ELSEVIER

European Journal of Operational Research 132 (2001) 666–680

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

An investigation of model selection criteria for neural network time series forecasting

Min Qi [a], Guoqiang Peter Zhang [b,*]

[a] *College of Business Administration, Kent State University, Kent, OH 44242-0001, USA*

[b] *Department of Management, J. Mack Robinson College of Business, Georgia State University, University Plaza, Atlanta, GA 30303-3083, USA*

Received 23 March 1999; accepted 6 June 2000

Abstract

Artificial neural networks (ANNs) have received more and more attention in time series forecasting in recent years. One major disadvantage of neural networks is that there is no formal systematic model building approach. In this paper, we expose problems of the commonly used information-based in-sample model selection criteria in selecting neural networks for financial time series forecasting. Specifically, Akaike's information criterion (AIC) and Bayesian information criterion (BIC) as well as several extensions have been examined through three real time series of Standard and Poor's 500 index (S&P 500 index), exchange rate, and interest rate. In addition, the relationship between in-sample model fitting and out-of-sample forecasting performance with commonly used performance measures is also studied. Results indicate that the in-sample model selection criteria we investigated are not able to provide a reliable guide to out-of-sample performance and there is no apparent connection between in-sample model fit and out-of-sample forecasting performance.

Keywords: Time series forecasting; Neural networks; Model selection criteria; Akaike's information criterion (AIC); Bayesian information criterion (BIC)

1. Introduction

In recent years, artificial neural networks (ANNs) have attracted more and more attention from both academic researchers and industrial practitioners. ANNs' powerful pattern recognition and flexible nonlinear modeling capabilities are the main reasons for their popularity. In contrast to many model-based forecasting methods, ANNs are data driven without any restrictive assumptions about the functional relationships between the predictor variables and the predicated variable. This unique characteristic of ANNs is highly desirable in various financial forecasting situations, where data are generally abundant but the underlying data generating mechanism is often unknown or untestable.

* Corresponding author. Tel.: +1-404-651-4065; fax: +1-404-651-3498.

E-mail addresses: mqi@bsa3.kent.edu (M. Qi), gpzhang@gsu.edu (G.P. Zhang).

0377-2217/01/$ - see front matter
PII: S0377-2217(00)00171-5

Although neural networks have been successfully used for numerous forecasting applications, several issues in ANN model building still have not been solved (Zhang et al., 1998). One of the most critical issues is how to select appropriate network architecture for a forecasting task at hand. Model selection is a nontrivial issue in traditional linear forecasting applications and is a particularly tricky one for nonlinear models such as neural networks. Due to a typically large number of parameters to be estimated, ANNs often suffer overfitting problems. That is, they fit in-sample (or training) data very well but forecast poorly out of sample (or on the test set).

To ameliorate the overfitting effect, two broad types of model selection approaches are often adopted in the ANN literature. The first is the cross-validation-based approach that divides the available data into three parts: training, validation, and test sets. The training and validation sets are used for ANN model building while the last part is for genuine out-of-sample evaluation. The training set is used for parameter estimation for a number of alternative neural network specifications (e.g., networks of different number of inputs, different number of middle layer units in a three-layer feed-forward neural network). The learned network is evaluated with the validation set. The network model that performs the best on the validation set is selected as the final forecasting model. The validity and usefulness of the model is then checked using the test set. This out-of-sample model selection and testing procedure is generally quite effective in conquering the overfitting inclination of neural network models. However, it has several limitations. First, data splitting may increase the variability of the estimates (Faraway, 1992). Using a bootstrap method, LeBaron and Weigend (1998) show that the variation due to different data splitting is significantly larger than the variation due to different network conditions such as parameter initialization, choice of number of hidden units, etc. In addition, there exist no general guidelines on how to split the data into three parts. Finally, data splitting requires fairly large sample size. For applications with small data set, this procedure reduces the already small number of training patterns and attenuates the reliability of the model.

The second method is using the in-sample model selection criteria borrowed from the conventional time series literature. This approach relies solely on certain in-sample criterion as a convenient computational shortcut, hoping that the in-sample criterion can help choose the best forecasting model among alternatives. Information-based criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC), which penalize large models that often tend to overfit, are the most widely used in-sample model selection criteria. For example, Franses and Draisma (1997) use BIC to select the middle layer units for the neural network model that is designed to recognize changing seasonal patterns. Cottrell et al. (1995) use BIC as a guideline to model identification for both neural networks and ARIMA models. De Groot and Würtz (1992) and Faraway and Chatfield (1998) use both AIC and BIC in their neural network model selection and diagnostic checking. Swanson and White (1995, 1997) have also studied the BIC-based model selection criterion in forecasting short-term interest rates and several real-time macroeconomic series.

The purpose of this study is to empirically investigate the potential of the in-sample model selection criteria in neural network modeling for time series forecasting. This investigation is useful as it is not at all clear which criterion (if there is any) should be used and different researchers apply different criteria for no obvious reasons. A comprehensive study on the effectiveness of these in-sample model selection criteria for neural network time series forecasting is warranted. Several research questions are of particular interest. First, whether in-sample model selection criteria can select the model that warrants the best out-of-sample performance and if so, what is the best or generally better in-sample model selection criterion? The identification of such a criterion can ease ANN model building and enhance the general acceptance of neural networks as a valuable forecasting tool for practitioners. Second, what is the relationship between the in-sample criteria and the out-of-sample performance? Third, how good are the commonly used nonpenalty-based performance criteria such

as the mean squared error in model selection? We examine these issues empirically by using three real financial time series: Standard and Poor's 500 index (S&P 500 index), US interest rate, and exchange rate between the British pound and the US dollar.

The rest of the paper is organized as follows. Section 2 describes the neural network model and estimation method. Section 3 reviews the in-sample model selection criteria. Section 4 outlines the research design and the data. Section 5 reports the empirical findings. Conclusions and some further discussions are given in Section 6.

2. Neural networks for time series forecasting

ANNs are computational methods that are inspired by the brain and nerve system. They can be considered as a class of generalized nonlinear nonparametric statistical models (Smith, 1993; White, 1989). The comparative advantage of neural networks over more conventional econometric models lies in the fact that they can flexibly model complex, possibly nonlinear relationship without any prior assumptions about the underlying data-generating process. These data-driven self-adaptive methods have been widely used in pattern classification, pattern recognition and forecasting. This section briefly describes the multilayer feed-forward neural network, which is by far the most popular network paradigm in forecasting applications and is used in the present study. A more detailed review on the methodology and its financial applications is provided by Qi (1996). Zhang et al. (1998) present a comprehensive survey of neural networks in forecasting applications.

Our model is a three-layer feed-forward neural network with a single output unit, k middle layer units and n input units (see Fig. 1). The input layer can be represented by a vector $X = (x_1, x_2, \ldots, x_n)'$, the middle layer can be represented by a vector $M = (m_1, m_2, \ldots, m_k)'$, and y is the output. Any middle layer unit receives the weighted sum of all inputs and a bias term (denoted by x_0, x_0 always equals one), and produces an output signal

$$m_j = F\Big(\sum \beta_{ij} x_i\Big) = F(X'\beta_j), \quad j = 1, 2, \ldots, k, \;\; i = 0, 1, 2, \ldots, n, \tag{1}$$

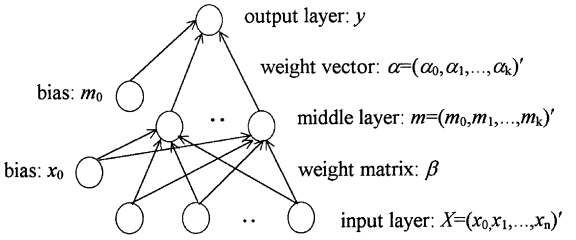

Fig. 1. A typical three-layer feed-forward neural network.

where F is a logistic function, x_i the ith input signal, and β_{ij} is the weight of the connection from the ith input unit to the jth middle layer unit. In the same way, the output unit receives the weighted sum of the output signals of the middle layer units, and produces a signal

$$y = G\Big(\sum \alpha_j m_j\Big), \quad j = 0, 1, 2, \ldots, k, \tag{2}$$

where G is an identity function, α_j the weight of the connection from the jth middle layer unit to the output unit, and $j = 0$ indexes a bias unit m_0 which always equals one. Substituting (1) into (2), we get

$$y = G\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j F\Big(\sum \beta_{ij} x_i\Big)\right) = f(X, \theta), \tag{3}$$

where X is the vector of inputs, and $\theta = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k, \; \beta_{11}, \beta_{12}, \ldots, \beta_{1n}, \ldots, \beta_{k1}, \beta_{k2}, \ldots, \beta_{kn})'$ is the vector of network weights.

For an extrapolative or time series forecasting problem, the inputs are the past observations of the data series and the output is the future value. The neural network performs the following functional mapping:

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \theta), \tag{4}$$

where y_t is the observation at time t. Thus, the neural network is equivalent to a nonlinear autoregressive (AR) model for time series forecasting problems.

For a time series forecasting problem, the training data usually consist of a fixed number of observations of the time series. Suppose there are N observations, $y_1, y_2, \ldots, y_N$, in the training set and we are interested in one-step-ahead forecasting. Using a neural network with p input nodes, we

have $N-p$ training patterns. The first training pattern will be composed of $y_1, y_2, \ldots, y_p$ as inputs and y_{p+1} as the target output. The second training pattern will contain $y_2, y_3, \ldots, y_{p+1}$ as inputs and y_{p+2} as the desired output. The last training pattern will have $y_{N-p}, y_{N-p+1}, \ldots, y_{N-1}$ as inputs and y_N as the target output. Then the parameters of the neural network can be determined by minimizing the following objective function of SSE in the training process:

$$\mathrm{SSE} = \sum_{i=p+1}^{N} (y_i - \hat{y}_i)^2, \tag{5}$$

where $\hat{y}_i$ is the output of the network.

3. Model selection criteria

In this section, we review the most commonly used model selection criteria in time series analysis and forecasting. For a more detailed account of the materials in this area, readers are referred to De Gooijer et al. (1985).

AIC (Akaike, 1974) is the most popular one for linear and nonlinear model identification. One common form of AIC is given below:

$$\mathrm{AIC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{2m}{T}, \tag{6}$$

where $\hat{\sigma}^2_{\mathrm{MLE}}$ denotes the maximum likelihood estimate of the variance of the residual term,

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{\mathrm{SSE}}{T} = \frac{\sum (y_i - \hat{y}_i)^2}{T},$$

m the number of parameters in the model and T is the number of observations. The first part in (6) measures the goodness-of-fit of the model to the data while the second part sets a penalty for model over-parameterization, i.e. overfitting. The optimal model is selected when AIC is minimized. It is clear that as the model becomes more complex, the first term of (6) will be smaller but the second term larger. Hence AIC is a reasonable criterion, which balances model fitting and model parsimony.

It should be noted that (6) is derived based on the assumption that the data are generated from a Gaussian (normal) process. Since in reality few time series are generated exactly from normal processes, AIC may not be the optimal criterion for a particular situation. In fact, AIC is not a consistent criterion in model selection. In addition, the AIC often leads to a model with unnecessarily large number of parameters. This is particular true when nonlinear models are considered by this criterion (De Gooijer and Kumar, 1992).

Another popular criterion is the BIC. Several forms of BICs have been proposed in the literature. One typical BIC is defined as follows:

$$\mathrm{BIC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{m \log(T)}{T}. \tag{7}$$

Eq. (7) is very similar to (6) in that BIC is also composed of two parts with the same first item as in (6). The difference is in the penalty term. It is clear that if $T > 7$, BIC imposes greater penalty for model complexity than AIC. Hence the use of BIC for model selection would result in a model whose number of parameters is no greater than that chosen by AIC. Eq. (7) is developed independently by Schwarz (1978) and Rissanen (1978). Rissanen (1980) shows that the BIC gives a consistent estimate of the order of an AR model. Therefore, in real applications, BIC is often preferred to AIC because it is a more reliable criterion for model selection.

There are some debates in the literature about the appropriateness of the penalty term used in both AIC and BIC. The linear function of the number of parameters in the penalty term is particularly controversial (De Gooijer et al., 1985). One of the extensions to these criteria proposed by De Gooijer and Kumar (1992) and advocated by Granger (1993) is

$$\mathrm{BIC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{m^d \log(T)}{T}, \tag{8}$$

where d is a constant usually set greater than 1 for nonlinear model identification. The introduction of the exponential d allows greater flexibility in the magnitudes of the appropriate penalty term. However, there is no investigation in the literature on the selection of an appropriate d in nonlinear modeling.

Another model selection criterion that is often used to chose the number of regressors is Theil's adjusted R^2 (or $\bar{R}^2$). Though the usual coefficient of determination (R^2) measures the goodness-of-fit of a model, it almost invariably increases and never decreases with the number of regressors. Therefore, if R^2 were used as a model selection criteria, it would always favor larger number of lags. The adjusted R^2 corrects the problem with an adjustment to the degrees of freedom. Sum of squared errors (SSE) or residual variance ($\hat{\sigma}^2$) have also been used to measure the goodness-of-fit. It should be noted that minimizing the estimated residual variance $\hat{\sigma}^2$ is equivalent to maximizing $\bar{R}^2$, thus $\hat{\sigma}^2$ is redundant in cases where $\bar{R}^2$ is used. Although $\bar{R}^2$ has been widely accepted as a model selection criterion, Cameron (1993) argues that $\bar{R}^2$ is not an effective tool for the prevention of data mining because it will rise on the addition of any variable whose t ratio is greater than one when entered into the model. Therefore, some empirical investigation on the effectiveness of using $\bar{R}^2$ as a model selection criterion, especially in the neural network context is necessary.

Other criteria, such as Mallows' (1973) C_p criterion, Hocking's (1976) S_p criterion, Amemiya's (1980) prediction criterion, and Phillips' (1992) posterior information criterion, have also been suggested and used for statistical model selection. However, these criteria have not been widely adopted in the forecasting literature and practice and hence are excluded in this study.

4. Research design

4.1. Model selection criteria

In this paper, we empirically study the effectiveness of several popular criteria in ANN model selection. In particular, we are concerned with the two most commonly used criteria, AIC and BIC defined in (6) and (7). In addition, several extensions to these criteria are also investigated. Specifically, the following general formulas for AIC and BIC are used in our study:

$$\mathrm{AIC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{2m^d}{T}, \tag{9}$$

and

$$\mathrm{BIC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{m^d \log(T)}{T}.$$

Granger (1993) recommends that $d > 1$ for nonlinear models and that experimental method may be used to determine d. We select the following experimental values for d: $\log_m(\log(m))$, $1/2$, 1, and 2. In previous versions of this paper, we have also worked with $d = 3$, and found that it always selects the most parsimonious model. Therefore, we omit the results for $d = 3$ to save space. Notice that our experimental values for d include two cases, where $d < 1$.

Following the suggestion of an anonymous referee, we also include a modified version of AIC (AICC) which is discussed extensively in Brockwell and Davis (1991) and Burnham and Anderson (1998). AICC is defined as

$$\mathrm{AICC} = \log(\hat{\sigma}^2_{\mathrm{MLE}}) + \frac{2m}{T-m-1}. \tag{10}$$

The purpose of using the AICC is to use a more appropriate penalty term to ameliorate the overfitting tendency of the AIC. It is clear from (10) that for large models, the AICC has much higher penalty than the AIC.

Table 1 defines all of the model selection criteria and performance measures examined in this study. These criteria are applied to neural network models with different lagged input variables and different middle layer units, and to linear AR models. For the criteria listed in Table 1, AIC3 is the original AIC defined by Eq. (6), AIC1, AIC2, and AIC4 are three extensions to AIC with increasing penalty to model complexity. Similarly, BIC3 represents the original BIC as defined in Eq. (7) and BIC1, BIC2, and BIC4 are three different variations of BIC. In general, we would expect that the higher the penalty associated with a model selection criterion, the simpler the model will be chosen by that criterion.

$\bar{R}^2$ and other nonpenalty related performance measures such as RMSE (root mean squared error), MAE (mean absolute error), MAPE (mean absolute percentage error), ME (mean error), DA (direction accuracy), Sign (percentage of correct

Table 1
List of the model selection criteria and performance measures

Model selection criterion	Definition	Model selection criterion	Definition
SSE	$\sum_{i=1}^{T}(y_i-\hat{y}_i)^2$	$\bar{R}^2$	$1-\frac{\mathrm{SSE}/(T-m)}{\sum(y_i-\bar{y})^2/(T-1)}$
AIC1	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{2\log(m)}{T}$	BIC1	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{\log(m)\log(T)}{T}$
AIC2	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{2\sqrt{m}}{T}$	BIC2	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{\sqrt{m}\log(T)}{T}$
AIC3	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{2m}{T}$	BIC3	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{m\log(T)}{T}$
AIC4	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{2m^2}{T}$	BIC4	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{m^2\log(T)}{T}$
AICC	$\log\left(\frac{\mathrm{SSE}}{T}\right)+\frac{2m}{(T-m-1)}$	MAE	$\frac{1}{T}\sum\lvert(y_i-\hat{y}_i)\rvert$
RMSE	$\sqrt{\frac{1}{T}\mathrm{SSE}}$	ME	$\frac{1}{T}\sum(y_i-\hat{y}_i)$
MAPE	$\frac{1}{T}\sum\left\lvert\frac{(y_i-\hat{y}_i)}{y_i}\right\rvert$	Sign	$\frac{1}{T}\sum z_i$, where $z_i=\begin{cases}1 & \text{if } y_{i+1}\cdot\hat{y}_{i+1}>0,\\ 0 & \text{otherwise.}\end{cases}$
DA	$\frac{1}{T}\sum a_i$, where $a_i=\begin{cases}1 & \text{if } (y_{i+1}-y_i)(\hat{y}_{i+1}-y_i)>0,\\ 0 & \text{otherwise.}\end{cases}$		

signs), are also used to examine the patterns between in-sample and out-of-sample performance based on various performance measures. The definitions of these performance measures are also given in Table 1.

4.2. Data

Three financial time series, the monthly S&P 500 index, the monthly one-month treasure-bill (T-bill) rate, and the weekly exchange rate between the British pound and the US dollar are used in this study. S&P 500 index at close on the last trading day of each month are taken from S&P Statistical Service. The sample period is 1954(1) to 1992(12). The one-month T-bill rates are measured on the last trading day of each month and computed as the average of the bid and ask yields and the data are taken from the Fama–Bliss risk-free rates file on the Center for Research in Security Prices, CRSP, tapes. The sample period is also from 1954(1) to 1992(12). The exchange rates are obtained from DataStream International and contain daily rates from the beginning of 1976 through the end of 1993. The rates are quotations at 3:00 p.m. eastern time from Banker Trust. Weekly observations are compiled by taking the rates on Wednesday as the representative rates of that week to avoid potential biases caused by the weekend effect. If a particular Wednesday happens to be a nontrading day, then either Tuesday or Thursday rate is used. Following Meese and Rogoff (1983), we apply natural logarithmic transformation to the S&P 500 index and the exchange rates to stabilize the series. The augmented Dickey–Fuller test shows that all the three series contain a unit root, thus the first order difference is applied.

For both S&P 500 index and one-month T-bill rate, the in-sample period goes from 1954(1) to 1984(12) and the out-of-sample period goes from 1985(1) to 1992(12). The model estimated from the in-sample data is used to forecast the series in the next one, four, and eight years. For the exchange rate between the US dollar and the British pound, the in-sample data contain weekly observations from the first week of 1976 to the last week of 1989 and the out-of-sample data contain weekly series from the first week of 1990 to the last week of 1993. The model estimated from the in-sample data is used to forecast the exchange rate in the next one, two, and four years.

All of the series are plotted in Fig. 2. While the logarithm of S&P 500, log (S&P500), is obviously nonstationary and has an upward trend, the first order difference, $\Delta \log$ (S&P500), appears to be stationary. Agreeing with the augmented Dickey–Fuller test, the original one-month T-bill rates and

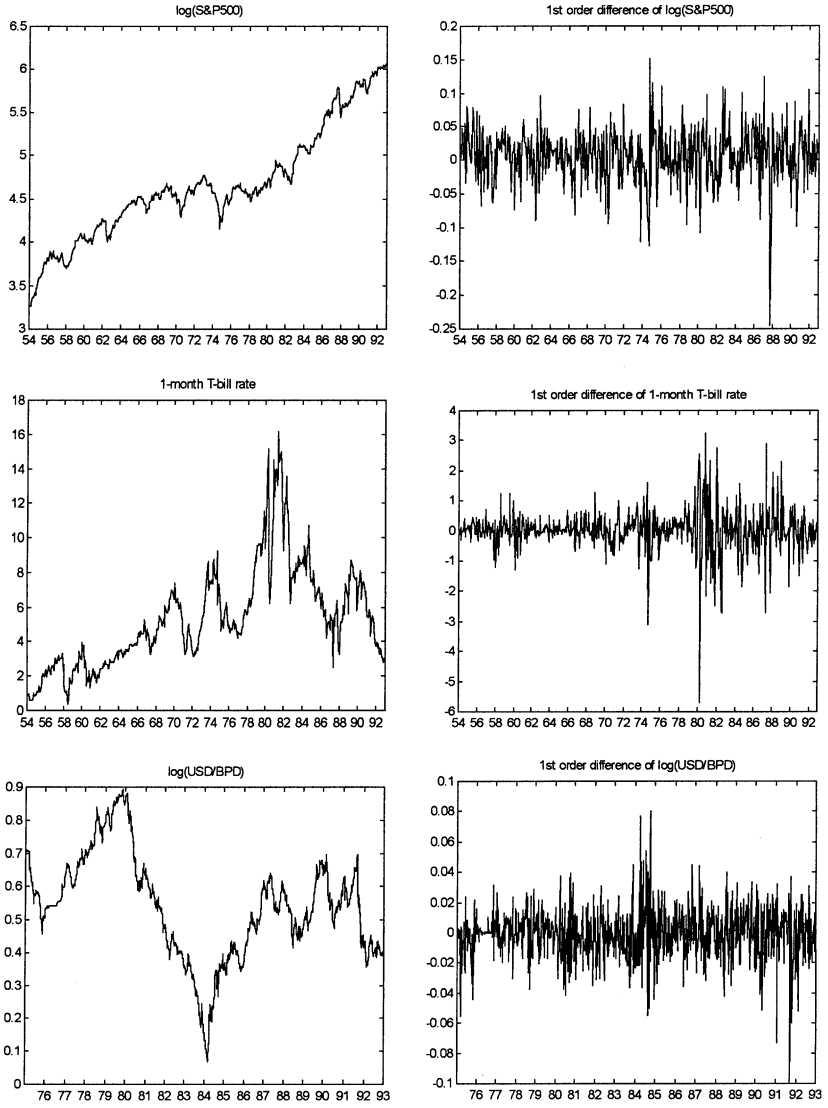

Fig. 2. Time series plot of the data.

logarithm of exchange rates are nonstationary, but the first order differences seem stationary.

4.3. Design of experiments

More model uncertainties are associated with neural networks than with traditional linear AR models. In addition to the number of lagged values (p), the construction of a multilayer feed-forward neural network typically involves specification of several other parameters or factors: the number of layers, the number of middle layer units, and the transfer functions associated with the middle and output layer units. It has been shown that a three-layer feed-forward network is able to approximate any continuous function. Furthermore, a neural network with logistic middle layer unit transfer function and identity output unit transfer function is the most popular choice for many successful applications of neural forecasters. Hence, the two key uncertainties associated with neural network model selection in the context of time series forecasting are the number of lagged values in the input vector (p) and the number of middle layer units (k). The total number of parameters in a neural network with p inputs and k middle layer units is $m = k(p+2)+1$.

For each data series, both the number of input units and the number of middle layer units will vary from 1 to 5. Thus there are 25 different neural network models to be applied to each problem. We also compare the neural network results with those of linear AR models and random walks. Five AR models with lags ranging from 1 to 5 are fitted to each series. Each linear AR model also contains a constant term, therefore, the total numbers of free parameters in these linear models range from 2 to 6.

5. Empirical findings

The notations for various models are as follows. AR(p) or (p) indicates linear AR model with p lagged dependent variables. NN(p,k) or (p,k) indicates neural network model with p lagged dependent variables as input and k middle layer units.

5.1. S&P 500 index

The in- and out-of-sample model selection results for the S&P 500 index are given in Table 2. From the in-sample results, different criteria in general yield different "best" model though there is some agreement. For example, by $\bar{R}^2$, the best in-sample model is (1,4), i.e., a neural network with one input (or lagged dependent variable) and four

Table 2
Best models selected for the S&P 500 index

Criterion	Best model (in sample)	Criterion	Best model (in sample)
SSE	(1,4)	$\bar{R}^2$	(1,4)
AIC1	(1,4)	BIC1	(1,4)
AIC2	(1,4)	BIC2	(1,4)
AIC3	(1,4) (1)	BIC3	(1,1) (1)
AIC4	(1,1) (1)	BIC4	(1,1) (1)
AICC	(1,1) (1)		

		Best model (out of sample)		
		One-year	Four-year	Eight-year
RMSE	(1,4)	(5,5)	(5,1) (2)	(5,1) (2)
MAE	(5,3) (5)	(5,5)	(5,3)	(5,3)
MAPE	(4,4)	(2,3)	(4,2)	(4,2)
ME	(4,5)	(5,5)	(5,3)	(1,4)
DA	(2,3)	(2,3)	(1,2)(2,4)(4,3)	(4,3)
Sign	(5,4)	(4,2)	(1,1)(2,1)(2,2)(2,4)(3,1)(3,2)(3,3)(4,1)(5,1)(5,3)	(4,1)(1)(2)

middle layer units. Although the same model is selected by AIC1–AIC3 and BIC1–BIC2, a different model (1,1) is chosen by AIC4 and BIC3–BIC4. Comparing the best models chosen by different AICs and BICs, it is always the case that for any in-sample model selection criterion used, the higher the penalty term, the simpler the model it picks, which is expected. The in-sample result is fairly consistent in that all of our penalty-related selection criteria pick either (1,4), or (1,1). It is also noticed from the table that the AICC chooses the same best model as the BIC_5.

However, judging from the performance measured by RMSE, MAE, MAPE, DA and Sign, none of the best model chosen by the in-sample penalty related criteria, i.e., (1,4) and (1,1) is the best for out-of-samples across three forecasting horizons. Furthermore, there is no convincing evidence of the correspondence between the in- and out-of-sample performance based on any of the six performance measures for all three forecasting horizons. For example, though model (1,4) has the smallest RMSE in sample, it does not have the smallest RMSE out of sample. Instead, neural network model (5,5) has the smallest RMSE for 1-year forecasting horizon, (2,2) and (5,1) for 4-year horizon, and (5,1) for eight-year horizon.

From the results for both in and out of samples, several observations can be made. First, the models selected based on in-sample data by neither the penalty-related criteria nor the no-penalty-related performance measures are consistent with the best performances in out of samples. Second, although the investigated AICs and BICs cover a relatively wide range of degrees of penalty for model complexity, it is not clear which one gives the most appropriate penalty since none of the models chosen by these criteria performs the best out-of-sample. Third, AICC and BIC have the similar effect in model selection. In fact, the same best in-sample models are selected by both AICC and BIC criteria. This result is in line with the finding of Faraway and Chatfield (1998). Finally, judging from different performance measures, there seems to be no clear pattern in the out-of-sample performance of alternative models. In general, the six out-of-sample performance measures do not seem to agree with one another.

Tables 3 and 4 compare the in-sample fit and out-of-sample prediction of various neural network, linear AR and random walk models. The "curse of complexity" is apparent in that as the number of lagged variables (p) and the number of middle layer units (k) increase, the number of parameters in the neural network model increases dramatically as compared to the linear AR model. As expected, for each of the 25 neural network and five AR models tabulated, as the penalty on model complexity (d) in both AIC and BIC increases, the values of AIC and BIC increase. For the same value of d, BIC is always larger than AIC, indicating that the BIC penalizes model complexity more than AIC does. The best model is NN(1,4) based on AIC1–AIC3 and BIC1–BIC2, and NN(1,1) based on AIC4 and BIC3–BIC4. Hence AIC and BIC give different "best" models. As expected, although the in-sample fit of various neural networks and linear AR models is generally quite close, most neural network models fit the data better than linear AR models.

From the out-of-sample prediction in Table 4, it is obvious that a good in-sample fit, i.e., a low AIC or BIC value, or a low RMSE value, is a poor guide to model selection for out-of-sample prediction. Neither NN(1,4) nor NN(1,1) has the best performance out-of-sample. An important observation in Table 4 is that the best models based on out-of-sample performance measures are more complex than those chosen by the penalty-related in-sample model selection criteria. This indicates that the best in-sample models may be too parsimonious. In other words, the model selection criteria may over-penalize the model complexity and make the model underfit the data. In general, we find the predictive performance of the best neural network model is better than or as comparable as the linear AR models and in many cases both ANN and AR models have more accurate predictions than the random walk.

5.2. Exchange rate between US dollar and British pound

Table 5 reports the in- and out-of-sample model selection results for the exchange rate between US

Table 3
Comparison of the in-sample fit for the S&P 500 index data

p	k	m	RMSE×100	AIC1	AIC2	AIC3	AIC4	BIC1	BIC2	BIC3	BIC4	AICC
1	1	4	4.024	−6.418	−6.415	−6.404	−6.339	−6.404	−6.394	−6.362	−6.170	−6.404
1	2	7	4.022	−6.416	−6.413	−6.389	−6.162	−6.396	−6.385	−6.315	−5.644	−6.388
1	3	10	4.020	−6.415	−6.411	−6.374	−5.887	−6.391	−6.377	−6.268	−4.829	−6.372
1	4	13	3.925	−6.462	−6.456	−6.405	−5.562	−6.435	−6.418	−6.268	−3.775	−6.403
1	5	16	4.004	−6.421	−6.414	−6.349	−5.052	−6.391	−6.372	−6.180	−2.344	−6.345
2	1	5	4.031	−6.414	−6.410	−6.395	−6.287	−6.397	−6.387	−6.342	−6.022	−6.395
2	2	9	4.028	−6.412	−6.408	−6.375	−5.985	−6.389	−6.376	−6.280	−5.126	−6.374
2	3	13	4.095	−6.377	−6.371	−6.320	−5.475	−6.350	−6.333	−6.183	−3.684	−6.318
2	4	17	4.031	−6.407	−6.400	−6.330	−4.856	−6.377	−6.356	−6.150	−1.793	−6.326
2	5	21	4.007	−6.418	−6.409	−6.320	−4.044	−6.386	−6.361	−6.098	0.630	−6.313
3	1	6	4.030	−6.413	−6.410	−6.390	−6.227	−6.394	−6.384	−6.327	−5.845	−6.390
3	2	11	4.015	−6.417	−6.412	−6.371	−5.773	−6.392	−6.377	−6.254	−4.488	−6.369
3	3	16	4.020	−6.413	−6.406	−6.341	−5.036	−6.383	−6.363	−6.171	−2.318	−6.336
3	4	21	4.021	−6.411	−6.402	−6.313	−4.030	−6.378	−6.354	−6.090	0.653	−6.306
3	5	26	3.978	−6.431	−6.421	−6.308	−2.775	−6.397	−6.367	−6.031	4.404	−6.296
4	1	7	4.033	−6.411	−6.407	−6.383	−6.154	−6.390	−6.379	−6.309	−5.633	−6.382
4	2	13	4.025	−6.411	−6.406	−6.354	−5.504	−6.384	−6.367	−6.216	−3.706	−6.352
4	3	19	3.992	−6.426	−6.418	−6.338	−4.474	−6.394	−6.372	−6.136	−0.633	−6.332
4	4	25	4.007	−6.417	−6.407	−6.298	−3.028	−6.382	−6.354	−6.032	3.623	−6.287
4	5	31	3.991	−6.423	−6.412	−6.273	−1.205	−6.387	−6.353	−5.943	9.021	−6.257
5	1	8	4.037	−6.408	−6.404	−6.376	−6.070	−6.386	−6.374	−6.290	−5.387	−6.375
5	2	15	4.036	−6.405	−6.399	−6.338	−5.191	−6.376	−6.358	−6.178	−2.791	−6.334
5	3	22	4.000	−6.421	−6.412	−6.318	−3.793	−6.388	−6.362	−6.083	1.368	−6.310
5	4	29	3.996	−6.421	−6.410	−6.281	−1.844	−6.386	−6.353	−5.972	7.123	−6.267
5	5	36	4.007	−6.415	−6.401	−6.237	0.648	−6.376	−6.337	−5.854	14.467	−6.215
AR												
1	0	2	4.025	−6.421	−6.418	−6.414	−6.404	−6.414	−6.403	−6.393	−6.361	−6.414
2	0	3	4.028	−6.418	−6.415	−6.408	−6.375	−6.406	−6.396	−6.376	−6.280	−6.407
3	0	4	4.024	−6.418	−6.415	−6.404	−6.339	−6.403	−6.393	−6.361	−6.169	−6.404
4	0	5	4.007	−6.425	−6.422	−6.407	−6.298	−6.408	−6.398	−6.354	−6.032	−6.406
5	0	6	3.996	−6.430	−6.427	−6.407	−6.243	−6.411	−6.400	−6.343	−5.859	−6.406

dollar and British pound (USD/BPD). The in-sample model selection results for exchange rate data are very similar to those for S&P 500 index. While $\bar{R}^2$, AIC1–AIC3, AICC, and BIC1–BIC2 chose (1,4) as the best model for both exchange rate and S&P 500 index, AIC4, and BIC4 chose (1,1) as the best model for both series. It is clear that the models selected by all penalty-related criteria are quite simple with only one lagged dependent variable in almost all cases. However, based on the out-of-sample performance most of these parsimonious models do not show any superiority over a little more complex models, suggesting possible underfitting of the neural networks selected by these in-sample criteria.

From Table 5, once again we see that the models selected by either the penalty-related criteria or the no-penalty-related performance measures do not necessarily have the best performance out of sample. However, the model (2,1) chosen by BIC3, the original version of BIC, does have the lowest out-of-sample MAPE for all three forecasting horizons and the lowest RMSE at a four-year horizon. Overall no criterion gives consistently the most appropriate model which has the best predictive ability. Model (1,5) is perhaps the most viable model from the out-of-sample perspective, since it has the best RMSE, MAE, ME, DA and Sign for one-year horizon, the best MAE, ME and Sign for two-year horizon, and the best MAE, DA and Sign for four-year horizon. However, none of the in-sample model selection criteria and performance measures indicates it is the best. The out-of-sample

Table 4
Comparison of the out-of-sample predictions for the S&P 500 index data

p	k	m	One-year				Four-year				Eight-year			
			RMSE ×100	MAE ×100	MAPE	DA ×100	RMSE ×100	MAE ×100	MAPE	DA ×100	RMSE ×100	MAE ×100	MAPE	DA ×100
1	1	4	3.510	2.797	1.206	58.333	5.611	3.927	1.044	64.583	4.890	3.528	1.228	63.542
1	2	7	3.528	2.862	1.304	58.333	5.629	3.966	1.073	62.500	4.897	3.544	1.270	62.500
1	3	10	3.456	2.759	1.228	58.333	5.627	3.935	1.073	62.500	4.897	3.521	1.267	62.500
1	4	13	3.460	2.822	1.377	58.333	8.853	4.854	1.217	60.417	6.850	3.932	1.402	61.458
1	5	16	3.536	2.816	1.212	58.333	5.672	3.959	1.034	58.333	4.913	3.516	1.160	60.417
2	1	5	3.579	2.872	1.225	58.333	5.624	3.947	1.017	64.583	4.890	3.528	1.153	63.542
2	2	9	3.532	2.786	1.089	58.333	5.593	3.887	0.971	64.583	4.884	3.510	1.097	63.542
2	3	13	3.892	2.933	0.847	50.000	5.813	4.117	1.020	43.750	5.080	3.703	1.201	39.583
2	4	17	3.457	2.796	1.348	58.333	5.663	3.949	1.119	64.583	4.914	3.521	1.360	63.542
2	5	21	3.460	2.762	1.263	58.333	5.811	3.969	1.141	58.333	5.036	3.587	1.421	55.208
3	1	6	3.541	2.841	1.224	58.333	5.612	3.926	1.020	64.583	4.882	3.515	1.169	63.542
3	2	11	3.693	3.054	1.538	58.333	5.678	4.035	1.214	64.583	4.927	3.582	1.511	63.542
3	3	16	3.609	2.932	1.319	58.333	5.612	3.865	1.037	64.583	4.903	3.505	1.208	63.542
3	4	21	3.524	2.815	1.179	58.333	5.624	3.945	1.057	62.500	4.907	3.540	1.240	61.458
3	5	26	3.939	3.236	1.596	25.000	5.637	3.968	1.141	47.917	4.937	3.591	1.471	45.833
4	1	7	3.535	2.851	1.265	58.333	5.614	3.924	1.034	64.583	4.880	3.510	1.201	63.542
4	2	13	3.590	2.739	0.901	83.333	5.693	3.946	0.927	62.500	4.970	3.571	1.054	55.208
4	3	19	3.697	2.930	1.090	66.667	5.745	3.988	1.021	62.500	5.010	3.585	1.231	55.208
4	4	25	3.849	3.080	1.204	25.000	5.862	4.054	1.042	52.083	5.101	3.634	1.060	58.333
4	5	31	3.672	2.939	1.191	58.333	5.706	4.055	1.129	62.500	4.979	3.595	1.198	60.417
5	1	8	3.521	2.794	1.137	58.333	5.587	3.898	0.989	64.583	4.875	3.514	1.145	63.542
5	2	15	3.568	2.878	1.289	58.333	5.683	3.996	1.055	58.333	4.928	3.552	1.211	60.417
5	3	22	3.239	2.673	1.354	58.333	5.620	3.814	1.238	64.583	4.937	3.487	1.986	61.458
5	4	29	3.537	2.882	1.425	58.333	5.615	3.988	1.258	62.500	4.923	3.605	1.546	57.292
5	5	36	3.149	2.629	1.395	58.333	5.725	3.842	1.268	60.417	5.020	3.522	1.958	59.375
AR														
1	0	2	3.519	2.810	1.213	58.333	5.599	3.916	1.036	66.667	4.880	3.519	1.212	64.583
2	0	3	3.503	2.785	1.155	58.333	5.584	3.896	1.013	66.667	4.875	3.512	1.172	64.583
3	0	4	3.553	2.858	1.233	58.333	5.613	3.934	1.051	62.500	4.897	3.533	1.230	62.500
4	0	5	3.571	2.857	1.191	58.333	5.755	4.000	1.050	58.333	4.994	3.568	1.244	55.208
5	0	6	3.374	2.674	1.094	58.333	5.659	3.873	1.076	58.333	4.961	3.537	1.509	54.167
RW		0	3.986	3.233	2.015	66.667	7.441	5.555	2.260	54.167	6.762	5.266	3.175	50.000

Table 5
Best models selected for the exchange rate (USD/BPD)

Criterion	Best model (in sample)	Criterion	Best model (in sample)
SSE	(1,4)	$\bar{R}^2$	(1,4)
AIC1	(1,4)	BIC1	(1,4)
AIC2	(1,4)	BIC2	(1,4)
AIC3	(1,4) (1)	BIC3	(2,1) (1)
AIC4	(1,1) (1)	BIC4	(1,1) (1)
AICC	(1,4) (1)		

		Best model (out of sample)		
		One-year	Two-year	Four-year
RMSE	(1,4)	(1,5)	(2,2)	(2,1)
MAE	(4,5)	(1,5)	(1,5) (2,2)	(1,5)
MAPE	(5,2)	(2,1)	(2,1)	(2,1)
ME	(4,2)	(1,5)	(1,5)	(1,3) (5,3)
DA	(4,1)	(1,5)	(2,2)	(1,5) (2,2)
Sign	(4,5)	(1,5)	(1,5)	(1,5)

performance seems to be consistent across different forecasting horizons.

5.3. One-month T-bill rate

The in- and out-of-sample model selection results for the one-month T-bill rate is given in Table 6. The in-sample results show that $\bar{R}^2$, AIC1–AIC3, and BIC1–BIC2 chose the same model (5,5), while other criteria with higher penalty for complexity chose simpler models. Although none of the models selected by the in-sample criteria is predominantly the best out of sample, the simplest model (1,1) chosen by BIC4 does have the best MAPE, DA at the four-year horizon, and the best MAE, MAPE, and DA at the eight-year horizon. Models (2,2) and (3,5) have some forecasting power out of sample. Nevertheless, neither has been selected as the best model by all of the in-sample criteria and performance measures.

6. Conclusions

ANNs have received more and more attention in time series forecasting in recent years. One major

Table 6
Best models selected for the one-month T-bill rate

Criterion	Best model (in sample)	Criterion	Best model (in sample)
SSE	(5,5)	$\bar{R}^2$	(5,5)
AIC1	(5,5)	BIC1	(5,5)
AIC2	(5,5)	BIC2	(5,5)
AIC3	(5,5)	BIC3	(3,4)
AIC4	(4,1) (1)	BIC4	(1,1) (1)
AICC	(3,4)		

		Best model (out of sample)		
		One-year	Four-year	Eight-year
RMSE	(5,5)	(2,2)	(3,5)	(3,5)
MAE	(3,4)	(2,2)	(1,5)	(1,1)
MAPE	(3,1)	(2,2)	(1,1)	(1,1)
ME	(1,3)	(2,5)	(2,2)	(4,4)
DA	(5,3)	(3,5) (4,4) (4,5)	(1,1)	(1,1)
Sign	(5,5)	(3,5) (4,5)	(3,5)	(3,5)

disadvantage of neural networks is that there is no formal systematic model building approach. Due to a typically large number of parameters to be estimated, ANNs have the tendency to overfit. Two broad types of model selection approaches are often adopted in the ANN literature to ameliorate overfitting: cross-validation and in-sample model selection based on certain criterion. The problems and limitations of the first approach have been extensively studied in the literature (see Faraway (1992), LeBaron and Weigend (1998), for example). In the present research we conduct a comprehensive investigation on the effectiveness of a variety of commonly used in-sample model selection criteria and their variations.

Through forecasting of three financial time series we show that the commonly used in-sample model selection criteria are not able to identify the best neural network model for out-of-sample prediction. Results clearly indicate the inconsistency between the best in-sample model selected by the popular model selection criteria and the best model out of sample. Furthermore, there seems to exist no agreement between the best in-sample model and the best out-of-sample model based on the same performance measure such as RMSE, MAE, MAPE, DA, and Sign. Therefore, neither model selection criteria nor the performance measures based on in-sample data alone can serve as a reliable guide for choosing the model that has the best out-of-sample performance. This finding suggests that the popular in-sample selection criteria are not quite useful in neural network time series forecasting.

The failure of the various penalty-based in-sample criteria in identification of the best model for forecasting indicates their inadequacy for practical use in neural network applications. The reason that they are not adequate may be that both AIC and BIC are originally derived for traditional statistical models, where the number of parameters is usually small. Because of the large number of parameters typically with neural network models, the penalty term in AIC or BIC can overly dominate, which results in a model that emphasizes too much on the parsimony part. In other words, the high penalty may cause underfitting of neural networks.

This study also shows that good in-sample fit has no direct relationship to the out-of-sample performance. Based on both penalty-based criteria and nonpenalty-related measures, the relationship is hardly discernible. This finding is in line with the observations made by several other researchers. For example, Makridakis (1986) and Makridakis and Winkler (1989) find that the correlation between in-sample fit and out-of-sample forecasting is only about 0.2 based on a vast amount of empirical evidence. In our opinion, the low connection between in-sample and out-of-sample performance measures is due to the model uncertainty commonly occurred in statistical data analysis (Chatfield, 1995) and particularly in time series analysis and forecasting (Chatfield, 1996). Chatfield (1996) points out that model uncertainty comes from three main sources: model structure, parameter estimation and data. The nonlinear nonparametric nature of ANNs may cause more uncertainties in neural network model building. That is, an ANN model can give a very good fit to the in-sample but poor forecasts out of sample. This learning and generalization dilemma or tradeoff has been studied extensively and is still an active research topic in the field.

Fildes and Makridakis (1995) have pointed out that "if a close relationship between model fit and out-of-sample forecasts does not exist it is hard to argue that model selection should be based on minimizing model fitting errors". To improve generalization performance of neural network models, one may need to go beyond the model selection methods including the cross-validation and model selection criteria approaches. Efforts can be made along the lines of hint (Abu-Mostafa, 1994; Garcia and Gencay, 2000), Bayesian regularization (Mackay, 1992), Vapnik–Chervonenkis dimension analysis (Cherkassky et al., 1999), and support vector machines (Raudys, 2000; Keerthi et al., 2000).

Acknowledgements

We wish to thank Ramazan Gencay, James G. MacKinnon, three anonymous referees, and the

conference participants of the 1998 North American Summer Meeting of the Econometric Society for comments, suggestions, and discussions. Financial supports for Min Qi from the College of Business Administration at Kent State University and G. Peter Zhang from the J. Mack Robinson College of Business at Georgia State University are gratefully acknowledged. Any remaining errors are our own responsibility.

References

Abu-Mostafa, Y., 1994. Learning from hint. Journal of Complexity 10, 165–178.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 716–723.

Amemiya, T., 1980. Selection of regressors. International Economic Review 21, 331–354.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd ed., Springer, New York.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York.

Cameron, S., 1993. Why is the *R* squared adjusted reported. Journal of Quantitative Economics 9, 183–186.

Chatfield, C., 1995. Model uncertainty: Data mining and statistical inference (with discussion). Journal of the Royal Statistical Society A 158, 419–466.

Chatfield, C., 1996. Model uncertainty and forecast accuracy. Journal of Forecasting 15, 495–508.

Cherkassky, V., Mulier, F.M., Shao, X., 1999. Model complexity control for regression using VC generalization bounds. IEEE Transactions on Neural Networks 10, 1075–1089.

Cottrell, M., Girard, B., Girard, Y., Mangeas, M., Muller, C., 1995. Neural modeling for time series: A statistical stepwise method for weight elimination. IEEE Transactions on Neural Networks 6, 1355–1364.

De Gooijer, J.G., Abraham, B., Gould, A., Robinson, L., 1985. Methods for determining the order of an autoregressive-moving average process: a survey. International Statistical Review 53, 301–329.

Gooijer, J.G., Kumar, K., 1992. Some recent developments in nonlinear time series modeling, testing, and forecasting. International Journal of Forecasting 8, 135–156.

De Groot, C., Würtz, D., 1992. Analysis of univariate time series with connectionist nets: A case study of two classical examples. Neurocomputing 3, 177–192.

Faraway, J., 1992. On the cost of data analysis. Journal of Computational and Graphical Statistics 1, 213–229.

Faraway, J., Chatfield, C., 1998. Time series forecasting with neural networks: A comparative study using the airline data. Applied Statistics 231–250.

Fildes, R., Makridakis, S., 1995. The impact of empirical accuracy studies on time series analysis and forecasting. International Statistical Review 63, 289–308.

Franses, P.H, Draisma, G., 1997. Recognizing changing seasonal patterns using artificial neural networks. Journal of Econometrics 81, 273–280.

Garcia, R., Gencay, R., 2000. Pricing and hedging derivative securities with neural networks and a homogeneity hint. Journal of Econometrics 94, 93–115.

Granger, C.W.J., 1993. Strategies for modelling nonlinear time-series relationships. The Economic Record 69, 233–238.

Hocking, R.R., 1976. The analysis and selection of variables in multiple regression. Biometrics 32, 1–49.

Keerthi, S.S., Bhattacharyya, C., Shevade, S.K., 2000. A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Transactions on Neural Networks 11 (1), 124–136.

LeBaron, B., Weigend, A.S., 1998. A bootstrap evaluation of the effect of data splitting on financial time series. IEEE Transactions on Neural Networks 9, 213–220.

Mackay, D.J.C., 1992. Bayesian interpolation. Neural Computation 4, 415–447.

Makridakis, S., 1986. The art and science of forecasting; an assessment and future directions. International Journal of Forecasting 2, 15–39.

Makridakis, S., Winkler, R.L., 1989. Sampling distributions of post-sample forecasting errors. Applied Statistics 38, 331–342.

Mallows, C.P., 1973. Some comments on Cp. Technometrics 15, 661–675.

Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: do they fit out-of-sample. Journal of International Economics 14, 3–24.

Phillips, P.C.B., 1992. Bayesian model selection and prediction with empirical applications, Cowles Foundation Discussion Paper No. 1023, Yale University, New Haven, CT.

Qi, M., 1996. Financial applications of artificial neural networks. In: Maddala, G.S, Rao, C.R., (Eds.), Handbook of Statistics 14: Statistical Methods in Finance, Elsevier, Amsterdam, pp. 529–552.

Raudys, S., 2000. How good are support vector machines. Neural Networks 13, 17–19.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14, 465–471.

Rissanen, J., 1980. Consistent order-estimates of autoregressive processes by shortest description of data. In: Analysis and Optimization of Stochastic Systems, Jacobs, Ed. O., Davis, M., Dempster, M., Harris, C, Parks, P. (Eds.), Academic Press, New York, pp. 451–461.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Smith, M., 1993. Neural Networks for Statistical Modeling. Van Nostrand Reinhold, New York.

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

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

White, H., 1989. Learning in artificial neural networks: A statistical perspective. Neural Computation 1, 425–464.

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