On SETAR Non-linearity and Forecasting

Michael P. Clements*
Department of Economics,
University of Warwick

Philip Hans Franses
Econometric Institute,
Erasmus University Rotterdam.

Jeremy Smith
Department of Economics,
University of Warwick

March 12, 1999

Abstract

We consider the usefulness of the two-regime SETAR model for out-of-sample forecasting, and compare it with a linear AR model. A range of newly-developed forecast evaluation techniques are employed. Our simulation results show that time-series data need to exhibit a substantial degree of non-linearity before the SETAR model is favoured on some of these criteria. We find only weak evidence that a SETAR model of US GNP provides more accurate forecasts than a linear AR model.

1 Introduction

Switching-regime models (such as the threshold and Markov switching autoregressive models) have been used in contemporary empirical macroeconomics to characterize asymmetries between the expansionary and contractionary phases of business cycles. However, from a forecasting perspective, there appears to be no clear consensus as to whether allowing for non-linearities of these types leads to an improved forecast performance (see, e.g., De Gooijer and Kumar, 1992). Part of the reason is the equivocal support from empirical forecast accuracy comparisons when the non-linear features that were prominent in-sample fail to carry over to the out-of-sample period. Forecast comparisons of US GNP where the out-of-sample period is taken to be 1992 onwards is a case in point (see, Clements and Smith, 1997). However, Clements and Krolzig (1998) and Clements and Smith (1999) attempt systematic, simulation-based evaluations of the forecast performance of non-linear models that seek to ensure that the future bears the same non-linear imprint as the past. The general finding turns out to be that linear autoregressive models are reasonably robust forecast devices.

A number of authors have suggested (e.g., Pesaran and Potter, 1997, Clements and Smith, 1999) that

*Financial support from the U.K. Economic and Social Research Council under grant L116251015 is gratefully acknowledged.

non-linear models may do better on forecast comparisons other than the (root) mean squared forecast error ((R)MSFE) of point forecasts: Pesaran and Potter (1997) look at conditional variances, and Clements and Smith (1999) consider MSFE calculations conditional on being in a particular regime, and direction-of-change statistics. The study by Pesaran and Potter (1997) in particular suggests that non-linear models may be better at forecasting higher-order moments, which may translate in to non-linear models producing superior interval and density forecasts to those of linear competitors. Recently, methods of evaluating such forecasts have been proposed.

The aims of this paper are to gauge the 'degree' of non-linearity that needs to be present before the forecasts of non-linear models outperform linear rivals, and to assess whether, given the sample sizes typically available in empirical macroeconomics, density and interval forecasts from non-linear models are likely to be demonstrably better than those of linear models. The 'degree of non-linearity' is measured in this paper by the significance level at which a test of the null of linearity is rejected, i.e., the p-value. The non-linear model we choose is the self-exciting threshold autoregressive (SETAR) model, which has been widely used in empirical research. It is a special case of the smooth 'threshold' or smooth 'transition' model popularised in the econometrics literature by Teräsvirta, see, e.g., Granger and Teräsvirta (1993).

The plan of the paper is as follows. In section 2 we briefly present the SETAR model. Section 3 notes the technical problems that arise when testing for threshold non-linearities, and the testing framework adopted. Section 4 reviews forecast evaluation techniques for point, interval and density forecasts. Section 5 describes how the various types of forecast can be constructed for the AR and SETAR models, and how the forecast evaluation techniques are applied. It also describes the design of the Monte Carlo, including the forecast sampling scheme, and the choice of data generating processes (DGPs) for the Monte Carlo. A discussion of the results follows in section 6, followed by an empirical illustration in section 7, and section 8 concludes.

2 SETAR models

The self-exciting threshold autoregressive (SETAR) model assumes that a variable y_t is a linear autoregression within a regime, but may move between regimes depending on the value taken by a lag of y_t, say, y_{t-d}, so that d is the length of the delay. Hence the model is linear within a regime, but liable to move between regimes as the process crosses a threshold: see Tong (1978, 1983, 1995) and Tong and Lim (1980), but a switching regression model with an observable switching variable appears in Goldfeld and Quandt (1972). When there are only two regimes, and the process is a p^{th} order autoregression in each, we convey this information more succinctly by writing SETAR$(2; p, p)$, and algebraically:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \sigma_1 \epsilon_t + \mathrm{I}\left(y_{t-d} > r\right)\left(\beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \left(\sigma_2 - \sigma_1\right)\epsilon_t\right). \tag{1}$$

Letting $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_p]'$, $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]'$ and $\mathbf{x}'_{t-1} = [1, y_{t-1}, \ldots, y_{t-p}]$ we can write:

$$y_t = \mathbf{x}'_{t-1}\boldsymbol{\alpha} + \mathtt{I}_d(r)\,\mathbf{x}'_{t-1}\boldsymbol{\beta} + (\sigma_1 + \mathtt{I}_d(r)(\sigma_2 - \sigma_1))\,\epsilon_t \tag{2}$$

where $\mathtt{I}_d(r) \equiv \mathtt{I}(y_{t-d} > r) = 1$ if $y_{t-d} > r$ and zero otherwise.

SETAR models have been used, sometimes with spectacular success, to forecast certain biological or physical processes, such as the Canadian lynx data and Wolf's sunspot numbers (see, for example Tong, 1995, chapter 7). SETAR models have also been applied to a number of economic and financial variables. For example, Kräger and Kugler (1993), Peel and Speight (1994) and Chappell, Padmore, Mistry and Ellis (1996) apply such models to the foreign exchange market, Tiao and Tsay (1994) and Potter (1995) to US GNP, and Montgomery, Zarnowitz, Tsay and Tiao (1998) and Rothman (1998) to unemployment rates.

3 Testing SETAR models

Hansen (1996) presents a general framework for testing the null of linearity against the alternative of threshold autoregression, that delivers valid inference when the threshold value r and delay d are unknown *a priori*, in the sense that they have to be learnt from the data (either by a formal estimation procedure or by casual inspection, as in Potter, 1995). r and d are nuisance parameters that are unidentified under the null hypothesis so that the testing procedure is non-standard. For example, the null of linearity in (1) is given by the parametric restriction $\boldsymbol{\beta} = \mathbf{0}$, in which case neither r nor d are identified. For given values of r and d the usual testing principles, LR, LM or Wald, will deliver test statistics which are asymptotically χ^2. But when the assumed values of r and d are far from their true values, under the alternative, $\boldsymbol{\beta} \neq \mathbf{0}$, the LR, LM and Wald tests lack power. The approach advocated by Hansen (1996) is to calculate the statistics over a grid of values for the nuisance parameters, and then to take the supremum or some average of the statistics – the supremum, supW_γ, was considered by Davies (1977, 1987) as a way of testing the null, and the average aveW_γ and expW_γ ($\ln$ of the average of $\exp(W_\gamma/2)$), have recently been considered by Andrews and Ploberger (1994). Here, γ is the vector of nuisance parameters, so $\gamma = [r, d]'$ for the SETAR model, and the averaging or $\sup$ is taken over all admissible values of γ, $\gamma \in \Gamma$. Hansen (1996) shows how asymptotic p-values can be simulated for these functions of W_γ

4 Forecast evaluation techniques

Until recently, macroeconomics has focused primarily on the production and evaluation of point forecasts (see, e.g., Wallis, 1995), with only summary information regarding the degree of uncertainty associated with forecasts, such as standard errors, being provided. However, Christoffersen (1998) suggests ways of evaluating interval forecasts, and continuing this trend toward providing a more complete description of the uncertainty surrounding forecasts, Diebold, Gunther and Tay (1998) propose methods for evaluating density forecasts. The following sections review these and other methods.

4.1 Point forecasts

Although there is a large literature on forecast accuracy comparisons, until Diebold and Mariano (1995) there were few formal comparisons of rival forecasts of the same phenomena, where by 'formal', we mean comparisons which attempt to assess whether differences between rival forecasts can be attributed to sampling variability, or whether they are 'significant'. We begin by reviewing the Diebold and Mariano (1995) test of the null of no difference between two forecasts. They propose a test of the null of equal forecast accuracy, for some arbitrary loss function, $g(e_{i,t})$, where here $e_{i,t}$ is an h-step ahead forecast error from using model i. The loss differential is defined as $d_t \equiv [g(e_{i,t}) - g(e_{j,t})]$ for rival forecasts i and j, so that equal forecast accuracy entails $\mathsf{E}[d_t] = 0$. Given a covariance-stationary sample realization $\{d_t\}$ of n observations, the asymptotic distribution of the sample mean loss differential $\overline{d}$ ($\overline{d} = n^{-1} \sum_{j=1}^{n} d_j$) is given by:

$$\sqrt{n}\left(\overline{d} - \mu\right) \xrightarrow{D} \mathsf{N}\left[0, 2\pi f_d\left(0\right)\right]$$

where:

$$f_d\left(0\right) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d\left(\tau\right),$$

is the spectral density of the loss differential at frequency zero, and γ_d is the autocovariance function.

Thus the large-sample statistic that Diebold and Mariano (1995) propose for testing the null of equal forecast accuracy is:

$$\frac{\overline{d}}{\sqrt{n^{-1} 2\pi \widehat{f}_d\left(0\right)}} \underset{app}{\sim} \mathsf{N}\left[0, 1\right],$$

where $\widehat{f}_d(0)$ is a consistent estimate of $f_d(0)$, based on a weighted sum of the sample autocovariances. Diebold and Mariano (1995) discuss the choice of lag window and truncation lag. Harvey, Leybourne and Newbold (1997) propose a modified version of this test statistic that corrects for the tendency of the statistic to be over-sized, and present simulation evidence that attests to the usefulness of the modification. But given the (relatively) large number of forecasts (minimum of a hundred) at our disposal, and since we consider only 1-step forecasts, so that serial correlation is ignored, we use:

$$\hat{\gamma}_0 = n^{-1} \sum_{j=1}^{n} \left(d_j - \overline{d}\right)^2$$

$$\frac{\overline{d}}{\sqrt{\hat{\gamma}_0}} \underset{app}{\sim} \mathsf{N}\left[0, 1\right].$$

4.2 Forecast encompassing

For two rival 1-step forecasts, $y_{i,t}$ and $y_{j,t}$, the forecast encompassing test is implemented by regressing $e_{i,t}$ ($= y_t - y_{i,t}$) on $y_{i,t} - y_{j,t}$:

$$y_t - y_{i,t} = \alpha\left(y_{i,t} - y_{j,t}\right) + \zeta_t \tag{3}$$

and using a t-test of H_0: $\alpha = 0$. This is equivalent to the regression:

$$y_t = (1+\alpha)y_{i,t} - \alpha y_{j,t} + \zeta_t. \tag{4}$$

and is a variant of the test due to Chong and Hendry (1986): see Clements and Hendry (1998). Harvey, Leybourne and Newbold (1998) show that this test may be over-sized if the forecast errors are not approximately normal, and suggest a number of alternatives, which include:

$$R_1 = \frac{1}{\sqrt{n}} \hat{Q}_1^{-\frac{1}{2}} \sum_{t=1}^{n} (e_{it} - e_{jt})^2 \, \hat{\alpha}$$

where:

$$\hat{Q}_1 = \frac{1}{n} \sum_{t=1}^{n} (e_{it} - e_{jt})^2 \, \hat{\zeta}_t^2$$

and $\hat{\alpha}$ and $\hat{\zeta}_t$ are the OLS estimates of α and the residuals from (3). Harvey *et al.* (1998) also show that the Diebold - Mariano test of equal MSFE is closely related to the notion of forecast encompassing.

4.3 Forecast failure

Forecast failure is said to occur when a model's forecast performance is significantly worse than was expected based on the model's explanation of the historical data, judged by its in-sample fit. We calculate the following test (see, e.g., Clements and Hendry, 1999):

$$\widehat{Q} = \frac{{e_i}^2_{T+l}}{\widehat{\sigma}_i^2 \left(\frac{T-p}{T}\right)} \sim \mathsf{F}^1_{T-p-1}, \tag{5}$$

where $\widehat{\sigma}_i$ is the in-sample standard error, p is the number of explanatory variables (for the AR: the number of lags plus the intercept; for the SETAR: the sum of the number of AR parameters in each regime plus four: two intercepts, r and d).

4.4 Interval evaluation

Christoffersen (1998) suggests that a 'good' interval forecast should have correct conditional coverage, such that, the interval is wider in volatile periods than in those of greater tranquility. Otherwise we expect that the occurrences of observations outside the interval would be clustered in volatile periods and completely absent in tranquil periods. Christoffersen (1998, p.844) develops a 'unified framework for testing the conditional coverage' using three LR tests for unconditional coverage, independence and conditional coverage.

Let $L_{t|t-1}(p)$ and $U_{t|t-1}(p)$ denote the lower and upper limits of the interval forecast of y_t made at time $t-1$, for a coverage probability p. Then define the indicator function, I_t, which takes the value one when y_t lies inside the interval range and zero otherwise. For $t = 1, \ldots, n$ 1-step ahead forecasts, we have a sequence $I_{t=1}^n$. The test of unconditional coverage uses an LR test to compare the coverage probability, p, with the sample proportion of times $\widehat{\pi}$ that the forecast interval includes y_t. The test for independence uses an LR test to compare the sample transition probabilities, $\widehat{\pi_{ij}} = Pr(I_t = j|I_{t-1} = i)$ for $i, j = 0, 1$, with the transition probabilities obtained assuming independence (e.g., $\pi_{01} = \pi_{11} = p$). Finally, the test for conditional coverage adds together the LR tests for unconditional coverage and independence.

In section 6 the interval forecasts for the Christoffersen (1998) test are based on $p = 0.75$ and the intervals are obtained by two alternative methods, either by drawing random errors assuming gaussianity, with the variance taken as the within-sample regime-specific error variance, or by bootstrapping the residuals.

4.5 Density evaluation

Diebold *et al.* (1998) suggest calculating the probability integral transforms of the actual realizations of the variables over the forecast period with respect to the forecast densities of the SETAR and AR models, denoted by $\{p_t(y_t)\}_{t=1}^n$. That is, we evaluate:

$$\{z_t\}_{t=1}^n = \left\{ \int_{-\infty}^{y_t} p_t(u)du \right\}_{t=1}^n . \tag{6}$$

When the model forecast density corresponds to the true predictive density (given by the DGP, and denoted by $f_t(y_t)$), i.e., $p_t(y_t) = f_t(y_t)$, then $\{z_t\}_{t=1}^n \sim \text{iid}U[0,1]$. In a Monte Carlo context, a simple way of testing the uniformity part of the hypothesis *conditional* on the assumption of independence is via a Kolmogorov-Smirnov (KS) test, which is based on the maximum deviation between the empirical distribution function of the z_t and the theoretical distribution function. When independence fails the appropriate critical values for the KS tests are unknown (but could be bootstrapped). Then, a potentially attractive approach is that of Berkowitz (1999), who suggests taking the inverse normal CDF transformation of the $\{z_t\}$ series, to give, say, $\{z_t^*\}$. Thus, the $\{z_t\}$, which are $\text{iid}U[0,1]$ under the null, become standard normal variates. Berkowitz argues that more powerful tools can be applied to testing a null of $\text{iid}N(0,1)$ compared to one of iid uniformity, and proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degree of freedom test of zero-mean, unit variance and independence. But in each the maintained assumption is that of normality, while in some circumstances a test of the normality assumption itself may prove more powerful.[1] Consequently, we report the normality test recommended by Doornik and Hansen (1994) based on Shenton and Bowman (1977).

4.6 Calculation of SETAR and AR model forecasts

Optimal 1-step ahead forecasts form the SETAR are simple to derive. From (2):

[1] We are grateful to Ken Wallis for bringing this to our attention.

$$\begin{aligned} \mathsf{E}\left[y_{t+1} \mid \mathcal{I}_t\right] &= \mathsf{E}\left[\mathbf{x}_t'\boldsymbol{\alpha} + \mathtt{I}_d\left(r\right)\mathbf{x}_t'\boldsymbol{\beta} + \left(\sigma_1 + \mathtt{I}_d\left(r\right)\left(\sigma_2 - \sigma_1\right)\right)\epsilon_{t+1}\right] & (7) \\ &= \mathbf{x}_t'\boldsymbol{\alpha} + \mathtt{I}_d\left(r\right)\mathbf{x}_t'\boldsymbol{\beta}. & (8) \end{aligned}$$

For $d = 1$ and horizons in excess of 1, though, exact analytic forecasts are not available: see, e.g., Granger and Teräsvirta (1993), De Gooijer and De Bruin (1997).

Interval and density forecasts can be calculated either by assuming that the disturbances are gaussian (conditional on the regime, for the SETAR) or by some type of bootstrapping. The 75% gaussian interval for the AR is defined by the lower and upper values $\{\widehat{y}_{t+1} - 1.15 \times \widehat{\sigma}_t, \quad \widehat{y}_{t+1} + 1.15 \times \widehat{\sigma}_t\}$, where $\widehat{y}_{t+1}$ is a 1-step ahead forecast of the value of the process in y_{t+1}, $\widehat{\sigma}_t$ is the standard error for the AR estimated up to period t. The gaussian forecast density is of course $p_{t+1}\left(y_{t+1}\right) = N\left(\widehat{y}_{t+1}, \widehat{\sigma}_t^2\right)$. The corresponding quantities for the SETAR follow in the same way, but now the standard error depends on the regime. Thus, the gaussian probability integral transforms (p.i.t.), i.e., the $\{z_t\}$ sequence, are given by $z_{t+1} = \Phi\left(\left(y_{t+1} - \widehat{y}_{t+1}\right)/\widehat{\sigma}_t\right)$. The Berkowitz series, $\{z_t^*\}$, are then just the original (scaled) errors, $\left(y_{t+1} - \widehat{y}_{t+1}\right)/\widehat{\sigma}_t$, since these are $N\left(0,1\right)$ under the null.

The bootstrap approach assumes that the forecast densities are the empirical distribution functions of the (regime specific for the SETAR) estimated residuals. We do not calculate $\{z_t^*\}$ series for p.i.t.s obtained in this way.

5 Monte Carlo

A number of sampling schemes (see, e.g., West and McCracken, 1998, pp.818-9) could be adopted: we use a 'recursive' scheme. As the forecast origin moves forward through the sample, the model order is re-specified and the parameters re-estimated in each period. Thus for an observation vector of length T, the model is first specified and estimated on the data up to period T_0 ($T_0 < T$), and a forecast (point, interval and density) of $T_0 + 1$ is made. The analysis is confined to 1-step ahead forecasts. Then, the model is re-specified and re-estimated on data up to and including $T_0 + 1$, and forecasts of $T_0 + 2$ are made. This continues up to a forecast of T made from models specified and estimated on data up to $T - 1$. Thus we generate a sample of $T - T_0$ 1-step forecasts. We choose $T = 200$ and 300, but set $T_0 = 100$. The former represents fifty years of quarterly data, with the sample split in half, so that 1-step forecasts are calculated recursively over the last twenty five years. This is just feasible for post War macroeconomic data sets, while the latter is possible for higher frequency data. However, a forecast sample of 200 observations is used to investigate whether even as the forecast sample becomes very large the alternative forecast evaluation techniques are capable of rewarding the non-linear model relative to the linear model. It would be of interest to allow both T and T_0 to change. Here we focus on $T - T_0$ for fixed T_0, i.e., the number of forecasts, but we note that the impact of parameter estimation (and model selection) uncertainty may be important (see, e.g., West, 1996, West and McCracken, 1998), and this could be assessed by varying T_0.

The choice of data generating process is based on Hansen (1996), who investigated the size and power properties of the test of non-linearity discussed in section 3. Tests of non-linearity for the DGPs used noted in table 1 have powers that range from around 5% (for the linear AR, i.e., size) up to unity, and so permit an assessment of the relative gain of the SETAR, over the AR, for varying 'degrees of non-linearity'.

The DGPs are all first order, so to reduce computational time we set the SETAR model order to one. The maximum lag order for the AR is set at three, to allow for the fact that it is mis-specified. $d = 1$, while r is chosen by a grid search to be close to the median of the generated series. For $T = 200$ we set the number of Monte Carlo replications to 1000, and for $T = 300$, the number of replications is 500.

6 Discussion

The testing for non-linearity is undertaken for the in-sample period, $T_0 = 100$. The results are shown in table 2 for the seven sets of design parameters listed in table 1. We report the averages across the Monte Carlo replications of the p-values of the null of linearity for each experiment. Experiment 1 is a linear AR. Experiments 2 to 4 are SETAR processes for which the intercept, but not the slope, changes between regimes. Exp. 2 is the 'small' intercept change, exp. 3 the 'large' intercept change, both for processes with regime-dependent error variances, and exp. 4 is a 'large' intercept change with homoscedastic disturbances. Going from exp. 1 to 4 the 'degree of non-linearity' increases. In particular, when the null of linearity is true, (exp. 1) the p-values are uniformly distributed on the $[0, 1]$ interval (Hansen (1996)): thus the average is close to 0.5. For exp. 2 the average p-value is around 0.35, for exp. 3, around 0.2. For exp. 4 it is in the region 0.05 to 0.10. The intercept-change SETAR with homoscedastic disturbances has lower average p-values than under heteroscedasticity. Experiments 5 to 7 are similar to those of experiments 2 to 4, but for a slope-change SETAR, and the average p-values are 0.2, less than 0.05, and 0.05 to 0.10. Now homoscedasticity makes it a little more difficult to reject linearity (c.f., exp. 7 and 6).

We now consider how these varying degrees of rejection of the null hypothesis of linearity translate in to forecast gains for the SETAR model relative to the misspecified AR model. We begin with the evaluation of the point forecasts set out in table 3. For $T = 300$ (200 forecasts) average p-values from table 2 of around 20% (exp. 3) allow the null, that the SETAR is no more accurate than the AR (col. c), to be rejected around 20% of the time at the 5% level using the Diebold - Mariano test of Equal MSFE, and around 25% of the time using the "$\%$ better" test (col. f). (The "$\%$ better" tests whether one model has a smaller absolute error statistically more/less often than the other). These rejection rates of the Diebold - Mariano test approximately halve when there are only 100 forecasts, a result which holds across almost all experiments 2 to 7. For average p-values of 0.05 to 0.10, (exp. 4) and $T = 300$, the rejection rates of these two tests (columns (c) and (f)) are around 50% and 40%. The small slope change (exp. 5) is similar to the large intercept change (exp. 3) in terms of the test of non-linearity, and leads to similar rejection frequencies (RFs) on the Diebold - Mariano and "$\%$ better" tests. The large slope change with heteroscedastic errors (exp. 6) leads to rejections in favour of the SETAR of over 70% and 50% on these two tests, compared to 40% and 30% when the errors are homoscedastic. There is no tendency for the AR model to be more accurate (col. b), or favoured on "$\%$ better" (col. e), except for exp. 1.

For experiments 2 to 7 the RFs of the null that the AR forecast encompasses the SETAR (col. g) are generally high, while those for the null that the SETAR encompasses the AR are generally less than 15% in most cases. Thus, some 85% of the time we find that a combination of the AR and SETAR forecasts is no more accurate than the SETAR forecast alone. A prominent feature of table 3 is that for the forecast failure test there is no evidence against the AR model for any of these DGPs (col. i): this is stark support for the contention in Clements and Hendry (1999) that model mis-specification is not sufficient forecast failure.

Table 4 records the results of the interval evaluation exercise. It is apparent that, despite having a large number of forecasts, this number is still not sufficient for the AR intervals to be rejected much above 25% of the time at best. This rejection rate occurs for the Gaussian AR interval for exp. 3 when $T = 300$. Interestingly, while the 'large' intercept change with homoscedastic errors DGP (exp. 4) (relative to when the errors are heteroscedastic, as in exp. 3) allows linearity to be rejected more convincingly (see 2), the interval RF drops back to the nominal size. There is little power when the intervals are bootstrapped (rather than calculated assuming gaussianity). These findings are not surprising given the discussions in Christoffersen (1998) and Berkowitz (1999), which suggest that interval forecast calculation and evaluation are only likely to be useful for a very large number of forecasts, implying these methods may be largely inappropriate even for monthly macroeconomic data.

Table 5 records the results of the density forecast evaluation. The greatest power for the KS tests is around 15% for exp. 3 when $T = 300$. This matches the finding for interval evaluation. The powers for the bootstrapped intervals are close to the nominal size. The KS test, which tests the assumption that the $\{z_t\}$ follow a uniform distribution, assuming that they are iid, has little power for AR model forecast densities of SETAR-generated data. Berkowitz's suggestion to transform the $\{z_t\}$ to $N(0,1)$ variables (under the null), and then test for independence, and jointly for $IN(0,1)$, conditional on the normality assumption, also receives little support. We achieve reasonable rejection rates (particularly for the intercept shifts: exps. 2 and 3) when the assumption of normality is itself tested, provided the SETAR has heteroscedastic disturbances (in exps. 4 and 7 the RF is close to the size). However, the rejection rates for the SETAR model exceed 5%.

7 Illustration: US GNP

We compare the forecast performance of AR and SETAR models of US post-War GNP to see whether the relationship between in-sample p-values of tests of non-linearity and relative forecast gains, analysed by simulation, is useful in practice. US GNP is an obvious choice — Tiao and Tsay (1994), Potter (1995) and Hansen (1996) have all fitted two-regime SETAR models for the period 1947 – 90. In order to have 70 1-step forecasts, the model is initially estimated on data up to 1973, with the AR and SETAR lag orders selected by AIC with 4 the highest allowed in both cases. The results are shown in table 6 for the initial model estimates, and for the full-sample estimates. The full-sample estimates are satisfactory and in accordance with the previous studies mentioned, particularly the large negative coefficient at lag 2 in the lower regime, suggesting a sharp movement from the lower regime to the upper. For the initial period the model is less easy to rationalise, probably due to the relatively small number of observations and the low information content of that data for modelling expansionary and contractionary regimes.

The results of applying the Hansen test for non-linearity to the initial (smallest) in-sample period, and to the full-sample, are recorded in table 7. The statistics suggest p-values of around 50%, so that it is not surprising that there is little to choose between the forecast performances of the two models – see table 8. The normality test of the forecasts densities (i.e., of the forecast errors) rejects at the 5% level for the SETAR and is borderline for the AR, but this could simply be due to a small number of outliers. Otherwise there are no clear rejections of the point, interval or density forecasts of either model. The results for the point forecasts tally with those of other authors, who only tend to find gains if one conditions on the regime, that is, if the forecasts made at times of 'recession' are evaluated by MSFE in isolation. The relatively poor performance of the SETAR model could result in part from parameter non-constancy over this period, exacerbated by the initial sample period being too small to reliably estimate the parameters.

One way of evaluating the forecast performance of the SETAR, conditional on the full-sample model being a reasonable description of the data generating process, is to simulate data from the estimated empirical SETAR model. We simulated 500 sets of data for $T = 350$, set $T_0 = 150$, and evaluated the forecast performance of the two models on this data. The models were again selected and estimated anew for each 1-step forecasts, but the maximum lag orders entertained for the AR and SETAR were 3 and 2, respectively (some experimentation suggested these were seldom binding). The average p-values for the Hansen test of non-linearity on the simulated data were around 0.2 (see table 7). The evaluation of the forecasts of the two models are collected in table 9. It is apparent that there is no evidence against the AR model from the interval and density forecasts. This is predicted by the simulation results given the sizes of the average p-values of the Hansen tests (which roughly correspond to exp. 3). The simulations also suggest that the normality test of the AR density should have reasonable power if the error variances are regime specific (see exp. 3, table 5) – that it does not here is because the regime error variances are quite similar (more closely matching exp. 4, table 5). As expected, the null that the AR model forecast encompasses the SETAR is rejected over 80% of the time, while the converse is rejected only 20% of the time.

8 Conclusions

In this paper we focused on the usefulness of SETAR models for out-of-sample forecasting relative to standard linear AR models. A range of forecast evaluation techniques were employed to judge the relative performances of the two models, and in so doing we critically appraised some of the newly-developed techniques.

Our simulations indicate that the time-series data need to exhibit a high degree of non-linearity before some of the evaluation criteria frequently indicate that the non-linear model outperforms the linear. Interval and density forecast evaluation techniques had relatively low power to reject the linear model, even when 200 forecasts were used. A simple amendment to the density evaluation procedures fared somewhat better, although a normality test of the within sample residuals may, in general, prove to be more powerful again. Forecast encompassing tests of the point forecasts turned out to be the most useful way of discriminating between the models. The results of an empirical application to SETAR and AR models for post-War US GNP growth were broadly in accordance with the simulation results. We found for all criteria, except perhaps for the forecast encompassing tests, only weak evidence that the SETAR model (that has been employed by a number of authors) gives more accurate forecasts than a linear AR model.

There are a number of potential explanations for the findings in this paper, suggesting directions for further research. One is the need for still better forecast evaluation techniques, notwithstanding the progress witnessed in recent years. Qualitative, 'direction-of-change' tests, and regime-dependent evaluation appear promising, see, e.g., Clements and Smith (1999). Alternatively, it may be that other non-linear models, such as Smooth Transition AR models and Artificial Neural Networks, would record more sizeable gains relative to a linear model, when the null of linearity is rejected at levels of probability comparable to those analysed in this paper.

References

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

Berkowitz, J. (1999). Evaluating the forecasts of risk models. mimeo, Federal Reserve Board, Washington, D.C.

Chappell, D., Padmore, J., Mistry, P., and Ellis, C. (1996). A threshold model for the French franc/Deutschmark exchange rate. *Journal of Forecasting*, **15**, 155–164.

Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press. The Marshall Lectures on Economic Forecasting.

Clements, M. P., and Hendry, D. F. (1999). Some methodological implications of forecasting failure. Mimeo, Department of Economics, University of Warwick.

Clements, M. P., and Krolzig, H.-M. (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal*, **1**, C47–75.

Clements, M. P., and Smith, J. (1997). The performance of alternative forecasting methods for SETAR models. *International Journal of Forecasting*, **13**, 463–475.

Clements, M. P., and Smith, J. (1999). A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics*. Forthcoming.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33–43.

De Gooijer, J. G., and De Bruin, P. (1997). On SETAR forecasting. *Statistics and Probability Letters*, **37**, 7–14.

De Gooijer, J. G., and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8**, 135–156.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.

Goldfeld, S. M., and Quandt, R. E. (1972). *Non-linear Methods in Econometrics*. Amsterdam: North-Holland.

Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, **64**, 413–430.

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

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

Kräger, H., and Kugler, P. (1993). Non-linearities in foreign exchange markets: a different perspective.

Journal of International Money and Finance, **12**, 195–208.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). Forecasting the U.S. unemployment rate. *Journal of the American Statistical Association*. Forthcoming.

Peel, D. A., and Speight, A. E. H. (1994). Testing for non-linear dependence in inter-war exchange rates. *Weltwirtschaftliches Archiv*, **130**, 391–417.

Pesaran, M. H., and Potter, S. M. (1997). A floor and ceiling model of US Output. *Journal of Economic Dynamics and Control*, **21**, 661–695.

Potter, S. (1995). A nonlinear approach to U.S. GNP. *Journal of Applied Econometrics*, **10**, 109–125.

Rothman, P. (1998). Forecasting asymmetric unemployment rates. *The Review of Economics and Statistics*, **80**, 164–168.

Shenton, L. R., and Bowman, K. O. (1977). A bivariate model for the distribution of $\sqrt{b_1}$ and b_2. *Journal of the American Statistical Association*, **72**, 206–211.

Tiao, G. C., and Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time-series. *Journal of Forecasting*, **13**, 109–131.

Tong, H. (1978). On a threshold model. In Chen, C. H. (ed.), *Pattern Recognition and Signal Processing*, pp. 101–141. Amsterdam: Sijhoff and Noordoff.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis*: Springer-Verlag, New York.

Tong, H. (1995). *Non-linear Time Series. A Dynamical System Approach*. Oxford: Clarendon Press. First published 1990.

Tong, H., and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of The Royal Statistical Society*, **B 42**, 245–292.

Wallis, K. F. (1995). Large-scale macroeconometric modelling. In Pesaran, M. H., and Wickens, M. R. (eds.), *Handbook of Applied Econometrics: Macroeconomics*: Basil Blackwell.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–1084.

West, K. D., and McCracken, M. W. (1998). Regression-based tests of predicitive ability. *International Economic Review*, **39**, 817–840.

Table 1 Monte Carlo design parmameter values.

Experiment	α_0	β_0	α_1	β_1	σ_1	σ_2
(1)	0	0	0.3	0	1	1
(2)	-0.75	0.75	0.3	0	1	2
(3)	-1.25	1.25	0.3	0	1	2
(4)	-1.25	1.25	0.3	0	1	1
(5)	0	0	-0.3	0.6	1	2
(6)	0	0	-0.7	1	1	2
(7)	0	0	-0.7	1	1	1

Given equation (1), α_0 and α_1 are the intercept and slope in regime 1 ($y_{t-d} \leq r$), and $\alpha_0 + \beta_0$ and $\alpha_1 + \beta_1$ are the same quantities for regime 2 ($y_{t-d} > r$).
σ_1 and σ_2 are the standard errors of the disturbance term in regimes 1, and 2, respectively.
Exp. (1) is a linear AR model. (2) and (3) have medium and large intercept shifts with regime-dependent disturbances; (4) has the large shift with homoscedastic disturbances. (5) and (6) have medium and large slope shifts with heteroscedastic disturbances; (7) is (6) but with homoscedastic errors.

Table 2 Hansen Test average p-values (across Monte Carlo replications) .

	Robust LM Statistics		
	SupT_T	ExpT_T	AveT_T
1	0.480	0.486	0.478
2	0.351	0.345	0.341
3	0.208	0.212	0.227
4	0.069	0.078	0.103
5	0.230	0.188	0.169
6	0.048	0.032	0.027
7	0.090	0.059	0.045

The sample size was 100. The results were obtained using Bruce Hansen's Gauss code `tar.prg`.

Table 3 Point forecast evaluation.

	Equal MSFE			% better			Forecast Encomp.		Forecast Failure	
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
					$T_0 = 100, T = 200$					
(1)	.097	.177	.004	.067	.102	.012	.072	.459	.046	.057
(2)	.039	.044	.051	.068	.013	.097	.212	.141	.070	.059
(3)	.066	.004	.116	.099	.009	.135	.423	.128	.067	.051
(4)	.194	.006	.293	.142	.001	.183	.814	.143	.053	.068
(5)	.072	.015	.115	.089	.017	.103	.448	.116	.062	.059
(6)	.305	.001	.441	.192	.000	.271	.925	.088	.053	.058
(7)	.135	.005	.210	.100	.009	.044	.749	.193	.044	.048
					$T_0 = 100, T = 300$					
(1)	.148	.232	.004	.060	.108	.004	.084	.674	.046	.058
(2)	.024	.006	.062	.076	.012	.124	.356	.152	.074	.054
(3)	.116	.000	.192	.178	.000	.258	.710	.144	.060	.066
(4)	.420	.000	.516	.286	.000	.394	.976	.154	.058	.074
(5)	.112	.002	.190	.126	.010	.170	.654	.126	.058	.042
(6)	.614	.000	.736	.424	.000	.536	.998	.082	.054	.056
(7)	.274	.000	.392	.180	.002	.290	.946	.240	.068	.062

Columns (a) — (c) report rejection frequencies (RFs) of Diebold – Mariano tests of equal forecast accuracy on MSFE. Col. (a) is a two-sided test at the 5% level. Col. (b) is a one-sided test of equal accuracy versus AR model more accurate; col. (c) is a one-sided test against the alternative of the SETAR being more accurate.

Columns (d) — (f) report RFs of tests of "$\%$ better" at the 5% level: a two-sided test; a one-sided test against the alternative that the AR is favoured; and a one-sided test against the alternative that the SETAR is favoured.

Columns (g) and (h) are RFs of forecast encompassing tests at the 5% level: that the AR encompasses the SETAR; and the SETAR encompasses the AR. We use the modified form of the test given in Harvey *et al.* (1998) as their equation (15), which makes an allowance for non-normal forecast errors.

Columns (i) and (j) are RFs of forecast failure tests, at the 5% level for the AR and SETAR models.

Full details are given in the text.

Table 4 Interval Forecast evaluation: 75% coverage intervals.

	AR model						SETAR model					
	Gaussian interval			Bootstrapped interval			Gaussian interval			Bootstrapped interval		
	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.
$T_0 = 100$, $T = 200$												
1	.072	.055	.054	.062	.042	.047	.116	.045	.091	.124	.047	.099
2	.113	.083	.118	.072	.086	.095	.082	.073	.079	.092	.067	.080
3	.142	.096	.165	.077	.102	.113	.068	.055	.069	.083	.059	.079
4	.050	.054	.053	.072	.043	.063	.076	.058	.072	.106	.055	.084
5	.101	.093	.117	.067	.087	.085	.064	.054	.065	.075	.055	.073
6	.088	.115	.129	.097	.113	.126	.077	.054	.063	.074	.049	.062
7	.074	.062	.073	.070	.071	.076	.097	.058	.092	.133	.052	.109
$T_0 = 100$, $T = 300$												
1	.044	.050	.048	.048	.046	.042	.080	.056	.084	.110	.054	.120
2	.166	.090	.150	.082	.082	.084	.080	.054	.082	.098	.056	.080
3	.234	.154	.264	.088	.152	.150	.058	.064	.082	.084	.064	.102
4	.064	.070	.066	.058	.052	.074	.080	.048	.080	.104	.060	.092
5	.140	.062	.116	.070	.086	.086	.080	.068	.096	.096	.060	.096
6	.148	.152	.172	.094	.120	.138	.068	.064	.080	.082	.052	.076
7	.076	.054	.084	.060	.076	.098	.104	.058	.088	.112	.058	.098

The elements are the rejection frequencies (at the 5% level) of the Christoffersen tests of unconditional coverage, independence, and conditional coverage of 75% intervals.

Table 5 Density Forecast evaluation.

	Kolmogorov-Smirnov 5% level				Berkowitz 5% level				Normality 5% level	
	AR		SETAR		AR		SETAR		AR	SETAR
	Gauss.	Bootst.	Gauss.	Bootst.	Indep.	Joint	Indep.	Joint		
					$T_0 = 100, T = 200$					
(1)	.028	.039	.040	.071	.097	.062	.094	.117	.044	.098
(2)	.097	.056	.043	.075	.100	.121	.114	.100	.477	.164
(3)	.086	.060	.037	.063	.079	.160	.088	.111	.564	.164
(4)	.041	.046	.045	.080	.077	.067	.102	.098	.070	.081
(5)	.068	.060	.044	.080	.139	.169	.091	.132	.389	.187
(6)	.062	.053	.036	.060	.159	.201	.083	.110	.286	.168
(7)	.037	.044	.033	.069	.086	.070	.097	.115	.053	.091
					$T_0 = 100, T = 300$					
(1)	.026	.028	.030	.058	.118	.054	.092	.128	.054	.098
(2)	.110	.050	.032	.088	.098	.142	.088	.150	.722	.206
(3)	.150	.060	.050	.084	.074	.164	.080	.126	.772	.204
(4)	.038	.050	.038	.088	.060	.062	.068	.116	.062	.088
(5)	.088	.062	.038	.068	.116	.138	.062	.106	.588	.214
(6)	.072	.042	.060	.074	.154	.226	.062	.154	.420	.212
(7)	.048	.048	.042	.080	.114	.108	.084	.016	.080	.108

The Berkowitz tests are rejection frequencies of the nulls of independence and iid(0,1), applied to the inverse Gaussian CDF transformed z_t sequence, for the forecast densities derived assuming gaussianity. That is, they are applied to the scaled forecast errors (see text). The last two cols. test the forecast errors for normality directly.

Table 6 Models of US GNP.

Linear AR models	1948 – 73	1948 – 90
α_1	0.3682	0.3449
α_2		0.1782
α_3		-0.1474
μ	0.9159	0.7910
σ	1.0432	0.9921
AIC	0.1440	0.0429

SETAR models	1948 – 73		1948 – 90	
	Lower	Upper	Lower	Upper
i	α_i	$\alpha_i + \beta_i$	α_i	$\alpha_i + \beta_i$
0	0.7187	-0.1089	-0.4693	0.4268
1	0.3496	0.7289	0.3936	0.3158
2			-0.8520	0.1718
σ	1.1758	0.8906	1.2684	0.8758
r	0.7161		-0.0580	
d	1		2	
AIC	0.0696		-0.0448	
n_L	42		35	
n_U	62		137	

Table 7 Hansen Test p-values for US GNP.

	Robust LM Statistics		
	SupT_T	ExpT_T	AveT_T
1948 – 73	0.796	0.516	0.378
1848 – 90	0.640	0.514	0.452
Simulated data			
	0.185	0.184	0.239

Table 8 Empirical US GNP forecast evaluation.

Point forecasts					
Equal MSFE	% better	Forecast Enc.		Forecast Fail.	
		AR	SETAR	AR	SETAR
.417	.831	.136	.087	.258	.201

Interval forecasts: 75% coverage intervals											
AR model						SETAR model					
Gaussian interval			Bootstrapped interval			Gaussian interval			Bootstrapped interval		
Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.
.483	.352	.507	.676	.284	.516	.890	.969	.990	.483	.432	.574

Density forecasts									
Kolmogorov-Smirnov 5% level				Berkowitz 5% level				Normality 5% level	
AR		SETAR		AR		SETAR		AR	SETAR
Gauss.	Bootst.	Gauss.	Bootst.	Indep.	Joint	Indep.	Joint		
.110	.113	.123	.116	.649	.586	.472	.329	.050	.019

The table records the forecast performance of AR and SETAR models of the Potter (1995) 1947 – 90 US GNP data, where $T - T_0 = 70$. All elements are p-values except for the KS statistics, which has a 10% critical value of 0.146, so no element is significant.

Table 9 Simulated US GNP data forecast evaluation.

Equal MSFE			% better			Forecast Encomp.		Forecast Failure	
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
.168	.006	.270	.064	.026	.090	.828	.198	.050	.052

Interval forecasts: 75% coverage intervals											
AR model						SETAR model					
Gaussian interval			Bootstrapped interval			Gaussian interval			Bootstrapped interval		
Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.	Uncond.	Ind.	Cond.
.068	.068	.074	.054	.054	.064	.068	.036	.062	.086	.024	.058

Density forecasts									
Kolmogorov-Smirnov 5% level				Berkowitz 5% level				Normality 5% level	
AR		SETAR		AR		SETAR		AR	SETAR
Gauss.	Bootst.	Gauss.	Bootst.	Indep.	Joint	Indep.	Joint		
.032	.032	.032	.098	.108	.102	.078	.108	.082	.076

The table records the rejection frequencies for the forecast performance of AR and SETAR models on simulated US GNP data, $T = 350$, $T_0 = 150$. An explanation of the information shown in the first panel of the table follows table 3.