Based on MSE criteria, the AR models always outperformed the ARMA models, except for IP in the h=1 case. We take this to constitute evidence that the cointegration among the data suggest that using differences is inappropriate. However, as mentioned above, our analysis is necessarily limited by the particular classes of econometric models which we consider.

re-estimate any potential cointegrating vectors at each point in time. This is particularly advantageous for the small finite samples which we are considering here, as the Johansen (1988) and Johansen and Juselius (1990) maximum likelihood method has been shown to produce widely varying (and perhaps imprecise) estimates using the rolling window approach adopted here (see Swanson (1996)). Furthermore, a number of recent papers suggest that using levels models may be reasonable in many cases, when the precise form of the cointegrating space is unknown. [5] Finally, as we are interested only in out-of-sample performance and model selection, we do not conduct any inference on the regression coefficients from our in-sample estimations, so that the non-standard distributions of the coefficients do not directly affect our analysis. However, given the above discussion, we suggest that further study into this and related issues seems to be warranted, and is the subject of ongoing research. Of final note is that the variables chosen as predictor variables in each of our regression models were obtained by using a "training" set of data from 1960:1-1982:2 to determine which macroeconomic variables were most closely related, in-sample, in terms of both cointegrating properties and in-sample fit.

The linear fixed specification regression models for 40, 58 and 76 quarter windows of data are evaluated using sequences of out-of-sample 1-quarter and 4-quarter ahead forecast errors. The forecast errors are generated by performing the regressions over a given window terminating at observation $t-h$, say, and then computing the error in forecasting y_t for $h=1,4$ using data available at time $t-h$ and the coefficients estimated using data in the window terminating at time $t-h$. Each time the window rolls forward one period, a new out-of-sample residual is generated, simulating true out-of-sample predictions and prediction errors made in real-time by this process. For our study, the smallest value for $t-h$ corresponds to the second quarter of 1983 for $h=1$ (and the third quarter of 1982 for $h=4$) while the largest corresponds to the second quarter of 1993 for $h=1$ (and the third quarter of 1992 for $h=4$). This

[5] In particular, Hoffman and Rasche (1996) conclude that if there is an advantage to using cointegration models (CIMs) instead of levels models (LMs), it is usually at longer forecast horizons. Lin and Tsay (1996), on the other hand, find that CIMs perform well in Monte Carlo, when the "correct" constraints are placed on the system. However, based on actual data, the results are very mixed, with CIMs doing quite poorly in many cases. Clements and Hendry (1995) also contains empirical evidence based on forecasts of M1 which suggests that there is little to choose between CIMs and LMs, although they find using Monte Carlo that gains from imposing long-run constraints become more apparent at small estimation sample sizes. One result which seems clear among these papers, though, is that pure difference models perform quite poorly.

staggered timing allows us to generate sequences of 45 out-of-sample 1-quarter and 1-year ahead forecast errors based on forecasts for the same period - 1982:3-1993:3.

The ANN estimation involves first performing forward stepwise linear regression, with regressors added one at a time until no additional regressor can be added to improve the SIC. The linear regression coefficients are thereafter fixed. Next, a single hidden unit is added (i.e. q is set to 1), and regressors are selected one by one for connection to the first hidden unit, until the SIC cannot be improved any more. More hidden units are sequentially attached to the network as needed, in order to maximize the in-sample SIC value at each point in time. This procedure is repeated using the various window lengths before each forecast is constructed. A final feature of our ANN estimation strategy is that it is possible to have no hidden units chosen. In this sense, the SIC-best model may be linear for some periods, and perhaps non-linear for others. [6] The linear flexible specification models were estimated in much the same way as the ANN models, except that no hidden units were allowed to enter into any final specifications. This is an important feature of our analysis, as it underscores a more subtle difference between our flexible specification and fixed specification linear models. In particular, the fixed specification models are fixed over time, with the "best" fixed specification models chosen based on a "training" set of data, while the flexible specification models evolve through time based on an in-sample model selection criterion. For this reason we "tip the scale" in favor of the linear fixed specification models by reporting results not only for the "best" model, but also for a random walk and a random walk with drift model. Note, however, that simply reporting the forecasting results from *all* of the estimated models would probably severely bias our experimental results, as the preferred forecasting models might mistakenly be equated with the models exhibiting the most appealing model selection statistics, ex post.

[6] Our approach to nonlinearity is to compare various models from a forecasting perspective. If the nonlinear models win, then we have direct evidence of the usefulness of the nonlinear model, without needing to initially test for nonlinearity. However, nonlinearity tests are useful in other respects, as they may help us to form initial expectations concerning whether or not our ANNs will perform well, for example. For this reason we constructed BDS, RESET, and White test statistics. The latter two of these tests may be viewed as tests of overall model specification. However, as nonlinearity is one form of misspecification against which they have been shown to have power, we feel that the tests are of interest. Our findings are based on initial linear ARX regressions which assume the lag order used in the paper. The sample used ends in 1982:2. White tests are constructed with and without cross product terms; RESET tests use 1, 2, and 3 fitted powers of the dependent variable; and BDS tests are based on m=2. For 7 of 9 series, we obtain rejections for at least two of three types of tests. The two exceptions are Π and *RGNP*. (Detailed results are available from the corresponding author.)

7. Evaluation Methods

7.1 Model Selection Criteria

The MSE is one of the more popular measures of forecast accuracy. We compute the forecast mean squared error of the 45 forecast errors for each model, window and horizon, *h*=1,4.

$$MSE = \sum_{t=1}^{T} \hat{u}_{I,t}^{2} / T, \tag{3}$$

where $\hat{u}$ is the forecast error, I corresponds to any one of the estimated econometric models, and T is the post-sample size, 45 in our case. (Related measures which are also calculated include mean absolute forecast error deviation - MAD, and mean absolute percentage forecast error - MAPE). Even though the MSE (as well as MAD and MAPE) is quite popular, at least two potential drawbacks of the criterion are worth mentioning. First, squared (or any other particular) error loss measures may not be closely related to other relevant cost functions. This is pointed out by Leitch and Tanner (1991), Stekler (1991), Diebold and Mariano (1994) and Swanson and White (1995), among others. Leitch and Tanner and Swanson and White find that the MSE is not closely related to profit measures in the context of an analysis of the term structure. Stekler examines methods for answering eight questions which can be asked about the quality of a set of macroeconomic forecasts, and discusses some of the criteria which we use here. Diebold and Mariano develop statistical tests for a broad class of error loss measures. A second potential drawback of the MSE is that it is not generally invariant to nonsingular scale-preserving linear transformations of the model in the context of multi-step forecasts or nonscalar processes, as pointed out by Clements and Hendry (1993) and discussed further in Clements and Hendry (1995). For these reasons, we also consider a number of other related (and unrelated) model selection criteria in our experiment. One closely related criterion which we calculate is Theil's U statistic, which is measured as:

$$U = \sqrt{\sum_{t=1}^{T} \frac{(\hat{y}_{I,t} - y_{I,t})^2}{(y_{I,t} - y_{I,t-h})^2}} ,$$

and can be viewed as the root MSE of a forecast divided by the root MSE of a naive no change forecast. The statistic takes the value 0 when the prediction is perfect, and unity when the MSE of the predicted change equals the MSE of the no change prediction. When comparing forecasts, Theil’s U is ordinally equivalent to the MSE selection criterion. Thus, in our context, Theil’s U only provides an alternative method by which the random walk model can conveniently be compared with our other methods based on the point MSE estimates. However, because Theil’s U has been reported in many past studies, we include it here.

An alternative model selection criterion, which may be particularly useful to market analysts trying to forecast the next economic turning point, for example, or to stock traders attempting to forecast the future price movements of particular stocks or futures contracts is the confusion rate. In our case, the confusion rate is calculated from a 2x2 contingency table. However, contingency tables can clearly be more complex. For detailed discussions of confusion rate measures, the reader is referred to Henriksson and Merton (HM, 1981), Schnader and Stekler (1990), Pesaran and Timmerman (1992, 1994b) and Stekler (1994).

An example of a confusion matrix taken from SW is

$$\begin{array}{ccc} & & \text{actual} \\ & & \begin{array}{cc} \textit{up} & \textit{down} \end{array} \\ \text{predicted} & \begin{array}{c} \textit{up} \\ \textit{down} \end{array} & \begin{bmatrix} 23 & 3 \\ 12 & 7 \end{bmatrix} \end{array} \tag{4}$$

The columns in (4) correspond to *actual* moves, up or down, while the rows correspond to *predicted* moves. In this way, the diagonal cells correspond to correct directional predictions, while off-diagonal cells correspond to incorrect predictions. We measure overall performance in terms of the model’s "confusion rate," the sum of the off-diagonal elements, divided by the sum of all elements. One question which we attempt to answer using our "least confused" models is whether such models are the same as those chosen as best based on other out-of-sample forecast performance measures such as the MSE, MAD and MAPE.

Related to the confusion matrix and to standard χ^2-test of independence in the context of confusion matrices (see discussion below), we report the ϕ coefficient. ϕ values are calculated as

$$\phi = \sqrt{\chi^2 / T}. \qquad (5)$$

For our contingency tables, in which the number of rows and columns are each two, ϕ ranges from 0 when the variables are independent to 1 when the variables are perfectly related, and is thus a measure which can be used in a loosely analogous way as the in-sample multiple coefficient of determination. It should be noted that the maximum value of ϕ exceeds unity when a contingency table contains more than two rows and two columns, and hence ϕ is used primarily with 2x2 tables. ϕ is sometimes referred to as a measure of the degree of diagonal concentration. This is because when two diagonally opposite cells are both identically 0, $\phi = 1$. One reason for reporting ϕ is that the value of our χ^2 statistic is directly proportional to the sample size, 45 in our case. Thus, the χ^2 test is particularly appropriate for ordinal rankings with T fixed. Overall, calculations using (5) provide a shortcut for determining how confused our models are, and for providing initial evidence as to whether the model is useful as a predictor of the sign of change in a particular macroeconomic variable.

Our final model selection criterion is an in-sample based complexity penalized likelihood measure. In particular, we report the geometric mean of the SIC (see Schwarz (1978)), which is hereafter referred to as the MSIC. For each of our 45 samples, the SIC is calculated as

$$\text{SIC} = \log s^2 + p(\log n)/n \ ,$$

where s^2 is the regression mean-squared-error and can be interpreted as a goodness of fit measure. The second term is the complexity penalty which depends on the number of parameters estimated, p, and on the window size (n=40, 58 or 76). However, as pointed out by Swanson and White (1996), the in-sample MSIC does not appear to offer a convenient shortcut to true out-of-sample performance measures for selecting econometric forecasting models, and as such is somewhat limited in the current context. We nevertheless report the MSIC values for a number of reasons. First, because our flexible specification models are chosen so as to minimize SIC values at each point in time, we expect that the MSIC values for our flexible specification models will be lower than those reported for our other forecasting models. Reporting the MSIC values allows us to confirm this expectation. Second, there are some cases where the flexible specification models do outperform the fixed specification models based on MSE, MAD, MAPE and Confusion Rate measures. As such, the evidence is not fully against using SIC measures as convenient shortcuts to selecting forecasting models. (One early treatment of in-sample model selection criteria and forecasting is given in Engle and Brown (1986).) In this vein, we include MSIC values at least partly because we feel that the use of in-sample model selection criteria in general (and not just the SIC) is still an open issue, and merits further attention.

7.2 Tests Based on Model Selection Criteria

There are a number of methods for testing which forecasting models are superior based on MSE. One such test is discussed in Granger and Newbold (1986). Assuming normality, and that the forecasts are unbiased, when $\varepsilon_{1,t} = \hat{u}_{I,t} - \hat{u}_{II,t}$ is contemporaneously uncorrelated with $\varepsilon_{2,t} = \hat{u}_{I,t} + \hat{u}_{II,t}$ the null hypothesis of equal forecast accuracy is the same as zero correlation between $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$, where I and II denote the two competing forecasting models. Setting $r = corr(\varepsilon_{1,t}\ , \varepsilon_{2,t})$, gives a standard z-statistic for correlation of the form: $z^{*} = \frac{1}{2}\log[(1 + r)/(1 - r)]$, where $(T-1)z^{*}$ is a standard normal random variable. Many generalizations of this basic test are available. For example, Meese and Rogoff (1988) modify the Granger-Newbold test by allowing for serial correlation, Mizrach (1991) further relaxes a Gaussianity assumption used in the Granger-Newbold and Meese-Rogoff tests, and Kolb and Stekler (1993) discuss a MSE regression test.

Here we use a test of equal forecast accuracy due to Diebold and Mariano (1995) which can easily be applied to a wide variety of criteria, including MSE, MAD, and MAPE. The "loss differential" test uses a sample path $\{d_t\}_{t=1}^{T}$ of a loss differential series, and is based on the following large sample statistic:

$$D \ = \ \bar{d} \,/\, [\, T^{-1} 2\pi \hat{f}_d(0) \,] \ \sim \ N(0,1)$$

where $\bar{d}$ is the sample average of d_t, and $2\pi f_d(0)$ is estimated in the usual way as a two-sided weighted sum of available sample autocovariances. In particular,

$$2\pi\hat{f}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} l\left[\frac{\tau}{S(T)}\right]\hat{\gamma}_d(\tau),$$

$$\hat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T}(d_t - \bar{d})(d_{t-|\tau|} - \bar{d}), \text{ and}$$

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(\hat{u}_{I,t}) - g(\hat{u}_{II,t})],$$

where $l\left[\frac{\tau}{S(T)}\right]$ is the lag window, $S(T)$ is the truncation lag, and examples of $g(\cdot)$ are given below. Following Diebold and Mariano's suggestion, we use a uniform lag window defined by

$$l\left[\frac{\tau}{S(T)}\right] = \begin{cases} 1 & \text{for } |(\tau/S(T))| \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

and assume $(h-1)$ dependence for our h-step ahead forecasts so that only $(h-1)$ sample autocovariances need be used in the estimation of $f_d(0)$ and $S(T)=(h-1)$. However, it should be noted that this assumption can be relaxed, and parametric methods can be used to estimate $f_d(0)$, for example. One appealing attribute of the loss differential test is that the asymptotic normality of D requires only that d_t is covariance stationary and short memory. We define our loss differential series (i.e. $g(\cdot)$) as

$d_t = \hat{u}^2_{I,t} - \hat{u}^2_{II,t}$, for the MSE test;

$d_t = |\hat{u}_{I,t}| - |\hat{u}_{II,t}|$, for the MAD test; and

$d_t = |(\hat{y}_{I,t}/y_t) - 1| - |(\hat{y}_{II,t}/y_t) - 1\ |$, for the MAPE test.

We also consider a nonparametric test which does not assume normality and essentially requires only that the distribution of d_t is continuous. In particular, Wilcoxon signed rank test statistics (see Bickel and Doksum (1977)) are calculated by pooling the results of various selection criteria (MSE, MAD and MAPE) across variables. This is meant to provide an overall flavor for the usefulness of one type of econometric model versus another, independent of which macroeconomic variable is used, and acts as a further guide for comparing the potential benefits of our various forecasting models.

Assume that we wish to compare the performance of the forecasts of each of our five final forecasting models with each other. We assume that we have independent pairs (i.e one pair is the MSE for the random walk model and the MSE for the random walk with drift model for a single variable over the entire forecast period) corresponding to the nine macroeconomic variables. Then we construct the difference between our "control" (the random walk model MSE values), and our "treated" model (the random walk with drift model MSE values). In this scenario we have ten different control-treatment combinations corresponding to the ten different pairwise model comparisons. The null hypothesis is of no "treatment effect", assuming that the differences between the MSEs for each variable are symmetrically distributed about 0. Define $Z_i = MSS_{I,i} - MSS_{II,i}$, i=1,...,NV, where MSS is the value of the particular model selection statistic being examined, and NV is nine. The Wilcoxon signed rank test uses the signs of the Z_i, and the ranks of the $|Z_i|$, and is thus a more sensitive distribution-free test than the sign test, for example. The signed ranks preserve the order relationship on each side of 0 and with respect to 0. Let R_i be the signed rank of Z_i. Then for example, if $Z = (200, -300, 400, 500, -600)$, we have $R = (1, -2, 3, 4, -5)$. Assuming that the population distribution of the R_i, say F, is continuous and symmetric about 0, then

$$W = ½\sum_{i=1}^{NV} R_i + \frac{NV(NV+1)}{4} = \sum_{i=1}^{S} T_i \ , \tag{6}$$

where T_i are the ordered positive R_i and S is the observed number of positive Z_i. The distribution of W in (6) can be approximated as $N(0,1)$ with approximate critical values for $NV > 16$ as follows:

$$\frac{1}{4}NV(NV+1) + \left[\frac{1}{24}NV(NV+1)(2NV+1)\right]^{½} z(1-\alpha),$$

after appropriately standardizing W using the variance given by Var(W) = $\frac{1}{24}NV(NV+1)(2NV+1)$. The distribution of W for cases where $NV \le 16$ can be tabulated by noting that

$$P[T_1 = t_1, \ldots, T_s = t_s, S = s] = \frac{1}{2^{NV}} \ ,$$

where $(t_1, \ldots, t_s)$, $t_1 < \cdots < t_s$ is the set of possible values of $(T_1, \ldots, T_s)$.

Turning now to our confusion rate criterion, recall that since (4) is simply a 2×2 contingency table, the hypothesis that a given model/window combination is of no value in forecasting the direction of spot rate changes can be expressed as the hypothesis of independence between the actual and predicted directions (as discussed in Pesaran and Timmerman (1994b) and Stekler (1994)). Pesaran and Timmerman (1994b) show that the test of market timing (in the context of forecasting the direction of asset price movements) proposed by HM is asymptotically equivalent to the standard χ^2-test of independence in a 2×2 contingency table, when the column and row sums are not *a priori* fixed (in our analysis the column and row sums are not fixed). When the column and row sums are fixed, Pesaran and Timmerman (1994b) further show that the HM-test of market timing is better interpreted as an exact test of independence within the framework of a 2×2 contingency table. We examine confusion matrices, confusion rates, and both the HM p-values and the standard χ^2-test of independence p-values. It should be noted that a finding that a model rejects the null hypothesis of independence is direct evidence that the model is useful as a predictor of the sign of change in a particular macroeconomic variable. The χ^2-test of independence is calculated as

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}} ,$$

where f_{o_i} is the observed number in a given cell of (4), f_{e_i} is the expected number in a given cell of (4), and k is the number of cells, 4 in our 2x2 case. In our example, for the uppermost left cell (with entry 23) we have that (i) the probability of *actual up* is 35/45=0.78, (ii) the probability of *predicted up* is 26/45=0.58, (iii) the joint probability of *actual up* and *predicted up* is 0.78x0.58=0.45, (iv) hence the expected number for the upermost left cell is 20.4.

Using all of the above tests, we compare each of our models in pairwise fashion, so that statistics are calculated for each of ten different pairwise "competitions". In this manner, we compare (i) linear fixed specification and linear flexible specification models, (ii) linear fixed specification and nonlinear flexible specification models and (iii) linear flexible specification and nonlinear flexible specification models. These comparisons are done on a variable by variable basis for each of the 9 macroeconomic series.

8. Experimental Results

8.1 Model Comparison Results Using MSE, MAD, MAPE and Confusion Rate

A number of interesting results emerge when comparing the random walk (L1), random walk with drift (L2), "best" linear (L3), flexible specification linear (L4) and flexible specification nonlinear (NN) forecasting models. Table 1 reports the results of loss differential tests based on MSE. (Complete empirical results are available upon request from the authors - e.g. an earlier version of the paper included loss differential test results for MSE, MAD, and MAPE.) There are 10 columns of p-values corresponding to ten pairwise model comparisons under the null hypothesis of equal forecast accuracy. For example, column 1 reports results for the comparison of NN (the ANN model) versus L1 (the random walk without drift model). For the first variable, unemployment (U), and for h=1, the p-value is 0.061. Thus, at a 6.1% level of significance we reject the null hypothesis that NN and L1 are equally accurate when used to forecast U, based on the MSE criterion. The "winning" model appears in brackets beside any p-values which are less than 0.10. Thus, in our example, L1 is seen to outperform NN for forecasting U based on the MSE. Whenever the flexible specification nonlinear model chose no hidden units for any of the 45 forecasts in the ex ante forecast period, then the L4 and NN results are necessarily identical. For these cases, and for all related duplicate cases (e.g. when L2 versus L4 is exactly the same as L2 versus NN), dashes appear in place of the usual p-values.

Upon examination of the MSE (as well as MAD and MAPE) -best models based on loss differential test p-values the following models appear to perform well against their competitors in the current context. Unemployment and interest rates appear to be well characterized as random walks at forecast horizon of h =1, although there is little to choose among the 5 competing forecasting models for these two variables at a 1-year forecast horizon. However, when examining MAD and MAPE (results not included here), the interest rate variable appears best characterized as a random walk at both forecast horizons. In particular, trending variables are generally characterized better as random walks with drift than as random walks (not surprisingly). It should be reiterated, though, that comparisons based on other criteria such as confusion rates may result in the choice of a different model. For instance, Swanson and White (1995) find that artificial neural network models are often least confused when used to forecast the direction of change of the spot interest rate. At the opposite end of the spectrum from interest rates is industrial production, where the random walk model is invariably beaten by all of L2, L3, L4 and NN, with little to choose between the latter models at h=1 and h=4. [7] However, at h=4 the flexible specification nonlinear model is often beaten, while the flexible specification linear model remains a strong candidate for providing the best forecasting model. At forecast horizons h=1,4, both of the GNP variables generally favor L3, L4 and NN, with L4 appearing to be a slightly stronger candidate for overall "winner" than L3 and NN. For corporate profits variable (Π), the random walk model seems to perform particularly poorly, while it is difficult to choose between L3 and L4 when attempting to specify a "best" model. For the change in business inventories variable there is little to choose among the models based on MSE, although the flexible specification linear model clearly dominates at both forecast horizons when MAD and MAPE loss differential test results are examined. Overall, then, there are cases where each of the models "wins". Nevertheless, based on the loss differential test statistic for MSE, MAD, and MAPE, the flexible specification linear and nonlinear models seem to be beating the rest of the models around 50% of the time. The exceptions to this are interest rates, net exports and unemployment.

Table 2 reports Wilcoxon ranked sign test statistic values, associated p-values, and lists the "winning" model in brackets for p-values of less than 0.25. (Analogous results based on the sign test are available from the authors.) This rather high p-value was used because we have only nine matched pairs (corresponding to the nine variables) in our sample. Thus, based on the binomial distribution, it would be very difficult to reject the null hypothesis (that the model perform equally well) any of the time at a 5% level, say, as statistic values of only 1 or 9 would suffice. Of course, observing a statistic value of 1 or 9 in our context would be startling, given the diverse nature of our group of variables. Furthermore, since

[7] For h=4, this assertion is made because L2,L3,L4, and NN "win" based on MAD and MAPE, even though only L2 and L3 "win" based on MSE.

we are attempting with these tests only to capture an overall feel for which models are performing well, setting the size of the test at 0.25 does not seem unreasonable. At the 1-quarter forecast horizon, there is little to choose between the models based on the MSE, although L4 and L2 appear to be overall winners based on MAD and MAPE, respectively. This result is only loosely consistent with the results of the loss differential tests, and can probably be attributed to the "pooled" nature of the ranked sign test. Based on a 1-year forecast horizon, NN and L2 perform well based on MAPE; L3 is the "winner" based on MAD; and L2, L3, L4, and NN are all "winners" based on MSE. These results, then, suggest that flexible specification models are useful (with NN models being more useful at longer forecast horizons), although each model performs well in a number of cases based on the different criteria examined.

In order to look at our models from another perspective, we also compared them based on simple point estimates of MSEs, MADs, MAPEs and Confusion Rates. While this approach does not constitute any sort of valid statistical test, it does help to shed light on the relative merits of the alternative forecasting models. In Table 3, "winners" from a competition between the flexible specification (L4 and NN) and the fixed specification (L1, L2 and L3) models are reported. The table clearly suggests that the flexible specification models are particularly good at the h=4 forecast horizon, forecasting more accurately and being less confused than their fixed specification counterparts for the majority of variables (including U, NGNP, RGNP, PCE, ΔBI and Net X). This is contrary to the h=1 case where flexible specification models only clearly dominate for the GNP variables, and are partial winners for R, IP and ΔBI (with a confusion rate tie for PCE). As suggested above, the flexible specification models appear to show some promise in the context of multi-step forecasting. However, even in the h=1 case, there is evidence that flexible specification models need to be considered when constructing economic forecasts. One avenue for further research in this respect is the construction of more complex and varied flexible specification models (i.e. which are not only chosen using an in-sample SIC criterion) using other means of adaptation and encorporating other forms of nonlinearity, for example. As an illustration, one alternative type of nonlinear specification involves estimating nonlinear cointegrating relationships when constructing the forecasting model (see Corradi, Swanson, and White (1995) and Granger and Swanson (1996)).

Because the flexible specification models were both picked in the same way (using an in-sample SIC), and because the flexible specification nonlinear model has a generally nonlinear functional form, even when 0 hidden units are selected, we also asked the following question based on the results of Table 3: "Does the flexible specification nonlinear forecasting model outperform the "best" fixed specification linear model based on point estimates, regardless of how many hidden units are selected?" (These results summarize various statistic values presented in Tables 4-5.) Counting up the total number of "wins" across MSE, MAD, MAPE and confusion rate criteria, we see that for h=1 the linear models "win" approximately 62% of the time, while for h=4 the nonlinear model "wins" around 61% of the time. Based on these results it appears that *both* the linear and the nonlinear models have something to offer (with the nonlinear model slightly preferred for h=4 and the linear model slightly preferred for h=1), and that econometric forecasting models should be constructed by entertaining both linear and nonlinear models, with the final outcome clearly being dependent on which particular macroeconomic variable is being modeled. Put another way, in the current context the artificial neural networks appear useful; however, they do not supplant more straightforward linear models, and further analysis of both types of models is warranted.

8.2 Choosing a Forecasting Model

Tables 4-5 summarize the experimental results for L1-L4 and NN by tabulating all of the model selection criteria for each variable and forecast horizon. In addition to MSE, MAD, MAPE and confusion rates, MSIC, ϕ, and Theil's U statistics are also tabulated. In general, each of the above criteria has something to offer when "choosing" a final forecasting model. Of course, there is no clear cut answer to which model selection criterion is the "most" useful, as that depends on the user's loss (or cost) function. Nevertheless, the following rough guidelines are useful. The MSIC is our only in-sample criterion, and since it is minimized for the nonlinear models (by construction), it would be the clear squared-error loss based criterion of choice, if the nonlinear model were always the winner out-of-sample. However, since the nonlinear model wins only about 50% of the time, the MSIC is not always selecting the best forecasting model. For this reason, the other squared error loss based criteria (MSE, MAD, MAPE, and Theil's U) should also be considered. Since Theil's U only serves to summarize the MSE statistics, it does not provide much extra information. However, if the only goal of the analysis is to determine whether or not it is possible to construct a model which outperforms the simple random walk model (based on squared error loss), then Theil's U provides a convenient short-cut to directly reporting the forecasting results from all models. (Note that the Theil's U statistic is always unity (by definition) for the random walk without drift models reported in Tables 4-5.) In Table 4 we see that the random walk with drift performs worse than the random walk when forecasting unemployment (Theil's U value for the random walk with drift model for h=1 is 1.049). However, the "best" linear, as well as both flexible specification models forecast unemployment better than the "no change" or random walk model, based on point estimates. Thus, based on Theil's U we might conclude that the random walk model is less appropriate for forecasting unemployment, as a number of MSE-better models are available.

As discussed above, however, squared error loss functions are not always of interest to forecasters. The confusion rate and ϕ values offer measures which are of interest when the objective is to forecast the direction of change in the variable. As an example of how choosing a model based on this criterion can differ from choices based on squared-error loss, consider the case of real GNP (see Table 4). Note that for h=1 the least confused model based on both the confusion rate and the ϕ coefficient is the flexible specification linear model, which is confused around 11% of the time. However, Theil's U statistic, the MSE, and the MAD statistics are all lower for the "best" linear model. Thus, while the "best" linear model is preferred based on squared-error loss, the flexible specification model is preferred based on ability to forecast direction of change. This suggests that the nonparametric tests discussed above may provide a guide as to which models perform well overall, but clearly do not provide convincing evidence as to which forecasting model will prevail on a case by case basis. Perhaps the best approach to this issue is to entertain an appropriate variety of selection criteria, noting any *patterns* of "winners" and "losers" which emerge, and weighting more heavily those criteria which are most closely related to some posited loss function or functions.

8.3 Model Adaptation and Evolution Results

In order to examine the parameter evolution and model adaptation of our various forecasting models, the estimated coefficients of each of the models were tracked throughout the sequence of 45 quarterly 1 and 4-step ahead forecasts. Parameter evolution arises because each model is re-estimated at each period of time, allowing the parameters to change before each new forecast is made. Both the flexible specification and fixed specification models are allowed to evolve over time. However, only the flexible specification models are allowed to "select" new specifications at each point in time. Panel A of Figure 1 provides plots of various parameters as they evolved over the forecast period. It is apparent that the most of the parameters plotted are changing over time. In this sense, the plots are indicative of the substantial evolution of parameter estimates in virtually all of the fixed specification models which we examined. This in turn suggests that many of these forecasting models are taking advantage of the evolutionary aspect of our estimation strategy. For example, drift and slope parameters tend to vary over time in a relatively erratic fashion, and do not appear to be "evolving" to some final fixed values. This can be taken as rather "loose" evidence that fixing the parameters of a forecasting model at the outset, and then constructing a sequence of 1 or 4-step ahead forecasts (as new data become available) may be sub-optimal in the sense of providing relatively inferior forecasts. This feature is emphasized by looking at a second feature of the plots in Figure 1. For the Π and GNP variables, the plotted parameter estimates are increasing steadily over time. In particular for the GNP case, the slope coefficient estimate is seen to be increasing more or less continually from 1985 to 1993. One of the reasons for the rich parameter evolution may be that fixed specification linear models have a tendency to become misspecified over time, as the economy changes. As possible evidence of model misspecification of this type we also examined the parameter evolution of the coefficients from the flexible specification models. To illustrate our findings, Panel B of Figure 1 contains plots of various lagged dependent variable coefficient values. The coefficients of the variables shown are clearly evolving over time. However, the coefficients do not appear to be evolving to such a great extent as those in the fixed specification models. In particular, in all 4 cases shown the plotted parameters exhibit rather long stable periods, where their values are seen to change little over time. Thus, it appears that flexible specification models may be less susceptible to evolutionary change than fixed specification models. This feature is perhaps tied in with the above discussion concerning the possible tendency of fixed specification models to become misspecified over time, and suggests that flexible specification models may be useful for the purpose of constructing forecasting models.

As a final question with regard to model adaptation, we ask: "Are the flexible specification models apt to choose large numbers of regressors, or to frequently change the number of regressors used?" If so we have evidence that, for example, the economy is undergoing rapid change, or the fixed specification models are overfitting the data, or that our flexible specification estimation strategy is sensitive to changing economic conditions. Panel C of Figure 1 plots the number of parameters (including the constant term) selected by various flexible specification models over time. From the plots, it appears that the answer to our question is no. In all cases shown, the number of selected parameters never exceeds nine, and is usually around five. Furthermore, the variation in the number of parameters selected is rather small, in particular for the Real GNP case, where 4 parameters are chosen throughout the 1982-1990 forecast period. In summary, our primary method for comparing forecasting models has been the use of model selection criteria and tests. As such, the above discussion of parameter evolution and model flexibility is not only rather subjective in nature, and is meant to be used only as an aid to understanding our model selection results. In light of this, we conclude by suggesting that there is some evidence supporting our model selection based finding that flexible linear and nonlinear models are potentially useful for forecasting, relative to simpler and fixed specification models.

9. Conclusions

Nine macroeconomic variables have been forecast using a variety of flexible specification and fixed specification econometric models, both linear and nonlinear. We have attempted to examine the usefulness of flexible specification and nonlinear forecasting models relative to more commonly used fixed specification and linear models, such as the random walk, random walk with drift and unrestricted vector autoregressive (estimated using ARX specifications) models. All models have been allowed to evolve through time, and our analysis focuses on real-time model selection and performance. In closing, we offer the following conclusions.

First, in the context of real-time forecasts, flexible specification and nonlinear models appear to offer a useful and viable alternative to less flexible fixed specification linear models. In particular at forecast horizons greater than one step ahead there seems to be some potential for improving macroeconomic forecasts using flexible specification econometric models, which have the feature that the model "specification" is allowed to vary over time, as new information becomes available. This model adaptation has the potentially useful feature of being able to capture shifts in the relationships among economic variables relatively quickly, and in a relatively painless manner computationally.

Second, while all (linear and nonlinear, flexible specification and fixed specification) models are re-estimated at each point in time, we provide initial evidence that the evolution of coefficients estimated using more rigid fixed specification models is somewhat more erratic than in the case of flexible specification models. We speculate that one reason for this result is that the economy is evolving (rather slowly) over time. This feature cannot easily be captured by fixed specification linear models, however, and manifests itself in the form of quickly evolving coefficient estimates. In this sense, it appears that flexible specification models have much to offer when forecasting economic time series, although smooth transition autoregressions and related nonlinear models are not examined in this paper.

A third result of our analysis is that while the flexible specification models perform well based on a number of model selection criteria (including mean squared forecast error and confusion rate), their nonlinear counterparts are somewhat less successful (although the artificial neural networks (ANNs) still dominate other models in a number of cases). This mixed evidence with regard to the ANN models suggests that further research using generalizations of the models considered here may be useful. For example, selecting networks based on model selection criteria other than the the Schwarz information criterion (SIC), using cross-validation techniques, and including more than one "hidden layer" in nonlinear architectures may be of potential interest, especially given the relative inability of the SIC to consistently pick optimal forecasting models based on out-of-sample model selection criteria.

Our fourth and last conclusion concerns how to "choose" a forecasting model. As is well known, models which "win" based on one model selection criterion (say a squared-error measure) do not necessarily win when an alternative selection criterion is used (say a confusion rate measure). This points to the need for the applied practitioner to carefully define a cost function prior to choosing a final forecasting model for a particular time series. Although cost functions based on error-loss are often used, other measures based on market timing and profitability, for example, are becoming increasingly more popular, as forecasters realize that model selection depends (sometimes crucially) on which model selection criteria are used. Overall, our results suggest that a variety of linear and nonliner models should be initially entertained in any forecasting exercise, and that final model selection based on a careful analysis of a number of model selection criteria is probably a good starting point. In particular, we were not surprised to find that each of the five alternative models (random walk, random walk with drift, "best" linear vector autoregression, flexible specification linear and flexible specification artificial neural network) is our chosen "winner" for one or another of the macroeconomic variables, and for one or another of our model selection criteria.

Acknowledgements:

Many thanks to an associate editor, 3 referees, John Moody, Jeff Racine, James B. Ramsey and H.O. Stekler for insightful comments and suggestions, and to Dina Nunez-Rocha and Tian Zeng for excellent research assistance. Parts of this paper were presented at the 1994 International Symposium of Forecasters in Stockholm, Sweden, and at the 1995 Symposium for Nonlinear Dynamics and Econometrics in New York, New York. Research was supported by NSF Grant SES #9209023, and by the Research and Graduate Studies Office at Penn State University.

References

Bickel, P.J. and Doksum, K.A., 1977, *Mathematical Statistics*, Englewood Cliffs, New Jersey: Prentice Hall.

Boschen, H.I. and Grossman, J.F., 1982, "Tests of Equilibrium Macroeconomics Using Contemporaneous Monetary Data," *Journal of Monetary Economics*, 10, pp. 309-333.

Brock, W.A., Hsieh, D.A. and LeBaron, B., 1991, *Nonlinear Dynamics, Chaos and Instability*, Cambridge Massachusetts: The MIT Press.

Carroll, S.M. and Dickinson, B.W., 1989, "Construction of Neural Nets Using the Radon Transform," in *Proceedings of the International Joint Conference on Neural Networks*, Washington DC. New York: IEEE Press, pp. 607-611.

Clements, M.P. and Hendry, D.F., 1993, "On the Limitations of Comparing Mean Square Forecast Errors," *Journal of Forecasting*, 12, pp. 617-37.

Clements, M.P. and Hendry, D.F., 1995, "Forecasting in Cointegrated Systems," *Journal of Applied Econometrics*, 10, pp. 127-46.

Corradi, V., Swanson, N.R. and White, H., 1995, "Testing for Stationarity-Ergodicity and for Comovements Between Discrete Time Markov Processes", mimeo, The University of Pennsylvania.

Croushore, D., 1993, "Introducing: The Survey of Professional Forecasters," *Business Review*, The Federal Reserve Bank of Philadelphia, November-December, pp. 3-15.

Diebold, F.X. and Mariano, R.S., 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, pp. 253-263.

Dorsey, R.E., Johnson, J.D. and van Boening, M.V., 1994, "The Use of Artificial Neural Networks for Estimation of Decision Surfaces in First Price Sealed Bid Auctions," in W.W. Cooper and A. Whinston eds., *New Directions in Computational Economics.* Boston: Kluwer, pp. 19-40.

Engle, R.F. and Brown, S.J., 1986, "Model Selection for Forecasting," *Applied Mathematics and Computation*, 20, pp. 313-327.

Engle, R.F. and Granger C.W.J., 1987, "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, pp. 251-76.

Fair, R.C. and Shiller, R.J., 1990, "Comparing Information in Forecasts from Econometric Models," *American Economic Review*, 80, pp. 375-89.

Franses, P.H., 1995, "Recent Advances in Modelling Seasonality," *Journal of Economic Surveys*, 10, pp. 299-345.

Funahashi, K., 1989, "On the Approximate Realization of Continuous Mappings by Neural Networks," *Neural Networks*, 2, pp. 183-92.

Ghysels, E., Granger, C.W.J. and P. Siklos, 1996, "Is Seasonal Adjustment a Linear or Nonlinear Data-Filtering Process?," *Journal of Business and Economic Statistics*, 14, pp. 386.

Granger, C.W.J., 1993, "Strategies for Modeling Nonlinear Time Series Relationships," *The Economic Record*, 69, pp. 233-8.

Granger, C.W.J. and Newbold, P., 1986, *Forecasting Economic Time Series*, San Diego: Academic Press.

Granger, C.W.J. and Swanson, N.R., 1996, "Further Developments in the Study of Cointegrated Economic Variables," *Oxford Bulletin of Economics and Statistics*, 58, pp. 537-551.

Granger, C.W.J. and Teräsvirta, T., 1993, *Modelling Nonlinear Economic Relationships*, New York: Oxford.

Häfke, C. and Helmenstein, C., 1994, "Neural Networks in Capital Markets: An Application to Index Forecasting," *International Journal of Forecasting*, forthcoming.

Henriksson, R.D. and Merton, R.C., 1981, "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills," *Journal of Business*, 54, pp. 513-33.

Hoffman, D.L. and Rasche, R.H., 1996, "Assessing Forecast Performance in a Cointegrated System," *Journal of Applied Econometrics*, 11, pp. 495-517.

Hornik, K., Stinchcombe, M. and White, H., 1989, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, pp. 359-66.

Hornik, K., Stinchcombe, M. and White, H., 1990, "Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks," *Neural Networks*, 3, pp. 551-60.

Johansen, S., 1988, "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, 12, pp. 231-54.

Johansen, S. and Juselius, K., 1990, "Maximum Likelihood Estimation and Inference on Cointegration - with Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, pp. 169-210.

Kolb, R.A. and Stekler, H.O., 1993, "Are Economic Forecasts Significantly Better Than Naive Predictions? An Appropriate Test," *International Journal of Forecasting*, 9, pp. 117-20.

Kuan, C.-M. and White, H., 1994, "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews*, 13, pp. 1-91.

Leitch, G. and Tanner, J.E., 1991, "Economic Forecast Evaluation: Profits Versus the Conventional Error Measures," *American Economic Review*, 81, pp. 580-90.

Lin, J.-L. and Tsay, R.S., 1996, "Co-integration Constraint and Forecasting: An Empirical Examination," *Journal of Applied Econometrics*, 11, pp. 519-538.

Maasoumi, E., Khotanzad, A. and Abaye, A., 1994, "Artificial Neural Networks for Some Macroeconomic Series: A First Report," *Econometric Reviews*, 13, pp. 105-22.

Maravall, A. and Pierce, D.A., 1983, "Preliminary-Data Error and Monetary Aggregate Targeting," *Journal of Business and Economic Statistics*, 2, pp. 337-339.

Mariano, Roberto S. and Hisashi Tanizaki, 1994, "Prediction of Final Data with Use of Preliminary and/or Revised Data," *Journal of Forecasting*, forthcoming.

Meese, R.A. and Rogoff, K., 1988, "Was it Real? The Exchange Rate - Interest Differential Relation over the Modern Floating-Rate Period," *Journal of Finance*, 43, pp. 933-48.

Mizrach, B., 1991, "Forecast Comparison in L_2," mimeo, Department of Finance, Wharton School, University of Pennsylvania.

Moody, J. and Utans, J., 1991, "Principled Architecture Selection for Neural Networks: Applications to Corporate Bond Rating Predictions," in J.E. Moody, S.J. Hanson and R.P. Lippmann, eds., *Advances in Neural Information Processing Systems* 4. San Mateo: Morgan Kaufman, pp. 683-690.

Moody, J., Levin, U. and Rehfuss, S., 1993, "Predicting the U.S. Index of Industrial Production," *Neural Network World*, 3, pp. 791-4.

Mulhern, F., Caprara, R.J., 1994, "A Nearest Neighbor Model for Forecasting Market Response," *International Journal of Forecasting*, 10, pp. 191-207.

Patterson, K.D., 1995, "An Integrated Model of the Data Measurement and Data Generation Processes with an Application to Consumers' Expenditure," *The Economic Journal*, 105, pp. 54-76.

Pesaran, M.H. and Timmerman, A.G., 1992, "A Simple Nonparametric Test of Predictive Performance," *Journal of Business and Economic Statistics*, 10, pp. 461-5.

Pesaran, M.H. and Timmerman, A.G., 1994a, "The Use of Recursive Model Selection Strategies in Forecasting Stock Returns," mimeo, University of California, San Diego.

Pesaran, M.H. and Timmerman, A.G., 1994b, "A Generalization of the Non-Parametric Henriksson-Merton Test of Market Timing," *Economics Letters*, 44, pp. 1-7.

Ramsey, J.B., 1996, "If Nonlinear Models Cannot Forecast, What Good Are They?," *Studies in Nonlinear Dynamics and Econometrics*, 1, pp. 65-86.

Rissanen, Jorma, 1978, "Modeling by Shortest Data Description," *Automatica*, 14, pp. 465-471.

Schnader, M.H. and Stekler, H.O., 1990, "Evaluating Predictions of Change," *The Journal of Business*, 63, pp. 99-107.

Schwarz, G., 1978, "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, pp. 461-4.

Soni, T., Otruba, H., Häfke, C. and Natter, M., 1995, "Can Neural Networks Capture Stylized Facts in Macroeconomic Time Series?," mimeo, Institute for Advanced Studies Vienna.

Stekler, H.O., 1991, "Macroeconomic Forecast Evaluation Techniques," *International Journal of Forecasting*, 7, pp. 375-84.

Stekler, H.O., 1994, "Are Economic Forecasts Valuable?" *Journal of Forecasting*, 13, pp. 495-505.

Swanson, N.R., 1996, "Money an Output Viewed Through a Rolling Window," Discussion Paper, Pennsylvania State University.

Swanson, N.R. and White, H., 1995, "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks," *Journal of Business and Economic Statistics*, 13, pp. 265-275.

Swanson, N.R. and White, H., 1996, "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," *Review of Economics and Statistics*, forthcoming.

Thoma, M.A., 1994, "Subsample Instability and Asymmetries in Money-Income Causality," *Journal of Econometrics*, 64, pp. 279-306.

White, H., 1988, "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns," in *Proceedings of the IEEE International Conference on Neural Networks*, San Diego. New York: IEEE Press, pp. 451-458.

White, H., 1989, "Learning in Artificial Neural Networks: A Statistical Perspective," *Neural Computation*, 1, pp. 425-64.

Biography:

Norman R. Swanson is assistant professor at Pennsylvania State University. He received his PhD from the University of California, San Diego in 1994. His research interests include time series methodology, forecasting, and nonlinearity. Halbert White is professor at the University of California, San Diego. He received his PhD from MIT in 1976. He currently serves on numerous editorial and advisory boards. His research interests include econometric theory, forecasting, neural networks, and financial markets.

fillllll

fillllll

Table 3: Winners from a Point Estimate Comparison of Various Models by Selection Criterion[1]

Summary of Results Tables 4–5

Table 3a: Flexible Versus Fixed Specification: *h*=1 (*h*=4)

Variable	Selection Criterion			
	MSE	MAD	MAPE	Confusion Rate
U	Fixed (Flex)	Fixed (Flex)	Fixed (Flex)	Fixed (Flex)
R	Fixed (Fixed)	Fixed (Fixed)	Fixed (Fixed)	Flex
IP	Fixed (Fixed)	Fixed (Fixed)	Flex (Fixed)	Fixed (Fixed)
NGNP	Flex (Flex)	Flex (Flex)	Flex (Flex)	----
Π	Fixed (Fixed)	Fixed (Fixed)	Fixed (Fixed)	Fixed (Fixed)
RGNP	Flex (Flex)	Flex (Fixed)	Fixed (Flex)	Fixed (Fixed)
PCE	Fixed (Flex)	Fixed (Flex)	Fixed (Flex)	----
ΔBI	Fixed (Flex)	Fixed (Flex)	Flex (Flex)	Fixed (Flex)
Net X	Fixed (Flex)	Fixed (Flex)	Fixed (Flex)	Fixed (Flex)

Table 3b: Linear Versus Nonlinear Specification: *h*=1 (*h*=4)

Variable	Selection Criterion			
	MSE	MAD	MAPE	Confusion Rate
U	Linear (NonLin)	Linear (NonLin)	Linear (NonLin)	Linear (NonLin)
R	Linear (Linear)	Linear (Linear)	Linear (Linear)	Nonlin (----)
IP	Linear (Linear)	Linear (Linear)	Nonlin (Linear)	Linear (NonLin)
NGNP	Nonlin (NonLin)	Nonlin (NonLin)	Nonlin (NonLin)	----
Π	Nonlin (Linear)	Nonlin (Linear)	Nonlin (Linear)	Linear (Linear)
RGNP	Nonlin (Linear)	Nonlin (Linear)	Nonlin (NonLin)	Nonlin (NonLin)
PCE	Linear (NonLin)	Linear (NonLin)	Linear (NonLin)	----
ΔBI	Linear (NonLin)	Linear (NonLin)	Nonlin (Linear)	Linear (NonLin)
Net X	Linear (NonLin)	Linear (NonLin)	Linear (NonLin)	Linear (NonLin)

[1] **The table summarizes the "winners" (based on point estimates) for all variable by forecast horizon (*h*), and for the out-of-sample model selection criteria as given. All statistics are calculated using the *true* ex-post observation period from 1982:3-1993:3. In the case of ties, dashes are shown in place of the "winner".**

Table 4: Summary Model Selection Statistics: Forecast Horizon, h=1[1]

$$dep_{t+h-1} = \alpha + \sum_{i=1}^{K1} \beta_i \, dep_{t-i} + \sum_{i=1}^{K2} \delta_i \, ind1_{t-i} + \sum_{i=1}^{K3} \gamma_i \, ind2_{t-i} + u_{t+h-1}$$

Var	Mod	MSIC	MSE	MAD	MAPE	CM	CR	HM	χ^2	ϕ	TU
U	L1	-1.531	0.102	0.241	3.323	-	-	-	-	-	1.000
(2,0,2)	L2	-1.538	0.113	0.256	3.534	13,25,4,1	0.674	0.994	0.138	0.226	1.049
[OW=76]	L3	-1.857	0.072	0.210	2.999	10,9,7,17	0.372	0.106	0.212	0.191	0.841
HU:2%	L4	-2.225	0.095	0.238	3.348	10,11,7,15	0.419	0.228	0.455	0.114	0.963
	NN	-2.224	0.095	0.238	3.341	10,11,7,15	0.419	0.228	0.455	0.114	0.963
R	L1	-1.268	0.284	0.396	3.919	-	-	-	-	-	1.000
(1,0,0)	L2	-1.286	0.314	0.423	4.170	13,30,0,2	0.667	0.501	0.901	0.019	1.052
[OW=76]	L3	-1.251	0.321	0.430	4.131	13,32,0,0	0.711	1.000	0.870	0.025	1.064
HU:0%	L4	-1.366	0.388	0.501	4.936	8,20,5,11	0.568	0.705	0.877	0.023	1.169
	NN	-1.366	0.388	0.501	4.936	8,20,5.11	0.568	0.705	0.877	0.023	1.169
IP	L1	1.250	2.191	1.183	1.223	-	-	-	-	-	1.000
(2,2,0)	L2	1.160	1.783	1.014	1.054	34,11,0,0	0.244	1.000	0.863	0.026	0.902
[OW=76]	L3	0.967	1.544	0.987	1.238	23,6,11,5	0.378	0.330	0.670	0.064	0.840
HU:0%	L4	0.677	2.011	1.186	1.172	19,4,15,7	0.422	0.219	0.436	0.116	0.958
	NN	0.677	2.011	1.186	1.172	19,4,15,7	0.422	0.219	0.436	0.116	0.958
NGNP	L1	8.167	6664	74.44	1.631	-	-	-	-	-	1.000
(4,0,4)	L2	6.956	1807	32.91	0.945	45,0,0,0	0.000	1.000	0.940	0.011	0.521
[OW=76]	L3	6.782	1789	33.25	0.762	45,0,0,0	0.000	1.000	0.938	0.011	0.518
HU:0%	L4	6.543	1652	30.27	0.685	45,0,0,0	0.000	1.000	0.938	0.011	0.498
	NN	6.543	1652	30.27	0.685	45,0,0,0	0.000	1.000	0.938	0.011	0.498
Π	L1	4.446	146.6	9.373	5.779	-	-	-	-	-	1.000
(5,0,5)	L2	4.420	137.9	8.994	5.597	31,14,0,0	0.311	1.000	0.872	0.024	0.970
[OW=76]	L3	4.693	177.6	10.33	6.572	23,9,8,5	0.378	0.367	0.746	0.048	1.100
HU:0%	L4	4.411	149.1	9.154	5.712	20,11,11,3	0.489	0.904	0.552	0.089	1.008
	NN	4.411	149.1	9.154	5.712	20,11,11,3	0.489	0.904	0.552	0.089	1.008
RGNP	L1	7.831	2033	38.04	0.857	-	-	-	-	-	1.000
(4,0,4)	L2	7.605	1156	25.47	0.573	40,5,0,0	0.111	1.000	0.940	0.035	0.754
[OW=58]	L3	7.719	1805	32.27	0.734	35,5,5,0	0.222	1.000	0.933	0.013	0.942
HU:20%	L4	7.126	1050	26.66	0.589	35,4,5,1	0.200	0.529	0.816	0.035	0.719
	NN	7.139	994.1	25.98	0.576	35,4,5,1	0.200	0.529	0.816	0.035	0.699
PCE	L1	6.946	1295	30.51	1.019	-	-	-	-	-	1.000
(1,0,1)	L2	6.585	814.2	22.38	0.743	35,10,0,0	0.222	1.000	0.858	0.027	0.793
[OW=58]	L3	6.626	827.3	22.71	0.776	35,10,0,0	0.222	1.000	0.858	0.027	0.799
HU:0%	L4	6.453	967.8	24.82	0.823	35,10,0,0	0.222	1.000	0.858	0.027	0.865
	NN	6.453	967.8	24.82	0.823	35,10,0,0	0.222	1.000	0.858	0.027	0.865
ΔBI	L1	6.273	649.0	21.24	447.2	-	-	-	-	-	1.000
(2,2,0)	L2	6.273	659.0	21.41	461.7	7,15,15,8	0.667	0.995	0.380	0.290	1.008
[OW=76]	L3	5.869	486.9	18.68	322.0	18,9,4,14	0.289	0.004	0.009	0.390	0.866
HU:0%	L4	5.742	557.3	19.68	266.1	15,7,9,14	0.356	0.049	0.098	0.247	0.927
	NN	5.742	557.3	19.68	266.1	15,7,9,14	0.356	0.049	0.098	0.247	0.927
Net X	L1	5.543	529.3	18.03	68.81	-	-	-	-	-	1.000
(1,0,0)	L2	5.535	535.5	18.11	69.34	1,6,14,23	0.455	0.959	0.441	0.116	1.006
[OW=76]	L3	5.578	557.6	18.73	69.64	6,13,9,16	0.568	0.733	1.000	0.002	1.026
HU:9%	L4	5.520	654.7	20.47	72.52	6,18,9,11	0.614	0.957	0.393	0.162	1.112
	NN	5.526	615.1	19.86	70.51	6,17,9,12	0.591	0.932	0.390	0.129	1.078

[1] The equation shown above is the general specification for the linear models, where K1,K2,K3 can take values from 1 to 5, and where all estimations are done using levels data. The "Mod" column lists the models estimated (see footnote to Table 1). The selected lag structure is shown underneath the variable name in the first column of the table, with notation (K1,K2,K3), and the optimal window (OW) chosen for the nonadaptive linear VAR model is given in square brackets underneath the lag specification. The percentage of forecasts which are made using specifications with hidden units is given underneath the variable name in the first column of the table, with notation HU: %. All statistics are calculated using the *true* ex-post forecast period from 1982:3-1993:3, and the associated forecast errors. The 2x2 confusion matrices reported in the CM column of the table have diagonal cells (a11 and a22) corresponding to correct directional predictions, while off-diagonal cells (a12 and a21) correspond to incorrect predictions. The matrix is reported as a vector in the following order: a11, a12, a21, a22. The HM (Henriksson and Merton (1981)) and χ^2 confusion matrix tests of independence *p*-values are based on the null hypothesis that a given model is of no value in predicting the direction of change in the dependent variable. The Yates correction is applied to the χ^2 calculations. MSIC is the mean of the SIC across all forecast estimation periods. MSE, MAD, and MAPE are mean square error, mean absolute deviation, and mean absolute percentage error, respectively.

Table 5: Summary Model Selection Statistics: Forecast Horizon, h=4[1]

$$dep_{t+h-1} \;=\; \alpha \;+\; \sum_{i=1}^{K1} \beta_i \, dep_{t-i} \;+\; \sum_{i=1}^{K2} \delta_i \, ind1_{t-i} \;+\; \sum_{i=1}^{K3} \gamma_i \, ind2_{t-i} \;+\; u_{t+h-1}$$

Var	Mod	MSIC	MSE	MAD	MAPE	CM	CR	HM	χ^2	ϕ	TU
U	L1	0.517	1.206	0.831	11.39	-	-	-	-	-	1.000
(1,1,1)	L2	0.503	1.330	0.905	12.58	11,28,5,0	0.750	1.000	0.008	0.399	1.050
[OW=76]	L3	-0.255	1.079	0.873	12.47	5,9,11,19	0.455	0.650	0.783	0.042	0.946
HU:29%	L4	-1.125	0.619	0.674	9.593	10,8,6,20	0.318	0.030	0.060	0.284	0.716
	NN	-1.091	0.615	0.659	9.510	10,8,6,20	0.318	0.030	0.060	0.284	0.714
R	L1	0.565	1.809	1.052	10.58	-	-	-	-	-	1.000
(1,1,1)	L2	0.504	2.294	1.146	11.50	13,31,0,1	0.689	0.711	0.638	0.070	1.126
[OW=76]	L3	0.369	2.583	1.211	12.05	11,21,2,11	0.511	0.183	0.363	0.136	1.195
HU:53%	L4	-0.670	3.222	1.416	14.22	8,18,5,14	0.511	0.506	1.000	0.001	1.335
	NN	-0.105	3.128	1.438	14.54	5,17,8,15	0.556	0.889	0.573	0.084	1.315
IP	L1	3.336	22.43	3.621	3.784	-	-	-	-	-	1.000
(5,5,5)	L2	3.150	16.87	2.977	3.167	38,7,0,0	0.156	1.000	0.838	0.031	0.867
[OW=76]	L3	3.182	31.54	4.540	4.761	26,6,12,1	0.400	0.926	0.636	0.071	1.241
HU:24%	L4	2.456	57.37	5.861	6.040	23,2,15,5	0.378	0.378	0.250	0.171	1.599
	NN	2.504	38.41	5.124	5.307	23,2,15,5	0.378	0.378	0.250	0.171	1.309
NGNP	L1	10.76	90232	288.9	6.305	-	-	-	-	-	1.000
(4,0,4)	L2	9.105	16520	111.2	2.490	45,0,0,0	0.000	1.000	0.940	0.011	0.428
[OW=76]	L3	8.541	12839	89.73	2.052	45,0,0,0	0.000	1.000	0.938	0.011	0.377
HU:2%	L4	8.110	9941	79.19	1.766	45,0,0,0	0.000	1.000	0.938	0.011	0.332
	NN	8.104	10043	80.49	1.796	45,0,0,0	0.000	1.000	0.938	0.011	0.334
Π	L1	5.683	617.0	18.24	10.32	-	-	-	-	-	1.000
(4,0,4)	L2	5.560	486.3	17.19	10.14	26,18,0,0	0.409	1.000	0.877	0.023	0.888
[OW=76]	L3	5.447	488.3	17.73	10.73	23,11,3,7	0.318	0.040	0.078	0.266	0.890
HU:2%	L4	5.164	611.3	20.31	12.20	19,10,7,8	0.386	0.189	0.378	0.133	0.995
	NN	5.165	545.14	19.30	11.74	19,10,7,8	0.386	0.189	0.378	0.133	0.940
RGNP	L1	10.00	23248	132.5	2.979	-	-	-	-	-	1.000
(2,2,0)	L2	9.423	10822	85.47	1.931	39,6,0,0	0.133	1.000	0.827	0.037	0.682
[OW=58]	L3	8.682	5914	64.08	1.435	37,4,2,2	0.133	0.080	0.136	0.222	0.504
HU:0%	L4	8.464	6379	64.71	1.421	38,4,1,2	0.111	0.043	0.053	0.288	0.524
	NN	8.464	6379	64.71	1.421	38,4,1,2	0.111	0.043	0.053	0.288	0.524
PCE	L1	9.063	3382	89.86	3.042	-	-	-	-	-	1.000
(1,1,1)	L2	8.065	3464	49.10	1.671	41,4,0,0	0.089	1.000	0.793	0.039	0.580
[OW=58]	L3	7.586	2832	45.80	1.542	41,4,0,0	0.089	1.000	0.793	0.039	0.525
HU:4%	L4	7.370	2784	44.30	1.468	41,4,0,0	0.089	1.000	0.793	0.039	0.520
	NN	7.386	2819	44.51	1.474	41,4,0,0	0.089	1.000	0.793	0.039	0.523
ΔBI	L1	7.046	1528	29.34	976.2	-	-	-	-	-	1.000
(1,0,1)	L2	7.046	1553	29.68	971.0	8,10,16,11	0.578	0.900	0.502	0.100	1.008
[OW=76]	L3	6.491	1009	26.12	362.3	13,6,11,15	0.378	0.076	0.152	0.213	0.812
HU:0%	L4	6.162	753.2	22.60	508.7	18,6,6,15	0.267	0.002	0.005	0.420	0.702
	NN	6.162	753.2	22.60	508.7	18,6,6,15	0.267	0.002	0.005	0.420	0.702
Net X	L1	7.161	2539	43.79	141.9	-	-	-	-	-	1.000
(1,1,1)	L2	7.375	2701	46.26	153.5	0,8,19,18	0.600	1.000	0.023	0.339	1.032
[OW=58]	L3	7.353	3103	49.92	151.7	6,15,13,11	0.622	0.980	0.152	0.213	1.106
HU:31%	L4	6.667	3322	41.03	103.8	16,7,3,19	0.222	0.000	0.000	0.521	1.144
	NN	6.706	2485	38.70	100.2	16,8,3,18	0.244	0.000	0.001	0.484	0.989

[1] See notes to Table 4.

Table 1: Loss Differential Model Comparison Test p-Values Based on MSE [1]

Variable	Forecast Horizon	NN vs. L1	NN vs. L2	NN vs. L3	NN vs. L4	L1 vs. L2	L1 vs. L3	L1 vs. L4	L2 vs. L3	L2 vs. L4	L3 vs. L4
U	h=1	0.061(L1)	0.004(L2)	0.134	0.311	0.000(L2)	0.377	0.062(L1)	0.061(L2)	0.004(L2)	0.137
R		0.628	0.480	0.367	-	0.000(L2)	0.000(L3)	-	0.387	-	-
IP		0.083(NN)	0.405	0.758	0.311	0.000(L2)	0.001(L3)	0.076(L4)	0.311	0.424	0.795
NGNP		0.000(NN)	0.000(NN)	0.846	-	0.000(L2)	0.000(L3)	-	0.000(L4)	-	-
Π		0.099(NN)	0.007(L2)	0.000(L3)	-	0.000(L2)	0.000(L3)	-	0.007(L3)	-	-
RGNP		0.000(NN)	0.361	0.005(L3)	0.422	0.000(L2)	0.000(L3)	0.000(L4)	0.029(L3)	0.458	0.010(L3)
PCE		0.000(NN)	0.321	0.124	-	0.000(L2)	0.000(L3)	-	0.958	-	-
ΔBI		0.575	0.578	0.953	-	0.819	0.536	-	0.541	-	-
Net X		0.609	0.704	0.210	0.531	0.000(L1)	0.004(L1)	0.420	0.173	0.883	0.281
U	h=4	0.438	0.262	0.870	0.906	0.000(L2)	0.336	0.475	0.168	0.297	0.864
R		0.647	0.827	0.759	0.266	0.000(L2)	0.872	0.639	0.694	0.230	0.365
IP		0.228	0.856	0.513	0.524	0.000(L2)	0.019(L3)	0.181	0.398	0.680	0.803
NGNP		0.000(NN)	0.000(NN)	0.002(L3)	0.274	0.000(L2)	0.000(L3)	0.000(L4)	0.000(L3)	0.000(L4)	0.005(L3)
Π		0.130	0.614	0.578	0.274	0.000(L2)	0.001(L3)	0.185	0.145	0.797	0.278
RGNP		0.000(NN)	0.901	0.669	-	0.000(L2)	0.000(L3)	-	0.923	-	-
PCE		0.000(NN)	0.715	0.488	0.303	0.000(L2)	0.000(L3)	0.000(L4)	0.984	0.730	0.518
ΔBI		0.824	0.835	0.297	-	0.869	0.447	-	0.468	-	-
Net X		0.910	0.869	0.997	0.252	0.006(L1)	0.774	0.462	0.638	0.282	0.256

[1] The following acronyms are used in the table: NN - Flexible Specification Artificial Neural Network; L1 - Fixed Specification Linear Random Walk; L2 - Fixed Specification Linear Random Walk with Drift; L3 - "Best" Fixed Specification Linear Model; L4 - Flexible Specification Linear Model. Values reported above are p-values for loss differential tests. The tests give pairwise comparisons of the competing econometric models (see above). In cases were the p-values are less than or equal to 0.10, the "winning" model appears in brackets directly below the associated p-value. Variable mnemonics are shown above, and the loss differential tests (d_t) are constructed as follows (for models I and II, say): $d_t = \hat{u}^2_{I,t} - \hat{u}^2_{II,t}$, for the MSE test; $d_t = |\hat{u}_{I,t}| - |\hat{u}_{II,t}|$, for the MAD test; and $d_t = |(\hat{y}_{I,t}/y_t) - 1| - |(\hat{y}_{II,t}/y_t) - 1|$, for the MAPE test, where the y_t are the observed data, the $\hat{y}_t$ are the forecast data, and the $\hat{u}_t$ are the forecast errors. MSE is the forecast mean squared error of the 45 observation ex-ante forecast sample, MAD is the mean absolute deviation, and MAPE is the mean absolute percentage error for the forecast sequence.

Table 2: Overall Pairwise Model Performance Results: Wilcoxon Signed Rank Test p-Values[1]

Model Selection Criterion	Forecast Horizon	NN vs. L1	NN vs. L2	NN vs. L3	NN vs. L4	L1 vs. L2	L1 vs. L3	L1 vs. L4	L2 vs. L3	L2 vs. L4	L3 vs. L4
MSE	1	12 (0.125) (NN)	24 (0.455)	25 (0.410)	30 (0.213) (L4)	32 (0.150) (L2)	34 (0.102) (L3)	34 (0.102) (L4)	18 (0.326)	24 (0.455)	20 (0.410)
MAD	1	16 (0.248) (NN)	28 (0.285)	24 (0.455)	30 (0.213) (L4)	22 (0.500)	27 (0.326)	29 (0.248)	19 (0.367)	17 (0.285)	21 (0.455)
MAPE	1	10 (0.082) (NN)	29 (0.248) (L2)	23 (0.500)	30 (0.213) (L4)	35 (0.082) (L2)	34 (0.102) (L3)	36 (0.065) (L4)	13 (0.150) (L2)	16 (0.248) (L2)	23 (0.500)
MSE	4	11 (0.191) (NN)	20 (0.410)	23 (0.500)	18 (0.326)	32 (0.150) (L2)	34 (0.102) (L3)	34 (0.102) (L4)	19 (0.367)	25 (0.410)	21 (0.455)
MAD	4	17 (0.285)	26 (0.367)	27 (0.326)	23 (0.500)	20 (0.410)	29 (0.248) (L3)	26 (0.367)	16 (0.248) (L3)	19 (0.367)	18 (0.326)
MAPE	4	11 (0.102) (NN)	28 (0.285)	27 (0.326)	19 (0.367)	33 (0.125) (L2)	33 (0.125) (L3)	38 (0.037) (L4)	12 (0.125) (L2)	16 (0.248) (L2)	18 (0.326)

[1] The table summarizes the results of Wilcoxon signed rank tests (see above) on the MSEs, MADs and MAPEs listed in Tables 4-5 for each of the variables and for forecast horizons h=1,4.

Reported statistics are the sum of ranks of positive differences, $W = \sum_{i=1}^{S} T_i$, where $S = \sum_{i=1}^{NV} I_i$, $I_i = 1$ if $MSS_{I,i} - MSS_{II,i} > 0$, $I_i = 0$ otherwise, where the T_i are the positive ranks (see above), MSS is the value of the particular model selection statistic being examined, and NV is the number of variables in the experiment. In this way, model I can be thought of as the "control". Bracketed p-values correspond to the probability of observing the reported number of positive differences between I and II, under the null that model I is "better", assuming the differences are symmetrically distributed about 0. Thus, for example, a low p-value indicates that model I outperforms the "control" (model II) across all variables. The "best" models are listed below the bracketed p-values for those cases where p-values are less than or equal to 0.25. All statistics are calculated using the *true* ex-post observation period from 1982:3-1993:3. The MSE is the forecast mean squared error of the 45 out-of-sample, 1-step-ahead (h=1) or 4-step ahead (h=4) forecasts. Similarly, MAD is the mean absolute deviation, and MAPE is the mean absolute percentage error.