Journal of Forecasting, Vol. 9, 25–36 (1990)

A Comparison of Tests for SETAR-type Non-linearity in Time Series

JOSEPH D. PETRUCCELLI
Worcester Polytechnic Institute, U.S.A.

ABSTRACT

Tests for SETAR-type non-linearity in time series have recently been proposed by Petruccelli and Davies (1986), W. S. Chan and Tong (1986), Tsay (1987), Luukkonen *et al.* (1988), Petruccelli (1987) and Moeanaddin and Tong (1988). In this paper we consider the relative performance of these tests.

KEY WORDS Non-linear time series SETAR-type non-linearity CUSUMS Lagrange-multiplier tests Likelihood ratio tests

INTRODUCTION

As more applications of non-linear time series models are found in areas in which forecasting plays a major role the importance of non-linear model building methodology to forecasters will continue to grow. Examples of such applications in hydrology, ecology and astronomy are found in Tong (1983) and in economics in Wecker (1977) and Maravall (1983).

Some of the structural features of non-linear models, such as limit cycles, which cannot be modeled by linear processes, can have a significant impact on the performance of forecasts. Jones (1976, 1978), Tong (1983), Pemberton (1987), Tong and Moeanaddin (1987) and Davies *et al.* (1988), among others, discuss multi-step forecasting using non-linear models.

As the use of non-linear models grows, so too will the need for tests to determine whether a given series is non-linear. In the past several years a number of such tests have been proposed by researchers. In the frequency domain, tests for general non-linearity have been outlined by Subba Rao and Gabr (1980) and Hinich (1982). In the time domain Keenan (1985) and Tsay (1986) have proposed tests based on the Volterra series representation of a time series. McLeod and Li (1983) have suggested a diagnostic portmanteau test statistic based on squared residuals from a linear fit. Lagrange multiplier tests for several specific classes of models have been considered by Saikkonen (1986), Saikkonen and Luukkonen (1986) and Luukkonen *et al.* (1987). W. S. Chan and Tong (1986) give a detailed discussion of several of these tests.

In this paper we investigate five tests designed to detect self-exciting threshold autoregressive (SETAR)-type non-linearity. In the next section we describe the five tests and discuss what we perceive to be the advantages and disadvantages of each. In the third section we compare their

0277–6693/90/010025–12$06.00

Received August 1988
Revised November 1988

performance via simulation on a large number of SETAR models. Our conclusions are given in the final section.

TESTS FOR SETAR-TYPE NON-LINEARITY

We assume the SETAR model of order p with delay d and thresholds $-\infty = r_0 < r_1 < \cdots < r_l = \infty$:

$$Y_t = \varphi_0^{(j)} + \sum_{i=1}^{p} \varphi_i^{(j)} Y_{t-i} + a_t, \; r_{j-1} < Y_{t-d} \leqslant r_j \tag{1}$$

where $t = \max\{d+1, p+1\}, \ldots, n$. Here the a_t are assumed i.i.d. with mean 0 and variance σ^2. Tong (1983) calls this a SETAR $(l; p, p, \ldots, p)$ model.

Let $h = \max\{1, p+1-d\}$ and let (i) denote the index of the ith smallest of $y_h, \ldots, y_{n-d}$. If we assume that equation (1) is linear, so that $\varphi_i^{(1)} = \varphi_i^{(2)} = \cdots = \varphi_i^{(l)}, i = 0, \ldots, p$, then we may write an ordered autogression for this equation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\varphi} + \mathbf{a} \tag{2}$$

where $\mathbf{Y}$ is a column vector containing $Y_{(i)+d}, i = 1, \ldots, n-d-h+1$, $\mathbf{X}$ is an $(n-d-h+1) \times (p+1)$ matrix whose first column is a vector of 1's and whose remaining columns contain appropriate lagged $Y_{(i)+d}$ values, $\boldsymbol{\varphi}$ is a column vector with entries $\varphi_0^{(1)}, \varphi_1^{(1)}, \ldots, \varphi_p^{(1)}$, and $\mathbf{a}$ is a column vector of noise terms. This ordered autoregression effectively divides equation (1) into the respective autoregressions in the l regions defined by the thresholds (see Petruccelli and Davies, 1986).

Test based on one-step-ahead forecast errors

Two of the tests we consider are based on successive one-step-ahead forecast errors from equation (2). Specifically, for each test:

(1) We choose the order p, delay d and a minimum number of startup observations $r_{\min} > p+1$.
(2) We then regress the first r rows of $\mathbf{Y}$ on the first r rows of $\mathbf{X}$ and compute z_{r+1}, the one-step-ahead standardized forecast error, successively for $r = r_{\min}, r_{\min}+1, \ldots, n-d-h$. This may be done very efficiently using regression updating methods (Brown *et al*, 1975).

The two tests differ in the use they make of the $\{z_{r+1}\}$. Petruccelli and Davies (1986) form the CUSUMS

$$Z_r = \sum_{i=r_{\min}+1}^{r} z_i, \qquad r = r_{\min}+1, \ldots, n-d-h+1$$

and from them their P-statistic,

$$P = \max_{r_{\min}+1 \leqslant r \leqslant n^* + r_{\min}} |Z_r| / \sqrt{n^*}$$

where $n^* = n - d - h + 1 - r_{\min}$. The asymptotic distribution of P is known if $\{Z_r\}$ is a random walk, which it will be approximately if the null hypothesis for linearity in equation (1) holds:

$$H_0: \varphi_i^{(1)} = \ldots = \varphi_i^{(l)} \qquad i = 0, 1, \ldots, p \tag{3}$$

However, we have recently found a more powerful test based on these CUSUMS, and it is this test that we will consider in this paper. The test is based on the idea that whereas the CUSUMS behave as a random walk under equation (3), if the observed process is SETAR, then it is the later rather than the earlier standardized forecast errors which should be biased. Thus we look at the reverse CUSUMS:

$$W_r = \sum_{i=1}^{r} z_{n^*+r_{\min}+1-i} \qquad r = 1, \ldots, n^*$$

In order to provide greater sensitivity to deviations in W_r for small r, we use boundaries of the form $ar + b$ and reject H_0 if $|W_r| > ar + b$ for some $1 \leqslant r \leqslant n^*$. Using a Brownian motion approximation (see e.g. Breiman, 1968, p. 289) and assuming the probability that the W_r cross both the lines $ar + b$ and $-ar - b$ is negligible, we have

$$P(\sup_r |W_r|/(ar + b) > 1) \approx 2P(\sup_r W_r/(ar + b) > 1) \approx 2e^{-2ab}$$

After a small amount of experimentation, we have found that for a level α test, choosing $b = \{\surd[-(n^*/2)\ln(\alpha/2)]\}/2$ and $a = 2\surd[-\ln(\alpha/2)/2n^*]$ gives satisfactory results and these values are used below. Undoubtedly, these choices can be improved upon. On virtually all simulated data we looked at, this text improved on the P-test, sometimes significantly so.

Remarks: We note that:

(1) The Petruccelli–Davies (1986) P-statistic uses boundaries of the above form for $a \equiv 0$.
(2) Brown *et al.* (1975) gave a more complicated approximation than we have used for computing a and b. However, in our experience, their approximation gives no better results than does the one used here.

The second test based on one-step-ahead forecast errors is that of Tsay (1987). He suggests performing the regression

$$\mathbf{z} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{z}$ is the vector of standardized one-step-ahead forecast errors with entries $z_i, i = r_{\min} + 1, \ldots, n^* + r_{\min}$, and $\tilde{\mathbf{X}}$ is the matrix $\mathbf{X}$ of equation (2) with the first $r_{\min}$ rows deleted, and computing the usual test statistic $\hat{F}_1$ for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$. He shows that under hypothesis (3), F_1 is asymptotically distributed as F_{ν_1,ν_2}, where $\nu_1 = p + 1, \nu_2 = n^* - p - 1$.

Remark: We have considered a simplification of Tsay's test, in which $\mathbf{z}$ is regressed on time rather than past values of the time series. That is, we perform the simple linear regression

$$\mathbf{z} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\delta}$$

where

$$\mathbf{W}' = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & n^* \end{bmatrix}$$

The test statistic $\hat{F}_2$ is the usual statistic for testing $H_0 : \boldsymbol{\gamma} = \mathbf{0}$. We can show that, under hypothesis (3), $\hat{F}_2$ is asymptotically distributed as $F_{2,\eta}$, where $\eta = n^* - 2$. This test has the advantage of simplicity and of being easier to interpret than the more complicated Tsay test. However, as it had virtually the same performance in simulations as did the Tsay test, we will not consider it further here.

Likelihood ratio-based tests

Suppose, for the moment, that $l = 2$ and that r_1, p and d are known in equation (1). Then, assuming normal errors, the likelihood ratio test statistic for equation (3) versus

$$H_1 : \varphi_i^{(1)} \neq \varphi_i^{(2)}, \qquad \text{some } 0 \leqslant i \leqslant p$$

is $\lambda = (\hat{\sigma}^2(NL; r_1)/\hat{\sigma}^2(L))^{(n-p+1)/2}$, where $\hat{\sigma}^2(L)$ and $\hat{\sigma}^2(NL; r_1)$ are, respectively, the residual variances from an AR(p) and SETAR (2; p, p) fit to the data, the latter with the known d and r_1 (see Petruccelli and Woolford, 1984; K. S. Chan and Tong, 1986). Under H_0, $-2\ln(\lambda)$ converges in distribution to a χ^2_{p+1}.

If the threshold r_1 is unknown, the likelihood ratio test statistic is

$$\lambda' = (\hat{\sigma}^2(NL; \hat{r}_1)/\hat{\sigma}^2(L))^{(n-p+1)/2},$$

where $\hat{r}_1$ is a least squares estimate of r_1. Until recently the asymptotic distribution of λ' was unknown.

For this reason, K. S. Chan and Tong (1986), who mention λ' as a possible test statistic, did not pursue its use further. Rather, W. S. Chan and Tong (1986) developed the following combined SETAR fitting/testing procedure:

(1) A 'best' (according to some criterion—Chan and Tong used BIC) AR model of order $1 \leqslant p \leqslant p_{\max}$ is obtained. $\hat{\sigma}^2(L)$ is computed from this fit.
(2) By searching over a grid of values of $1 \leqslant d \leqslant d_{\max}$, $1 \leqslant p \leqslant p_{\max}$, $-\infty = r_0 < r_1 < \cdots < r_l = \infty$ for $2 \leqslant l \leqslant l_{\max}$, a 'best' SETAR model is selected by the same criterion as in (1). $\hat{\sigma}^2(NL)$ is computed from this fit.
(3) The test statistic $\lambda'' = (\hat{\sigma}^2(NL)/\hat{\sigma}^2(L))^{(n-p+1)/2}$ is computed. Call it λ''_{obs}.
(4) An empirical significance level for λ''_{obs} is obtained by simulating realizations from the fitted AR model obtained in (1) and comparing λ'' for these realizations with λ''_{obs}.

In what follows, we consider the test λ'' as described above. Petruccelli (1987) implemented the test λ', obtaining empirical significance levels as in (4) above by simulating realizations from the AR(p) fit to data. This implementation was feasible (though computation time was still large) only through the use of an efficient regression updating algorithm.

Recently K. S. Chan and Tong (1988) solved this difficulty by obtaining the asymptotic distribution of a function of λ' under H_0 as the first passage probability of a $(p+1)$-dimensional Brownian bridge. In what follows we use critical regions based on this result. The regions are determined from the exact distribution for SETAR (2; 0, 0) processes and SETAR (2; 1, 1) processes with zero intercept and are empirically determined for all other SETAR processes, as described in Moeanaddin and Tong (1988).

A Lagrange multiplier test

Luukkonen *et al.* (1988) have devised an approximate Lagrange multiplier test for linearity versus smooth threshold autoregressive (STAR) alternatives, which are essentially, as the name implies, SETAR models whose autoregressive functions have a smooth transition between regimes. (See K. S. Chan and Tong, 1986b, for more on STAR models.)

Luukkonen *et al.* offer three versions of their test but all have a similar form:

(1) Regress Y_t on $\{1, Y_{t-j}, j = 1, \ldots, p\}, t = p+1, \ldots, n$, form the residuals $\{\hat{a}_t\}$ and the residual sum of squares SSE_0.
(2) Regress $\hat{a}_t$ on 1 and certain powers and cross-products of Y_{t-j}, $j = 1, \ldots, p$; obtain the residual sum of square SSE_1.

(3) Form the test statistic $S = (n - p)(1 - SSE_1/SSE_0)$.

The three versions of their statistic and the powers and cross-products in (2) which give them are

$$S_1: \{ Y_{t-j}, Y_{t-i}Y_{t-j}; \quad i, j = 1, \ldots, p \}$$
$$S_2: \{ Y_{t-j}, Y_{t-i}Y_{t-j}, Y_{t-i}Y_{t-j}^k; \quad i, j = 1, \ldots, p, k = 2, 3 \}$$
$$S_3: \{ Y_{t-j}, Y_{t-i}Y_{t-j}, Y_{t-j}^3; \quad i, j = 1, \ldots, p \}.$$

Under the null hypothesis that the process is AR(p),

$$S_1 \sim \chi^2_{p(p+1)/2}, S_2 \sim \chi^2_{p^2 + p(p+1)/2}, S_3 \sim \chi^2_{p+p(p+1)/2}$$

We note that:

(1) S_1 is the Tsay (1986) test statistic. (Note: this is not the Tsay, 1987, statistic considered below.)
(2) S_3 is a simpler version of S_2 designed to cut down on the large number of independent variables in the regression which yields SSE_1 for S_2.
(3) The tests do not assume the delay, d, is known. If it is, the second regressions can be done with far fewer independent variables in all tests. (See Luukkonen *et al.*, 1988, for details.)

Strengths and weaknesses of the tests
For convenience let us label the tests as:

LM, the Lagrange multiplier test;
L, the likelihood ratio test based on λ';
CT, the Chan and Tong test based on λ'';
T, the Tsay test; and
C, the reverse CUSUM test.

We summarize what we see as the strengths and weaknesses of each test in Table 1.

Some comments on the entries in Table 1 are in order. First CT differs from the other tests in that, in addition to varying the thresholds, it selects an 'optimal' SETAR model with respect to parameters p and d over a specified range. While LM requires a fixed value of p, it can be run in one version with d unspecified. While L, C and T all require p and d to be fixed, repeated applications of these tests for varied p and d can identify p and d values for model fitting.

Finally, LM is different from the other four tests in that, whereas they all explicitly incorporate the form of the SETAR model, LM seems to be testing for the presence of certain polynomial terms in the autoregressive function. Because of this, there does not appear to be clear and direct relation between the test and a particular SETAR fit to the data as there is with the other four tests.

With regard to the speed of the procedures, CT is very slow, primarily because of the use of simulations to obtain empirical significance levels, while the other four tests are very fast and of comparable speed. As an example, to process a series of length 100, CT averaged nearly one minute of CPU time compared to less than one half-second for the other four procedures! We have not tried CT or L for more than one threshold, but we would expect run times to increase exponentially for both tests with the number of thresholds tried.

Table 1. Strengths and weaknesses of the tests

Test	Strengths	Weaknesses
LM	(1) Can be used with either known or unknown d	(1) Size of second regression can grow quickly (2) Does not identify threshold, or delay (3) Must fix p
L	(1) 'Optimal' model selected is truly optimal in least squares sense	(1) Must fix p, d (2) Must prespectify a number of thresholds (3) Not a pre-test, as SETARs must be fitted
CT	(1) Gives 'optimal' model over a range of r, p and d values	(1) Very slow (2) 'Optimal' model has no known optimality properties (3) 'Optimal' model is only as good as the search procedure (4) Not a pre-test
C	(1) Does not require full model identification and fit (2) CUSUMS can be used to identify numbers and locations of thresholds	(1) Must fix p, d
T	As for C	As for C

PERFORMANCE OF TESTS ON SIMULATED DATA

We ran a simulation experiment to assess the performance of the five tests. In all cases the version of the test statistic S_3 with the correct known value of d was used for the LM procedure. This was done to keep the assumptions for its use in line with those for the L, C and T procedures. Its performance was comparable to that of S_2, which further justified the use of this version. The CT procedure was allowed to search over the range of parameter values $1 \leqslant d \leqslant p \leqslant 3$ and $1 \leqslant l \leqslant 2$.

A number of SETAR processes were simulated. For each process and test considered, 100 simulated series of length 100 were generated, using $N(0, 1)$ noise terms (except where noted otherwise) created by the IMSL routing GGNML. The slowness of CT made it necessary to use only 100 simulations. For each series startup values were set to 0 and the first 500 observations discarded. All simulations were performed on the DEC 20-60 computer at WPI.

We used $r_{\min} = p + 2$ in the T and C tests. Even though Tsay recommends $r_{\min} = n/10 + p$, we found our choice gave better results for both tests.

Empirical significance levels

To assess empirical significance levels, the tests were performed on the AR(1) processes

$$Y_t = \varphi Y_{t-1} + a_t$$

for $\varphi = \pm 0.9$, ± 0.5, ± 0.1 and 0. Table 2 shows the results for the 0.05 significance level. It can be seen that test C produces some high values and that CT tends toward low ones, but all empirical significance levels lie within the two standard error boundaries. This same pattern

Table 2. Empirical significance levels at nominal 5% level; 100 observations. Two standard error boundaries: (0.006, 0.094)

AR parameter	Significance				
φ	C	L	T	CT	LM
−0.9	0.09	0.04	0.06	0.05	0.04
−0.5	0.09	0.03	0.06	0.04	0.04
−0.1	0.05	0.06	0.03	0.03	0.07
0	0.03	0.07	0.04	0.01	0.07
0.1	0.09	0.04	0.06	0.05	0.05
0.5	0.04	0.09	0.04	0.04	0.04
0.9	0.05	0.04	0.03	0.03	0.05

shows itself in empirical significance results at the 0.01 and 0.10 nominal significance levels as well, and for sample size 50.

Power of the tests versus SETAR alternatives

In what follows we present the results of power studies conducted on simulated realizations of SETAR models. Since the number of realizations per simulation was relatively small, we also conducted an LSD (least significant difference)-type procedure to provide some guidance as to the true magnitude of differences in empirical power levels. The power levels for each model are ordered from highest to lowest and those groups having non-significant (at the 0.05 level) differences are underlined. All tests were conducted at the 0.05 level of significance.

Table 3 presents power results for data simulated from equation (1) with $p = 1, d = 1, l = 2, r_1 = 0$ and $\varphi_0^{(1)} = \varphi_0^{(2)} = 0$; that is, a SETAR (2; 1, 1) with zero threshold and intercepts. Although the autoregressive functions for these processes are non-linear, they are still continuous. The models in Table 4 have discontinuous autoregressive functions and threshold at the origin.

Since tests L and CT both fit single threshold models it is of interest to consider their performance on models with more than one threshold. Table 5 presents results for two such models. Others are found in Tables 6 and 7 as described below.

In addition to the above, we ran the tests on data simulated from three models from the literature that were fitted to the Canadian lynx data (see Tong, 1983). These models are:

Tong 1 (Tong, 1983, p. 102):

$$\begin{aligned} Y_t &= 0.62 + 1.25Y_{t-1} - 0.43Y_{t-2}, && Y_{t-2} \leqslant 3.25 \\ &= 2.25 + 1.52Y_{t-1} - 1.24Y_{t-2}, && Y_{t-2} > 3.25 \end{aligned}$$

with residual variances 0.0381 and 0.0626, respectively.

Tong 2 (Tong, 1983, p. 115):

$$\begin{aligned} Y_t &= 0.733 + 1.047Y_{t-1} - 0.007Y_{t-2} - 0.242Y_{t-3}, && Y_{t-2} \leqslant 3.083 \\ &= 1.983 + 1.52Y_{t-1} - 1.162Y_{t-2}, && Y_{t-2} > 3.083 \end{aligned}$$

with residual variances 0.0357 and 0.0586, respectively.

Table 3. Power of tests from equation (1): $p = d = 1$, $l = 2$, $r_1 = 0$, $\varphi_0^{(1)} = \varphi_0^{(2)} = 0$; 5% level; 100 observations

$\varphi_1^{(1)}$	$\varphi_1^{(2)}$	Power				
0.9	−0.1	C 0.60	LM 0.59	L 0.55	CT 0.26	T 0.20
0.9	−0.77	L 0.87	LM 0.87	C 0.79	CT 0.48	T 0.36
0.5	−0.5	LM 0.59	C 0.59	L 0.54	T 0.45	CT 0.21
0.5	−1	C 0.95	LM 0.93	L 0.93	T 0.80	CT 0.29
0	−1	LM 0.80	C 0.74	L 0.65	T 0.58	CT 0.38
−0.5	−1	LM 0.52	C 0.45	L 0.31	T 0.26	CT 0.21
1	−0.5	LM 0.67	C 0.48	L 0.36	T 0.32	CT 0.17
−1	0	C 0.81	T 0.71	L 0.64	LM 0.62	CT 0.42
−1	0.5	C 0.93	LM 0.91	L 0.90	T 0.90	CT 0.23
−0.5	0.5	T 0.67	C 0.66	L 0.61	LM 0.46	CT 0.26
−0.77	0.9	T 0.84	L 0.84	C 0.78	LM 0.72	CT 0.55
−0.1	0.9	LM 0.61	T 0.61	C 0.55	L 0.48	CT 0.28

Table 4. Power of tests from equation (1): $p = d = 1$, $l = 2$, $r_1 = 0$; 5% level; 100 observations

$\varphi_0^{(1)}$	$\varphi_1^{(1)}$	$\varphi_0^{(2)}$	$\varphi_1^{(2)}$	Power				
1	0.5	−1	1	L 0.99	LM 0.91	C 0.89	T 0.50	CT 0.44
1	−0.5	−1	1	C 1.0	L 1.0	T 1.0	LM 1.0	CT 0.35
−1	0.5	1	−1	L 0.50	LM 0.40	CT 0.14	T 0.11	C 0.08
1	0.5	−1	−1	CT 1.0	LM 1.0	T 1.0	L 1.0	C 1.0
−1	−0.5	1	−1	L 1.0	LM 0.78	C 0.61	CT 0.31	T 0.18
1	−0.5	−1	−1	LM 1.0	L 1.0	C 0.84	CT 0.81	T 0.69
1	0	−1	−0.5	L 1.0	LM 0.96	C 0.96	CT 0.88	T 0.75
1	0.5	1	−1.5	L 1.0	C 1.0	T 1.0	CT 1.0	LM 1.0
1	0.5	−1	−1.5	L 1.0	C 1.0	T 1.0	CT 1.0	LM 1.0

Table 5. Power of tests from equation (1); $p = d = 1$, $l = 3$, $r_1 = -1$, $r_2 = 1$; 5% level; 100 observations

$\varphi_0^{(1)}$	$\varphi_1^{(1)}$	$\varphi_0^{(2)}$	$\varphi_1^{(2)}$	$\varphi_0^{(3)}$	$\varphi_1^{(3)}$	Power				
						L	C	CT	LM	T
0	0.5	0	−0.5	0	0.5	0.61	0.37	0.36	0.32	0.10
						LM	L	C	CT	T
−1	0	0	1	1	0	0.53	0.46	0.32	0.27	0.24

Table 6. Power of tests on three data models for Canadian lynx data; 5% level; 100 observations

Model	Power				
	C	L	LM	T	CT
Tong 1	1.0	0.99	0.98	0.94	0.85
	L	C	LM	CT	T
Tong 2	1.0	1.0	0.95	0.94	0.94
	C	L	T	LM	
Tsay	1.0	1.0	0.96	0.94	

Table 7. Power of tests on three models for Canadian lynx data with equal variances; 5% level; 100 observations

Model	Variance	Power				
		LM	L	C	CT	T
Tong 1	0.05	1.0	0.99	0.97	0.89	0.80
		C	L	LM	T	CT
Tong 2	0.05	1.0	0.98	0.97	0.90	0.90
		C	L	LM	T	
Tsay	0.03	1.0	0.99	0.97	0.87	

Tsay (Tsay, 1987):

$$
\begin{aligned}
Y_t &= 0.083 + 1.096Y_{t-1}, && Y_{t-2} \leqslant 2.373 \\
&= 0.63 + 0.96Y_{t-1} - 0.11Y_{t-2} + 0.23Y_{t-3} - 0.61Y_{t-4} \\
&\quad + 0.48Y_{t-5} - 0.39Y_{t-6} + 0.28Y_{t-7}, && 2.373 < Y_{t-2} \leqslant 3.154 \\
&= 2.323 + 1.53Y_{t-1} - 1.266Y_{t-2}, && Y_{t-2} > 3.154
\end{aligned}
$$

with residual variances 0.015, 0.025 and 0.053, respectively. These models allow us to learn something of the effects of different variances on the performance of the tests. In addition, the TSAY model adds the effect of multiple thresholds. Tests, LM, L, C and T were run with the appropriate p and d values for each model. Test CT could not be run on the TSAY model due to the inordinate amount of CPU time required. The results are found in Table 6.

In order to gauge the effects due to unequal variances in the different regions we simulated these same three models with equal variances in all regions. The variances were chosen to be close to the average of the regional variances. Table 7 displays the results.

CONCLUSIONS

A number of conclusions can be drawn from the power studies described in the previous section. Our conclusions are based on differences in power significant at the 5% level.

First, not one of the tests considered performs best for all models considered. That having been said, there are clear differences apparent in the performance of the five tests.

For the SETAR (2; 1, 1) processes with zero intercept and threshold at the origin (Table 3) the C and LM statistics outperform the others. On the other hand, the CT test is clearly worst overall, while L seems to outperform T.

For the SETAR (2; 1,) models with non-zero intercepts and thresholds at the origin (Table 4), and for the two threshold models (Table 5), test L is clearly the best performer followed by LM and then by C. T is below these and CT worse still.

For data simulated from the three Canadian lynx models, C and L perform best in both the unequal and equal variance cases. CT is again worst. The only test affected by the unequal variances seems to be T, which performs better when the variances are unequal.

While no one test performs best for all series considered, a remarkable result of these simulations is the uniformly poor performance of CT. The reason for this performance has to be the search procedure for choosing the 'best' SETAR model from among a range of such models, for in all our simulations the search procedure had the option of duplicating the model chosen by L.

In order to overcome any bias in the results due to the CT search procedure choosing over a range of p and d values we re-ran a number of the simulation experiments on a version of CT that restricted its search to the correct p, d values. CT showed some improvement overall, but still performed no better than L and often much worse. Perhaps an improved search procedure would result in some real improvement in the performance of the CT test.

As regards C, L and LM, the best performers in our study, C and LM performed very well on the SETAR (2; ,1, 1) models with continuous autoregressive functions, while L dominated LM and LM dominated C when the autoregressive functions had a jump discontinuity. This suggests that C and LM may be more sensitive to a smooth change of regime while L may be more powerful when the change is abrupt. Perhaps there is some guidance here to the most profitable use of these tests.

Simulation results were obtained for sample size 50 for all models studied here. With the exception that C was clearly the best performer on the models of Table 3, the results were similar in general, though not in each particular, to those presented here. The similarity in performance extends to the same tests performed at the 1% and 10% levels.

Finally, we give some observations drawn from our experience in using these tests with simulated data not presented here and with real data.

(1) Given any of these tests, we can always find SETAR models on which it will perform poorly.

(2) The power of the tests quite clearly depends on the separation of the linear regimes within the SETAR model and on the signal-to-noise ratio of the data. For example, for the SETAR (2; 1, 1) models with $r_1 = 0$ and zero intercept (such as those in Table 3), the power depends in a clear way on $|\varphi_1^{(1)} - \varphi_1^{(2)}|$ for fixed σ^2, and on σ^2 for fixed $\varphi_1^{(1)}$, $\varphi_1^{(2)}$. For higher-order models the dependence is more complicated.

(3) For real data, which are often noisy, it is frequently difficult to detect true SETAR structure. Partly this is for the reasons described in (2), but just as often it is due to

outliers (with respect to a linear model). Simulation results indicate that a linear series with one or more added outliers is often detected to be non-linear. Of course, this may be reasonable in that the resulting data may certainly have non-linear features, yet unreasonable in that we would not wish to fit a non-linear model to the data. Similarly, SETAR series to which one or more outliers have been added are often detected as linear.

ACKNOWLEDGEMENT

The author would like to thank Professor Howell Tong for making available his and W. S. Chan's computer program for the computation of the CT statistic.

REFERENCES

Breiman, L., *Probability*, Reading, MA: Addison-Wesley, 1968.

Brown, R. L., Durbin, J. and Evans, J. M. 'Techniques for testing the constancy of regression relationships over time', *J. Royal Stat. Soc.*, Series B, **37**, (1975), 149–92.

Chan, K. S. and Tong, H., 'On a likelihood-ratio test for thresholds in time series', Technical Report, Mathematical Institute, University of Kent at Canterbury, UK, 1988.

Chan, K. S. and Tong, H., 'A survey of the statistical analysis of univariate threshold autoregressive models', to appear in Mariano, R. S. (ed.), *Advances in Statistical Computing: Theory and Applications*, JAI Press, USA, 1986a.

Chan, K. S. and Tong, H., 'On estimating thresholds in autoregressive models', *J. Time Ser. Anal.*, **7** (1986b), 179–90.

Chan, W. S. and Tong, H., 'On tests for non-linearity in time series analysis', *J. Forecasting*, **5** (1986) 217–28.

Davies, N., Pemberton, J. and Petruccelli, J., 'An automatic procedure for identification, estimation and forecasting univariate self exciting threshold autoregressive models', *The Statistician*, **37** (1988), 199–204.

Hinich, M. J., 'Testing for Gaussianity and linearity of a stationary time series', *J. Time Series Anal.*, **3** (1982) 169–76.

Jones, D. A., PhD, Thesis (unpublished), University of London, 1976.

Jones, D. A., 'Nonlinear autoregressive processes', *Proceedings of the Royal Society of London*, **A360** (1978), 71–95.

Keenan, D. M., 'A Tukey non-additivity-type test for time series nonlinearity', *Biometrika*, **72** (1985) 39–44.

Luukkonen, R., Saikkonen P. and Terasvirta, T., 'Testing linearity against smooth transition autoregressive models', *Biometrika*, **75** (1988), 491–500.

Luukkonen, R., Saikkonen P. and Terasvirta, T., 'Testing linearity in univariate time series models', unpublished manuscript 1987.

Maravall, A., 'An application of nonlinear time series forecasting', *J. Bus. Econ. Stat.*, **1**, (1983), 66–74.

McLeod, A. I. and Li, W. K., 'Diagnostic checking ARMA time series models using squared-residual autocorrelations', *J. Time Series Anal.*, (1983), 269–73.

Moeanaddin, R. and Tong, H., 'A comparison of likelihood ratio test and CUSUM test for threshold autoregression', Technical Report, Mathematical Institute, University of Kent at Canterbury, UK, 1988.

Pemberton, J., 'Exact least squares multi-step prediction from nonlinear autoregressive models', *J. Time Ser. Anal.*, **8** (1987), 443–8.

Petruccelli, J., 'On tests for SETAR-type nonlinearity in time series', Technical Report, WPI, Worcester, MA, 1987.

Petruccelli, J. D. and Davies, N.,'A Portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series', *Biometrika*, **73** (1986) 687–94.

Petruccelli, J. D. and Woolford, S. W., 'A threshold AR(1) model', *J. Appl. Prob.*, **21** (1984), 270–86.
Saikkonen, P., 'Properties of the Lagrange multiplier test under misspecification', Research Report No. 59, Department of Statistics, University of Helsinki, 1986.
Saikkonen, P. and Luukkonen, R., 'Lagrange multiplier tests for testing non-linearities in time series models', Research Report No. 57, Department of Statistics, University of Helsinki, 1986.
Subba Rao, T. and Gabr, M. M., 'A test for linearity of stationary time series', *J. Time Ser. Anal.*, **1** (1980), 145–52.
Tong, H., *Threshold Models in Non-linear Time Series Analysis*, Berlin: Springer-Verlag, 1983.
Tong, H. and Moeanaddin, R., On multi-step non-linear least squares prediction', Paper presented at the IoS conference, Cambridge, July 1987.
Tsay, R. S., 'Non-linearity tests for time series', *Biometrika*, **73** (1986). 461–6.
Tsay, R. S., 'Testing and modeling threshold autoregressive processes, Technical Report 406, Department of Statistics, Carnegie–Mellon University, to appear in *J. Amer. Statist. Assoc.* (1987).
Wecker, W. E., 'Asymmetric time series', *Amer. Statist. Assoc. Proc. Bus. Eco. Sec.*, (1977), 417–22.

Author's biography

Joseph D. Petruccelli is Associate Professor in the Mathematical Sciences Department at WPI, where he has taught since receiving his PhD in statistics from Purdue University in 1978. Since 1981 he has pursued research in time series analysis, where his current interests include non-linear models.

Author's address:
Joseph D. Petruccelli, Worcester Polytechnic Institute, Worcester, MA 01609, USA.