Econometrics Journal (1998), volume **1**, pp. C47–C75.

A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP

MICHAEL P. CLEMENTS[1], HANS-MARTIN KROLZIG[2]

[1]*Department of Economics, University of Warwick, UK*
E-mail: M.P.Clements@warwick.ac.uk
[2]*Institute of Economics and Statistics, University of Oxford, UK*
E-mail: Hans-Martin.Krolzig@nuffield.oxford.ac.uk

Received: October 1997, in final form January 1998

Summary While there has been a great deal of interest in the modelling of non-linearities in economic time series, there is no clear consensus regarding the forecasting abilities of non-linear time-series models. We evaluate the performance of two leading non-linear models in forecasting post-war US GNP, the self-exciting threshold autoregressive model and the Markov-switching autoregressive model. Two methods of analysis are employed: an empirical forecast accuracy comparison of the two models, and a Monte Carlo study. The latter allows us to control for factors that may otherwise undermine the performance of the non-linear models.

Keywords: *Business cycles, Monte Carlo simulation, Nonlinear time series, Prediction, Regime shifts.*

1. INTRODUCTION

In recent years there has been a great deal of interest in the modelling of non-linearities in economic time series. While the usefulness of linear time-series models in the tradition of Box and Jenkins (1970) is usually gauged by their predictive ability, there does not appear to be a clear consensus as to whether allowing for non-linearities has led to an improved forecast performance.[1] Much of this activity has focused on modelling either financial time series or output series such as GNP and industrial production.

In this paper we evaluate the forecast performance of two leading non-linear models that have been proposed for US GNP, the self-exciting threshold autoregressive (SETAR) model and the Markov-switching autoregressive (MS-AR) model. These two models have been used in contemporary empirical macroeconomics to characterize certain features of the business cycle, such as asymmetries between the expansionary and contractionary phases. The question

[1]See, for example, the review of non-linear time-series models by De Gooijer and Kumar (1992), especially p.151–2, and Diebold and Nason (1990) who in the context of exchange-rate prediction proffer four reasons why non-linear models may fail to forecast better than simple linear models, even when linearity is rejected statistically. Clements and Hendry (1996) and (1998) argue that a satisfactory in-sample fit is no guarantee of out-of-sample forecast performance even for linear models, a view echoed in a different context by Fildes and Makridakis (1995).

we address is whether in addition these models offer a much improved forecast performance. Our approach is univariate. The MS-AR framework can be readily extended to multivariate settings (see Krolzig (1997c) for an overview), but there are few applications of multivariate threshold autoregressive (TAR) models (but see, for example, Koop *et al.* (1996)).

Two methods of analysis are presented: an empirical forecast accuracy comparison of the two models, and a Monte Carlo study. For the empirical study, each model is formulated and estimated on a sub-sample of the historical data, and its forecasts of the observations held back at the model specification stage are then evaluated. The MS-AR model that we consider is based on the model of Hamilton (1989), which has been extensively discussed in the literature, and has been tested against linear autoregressive (AR) model) by for example Hansen (1992) and Hansen (1996a). The SETAR model of US GNP is similar to the models of Tiao and Tsay (1994) and Potter (1995) which have been formally tested by Hansen (1996b). The methods of model specification, estimation and forecasting for SETAR and MS-AR models are briefly described in Sections 3 and 4. For completeness, we begin in Section 2 by briefly reviewing model selection, estimation and forecasting for AR models.

To assess whether the results are unduly sensitive to a particular sample period and data vintage, we carry out the empirical forecast comparison on two data sets: the original (Potter 1995) data set, which contains real seasonally adjusted quarterly GNP measured at 1982 prices for the period 1948:1–1990:4, and on a recent vintage of data from 1959:1–1996:2 at 1992 prices. A longer series could be obtained by splicing the two series together, but we prefer to have the two separate sample realizations to provide an indication of the robustness of our findings.

It has often been argued (e.g. Granger and Teräsvirta (1993, Ch. 9) and Teräsvirta and Anderson (1992)) that the superior in-sample performance of non-linear models will only be matched out-of-sample if the 'non-linear features' also characterize the later period. Thus, we relate the discussion of the empirical forecast accuracy results to the business cycle characteristics realized over the forecast period (Section 5). For example, as we discuss in Section 3, a prominent non-linear feature of US GNP highlighted by the SETAR model is the robustness of the economy to negative shocks, so that once in recession the economy tends to return quickly to trend. Hence the SETAR model might be expected to perform well relative to linear models if the forecast period is characterized by a number of recessionary regimes. Some authors such as Tiao and Tsay (1994) then go on to evaluate forecast performance conditional on the state at the time the forecasts were made: for US GNP this entails evaluating forecasts made in the contractionary phase of the business cycle separately from those made during an expansion. Alternatively, since non-linear models typically contain many more parameters than linear models, a case can be made for requiring that the greater model complexity yields improved forecast performance 'on average' across states of nature, and this is the view we take in this paper.

The Monte Carlo study takes each of the estimated non-linear models in turn as the data generating process (DGP), to ensure that the non-linearities captured in the model on the past data do indeed persist into the future, and we assess the gains relative to a linear model, as well as the costs to using the 'wrong' non-linear model: that is, how much less accurate our forecasts would be if we used a SETAR model when the process generating the data is an MS-AR, and vice versa. Put bluntly, does the choice of non-linearity matter in this instance? In the Monte Carlo we also consider a small number of variants on the estimated models as the DGP, to assess the sensitivity of the costs/benefits to certain features of the design. Section 6 reports the Monte Carlo study, and Section 7 offers explanations of some of the key findings of the simulation study. Finally, Section 8 summarizes and concludes.

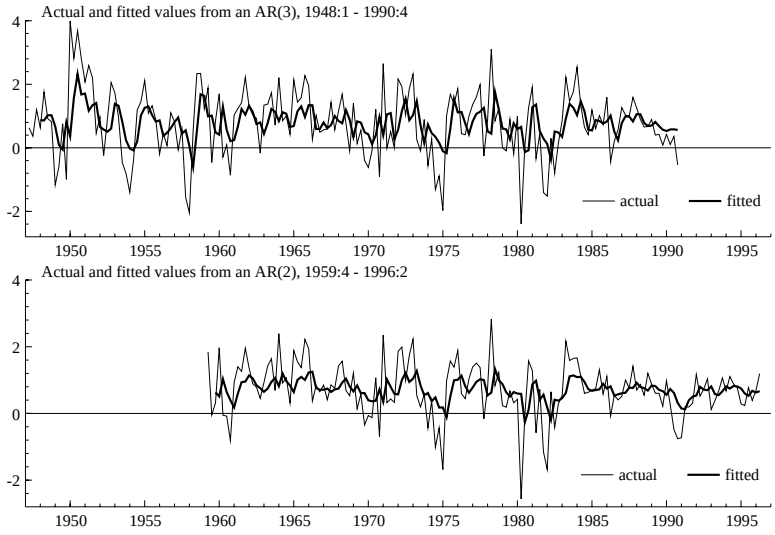

Figure 1. Full-sample AR model fitted values.

2. LINEAR AUTOREGRESSIVE MODELS

The linear model class we consider is in the Box and Jenkins (1970) time-series modelling tradition. Since the US GNP series appears to be integrated of order 1, we consider autoregressive-moving average (ARMA) models for Δy_t. We further restrict the model class to AR models. Without loss of generality, we report results for models in the mean-adjusted form:

$$\Delta y_t - \mu = \sum_{i=1}^{p} \alpha_i (\Delta y_{t-i} - \mu) + \epsilon_t. \tag{1}$$

The AR lag order, p, is selected to minimize the AIC (see, Akaike (1973)):

$$AIC(p) = \ln\{\hat{\sigma}^2(p)\} + 2(p+2)/T$$

where $\hat{\sigma}^2 = \hat{u}'\hat{u}/T$, and $\hat{u}$ denotes the vector of residuals.

Estimation and forecasting are straightforward. Once the parameters of the model have been estimated by OLS, the forecasts can be calculated recursively from:

$$\widehat{\Delta y}_{T+h|T} - \hat{\mu} = \hat{\alpha}_1(\widehat{\Delta y}_{T+h-1|T} - \hat{\mu}) + \cdots + \hat{\alpha}_p(\widehat{\Delta y}_{T+h-p|T} - \hat{\mu}) \tag{2}$$

where a '$\widehat{\ }$' denotes both a parameter estimate and a forecast value. T is the origin and h the forecast lead. The $\widehat{\Delta y}_{T+h-j|T}$, $0 \leq j \leq p$, are forecasts for $j < h$, and otherwise are known values. The forecast function (2) is the conditional expectation of Δy_{T+h} given information at period T and the form of the model (1).

Table 1 reports the results of estimating models of this form on the two vintages of US GNP data. Note that the AIC selects a different lag order for each of the four estimation periods. Figure 1 provides a graphical illustration of how well the selected linear models fit the data within sample.

Table 1. AR (p) models.

Sample	47:2–84:4	47:2–90:4	59:2–90:4	59:2–96:2
μ	0.7958	0.7775	0.7350	0.7106
α_1	0.3331	0.3472	0.3343	0.3039
α_2	0.1947	0.1780		0.1241
α_3	−0.1052	−0.1439		
α_4	−0.1198			
σ^2	1.0631	0.9635	0.6651	0.5767
AIC	0.1434	0.0217	−0.3586	−0.4950
Observations	146	170	122	144

3. SELF-EXCITING THRESHOLD AUTOREGRESSIVE MODELS

TAR models were first proposed by Tong (1978), Tong and Lim (1980) and Tong (1983) (see also Tong (1995a) for a detailed account). The idea is that the evolution of a process is governed by a series of distinct linear autoregressions, where the linear autoregression that generates the values of the time series at any instant depends upon the value taken by a 'threshold variable'. When the threshold variable is the value of the process in a previous period the process is described as being 'self-exciting', hence the acronym SETAR. Thus for a SETAR model of x_t the threshold variable would be a lag of x_t, say, x_{t-d}, where d is known as the length of the delay.

Formally, x_{t-d} is continuous on $\Re$, so that partitioning the real line defines the number of distinct regimes, say N_r, where the process is in the i-th regime when $r_{i-1} \leq x_{t-d} < r_i$, and in that case the p-th order linear autoregression is defined by,

$$x_t = \alpha_{0i} + \alpha_{1i} x_{t-1} + \cdots + \alpha_{pi} x_{t-p} + \epsilon_{ti}, \qquad \epsilon_{ti} \sim IID(0, \sigma_i^2), \qquad i = 1, 2, \ldots, N_r \quad (3)$$

where the parameters sub-scripted by i may vary across regimes. This model is sometimes written as a SETAR(N_r; $p, \ldots, p$). A lag order that varies over regimes can be accommodated within this framework by defining p as the maximum lag order across the regimes, and noting that some of the α_{ji} may be zero.

The model may be non-stationary within a regime, in the sense that some of the roots of

$$\alpha_i(L) = 1 - \alpha_{1i} - \cdots - \alpha_{pi}$$

may lie on the unit circle, but nevertheless stationary overall due to the alternation of 'explosive' and 'contractionary' regimes, generating limit cycle behaviour.

For modelling US GNP, x_t above is usually taken to be Δy_t, where Δy_t is 100 times the first difference of the log of real GNP.

3.1. Estimation

Conditional on the number of regimes, the regime r (assuming $N_r = 2$), and the delay, d, the sample can simply be split into two and an OLS regression can be run on the observations belonging to each regime separately, or indicator functions can be used in a single regression, constraining the residual error variance to be constant across regimes (see, for example, Potter, (1995, p. 113)).

In practice r is unknown, and the model is estimated by searching over r and d: r is allowed to take on each of the sample period values of x_{t-d} in turn,[2] and d typically takes on the values $0, 1, 2, \ldots$ up to the maximum lag length allowed. For a known lag order, the selected model is that for which the pair (r, d) minimize the overall residual sum of squares (RSS), equal to the sum of the RSS in each regime.

When p is unknown, fit is usually traded against parsimony. A search is made over all values of p less than some maximum, and the preferred order is often taken to be that which minimizes AIC.

In Sections 5 and 6 we set the maximum lag length at 5, require the lag orders to be the same across regimes, and do not allow 'holes' in the lag distributions. The selected model is the combination of p, d and r which minimizes

$$N_L \times \ln \hat{\sigma}_L^2 + N_U \times \ln \hat{\sigma}_U^2 + 2 \times (p + 1) + 2 \times (p + 1)$$

where $\hat{\sigma}_L$ and $\hat{\sigma}_U$ are the standard errors of the 'lower' $(_L)$ and 'upper' $(_U)$ regimes, and N_L and N_U are the number of observations in each.

3.2. Forecasting

Constructing multiperiod forecasts is considerably more difficult than for linear models, and exact analytical solutions are not available.[3] Exact numerical solutions require sequences of numerical integrations (see, for example, Tong (1995a, Sections 4.2.4 and 6.2)) based on the Chapman–Kolmogorov relation. Clements and Smith (1997) compare a number of alternative methods of obtaining multiperiod forecasts, including the normal forecast error (NFE) method suggested by Al-Qassam and Lane (1989) for the exponential AR model, and adapted by De Gooijer and De Bruin (1997) to forecasting SETAR models. They conclude that the Monte Carlo method performs reasonably well. In this paper SETAR forecasts are generated by Monte Carlo.[4]

3.3. SETAR models of US GNP

A number of authors have estimated SETAR models of US GNP. Tiao and Tsay (1994) consider a two-regime SETAR model and a four-regime refinement, where $p = 2$. Potter (1995)

[2]The range of values of x_{t-d} is restricted to those between the 15th and 85th percentile of the empirical distribution, following Andrews (1993) and Hansen (1996b).

[3]See Granger and Teräsvirta (1993) for an introductory account and a discussion of a number of methods of obtaining forecasts for a general non-linear model.

[4]Tiao and Tsay (1994) use this method for their SETAR model of US GNP.

estimates a SETAR(2; 5, 5) but with the third and fourth lags restricted to zero under both regimes, and $d = 2$ and $r = 0$. The model selection procedure in Potter (1995) is fairly *ad hoc*, but the selected model is similar to that of Tiao and Tsay (1994) except for the lag 5 terms. The estimation period in both instances is 1947–1990. A noteworthy feature of SETAR models of US GNP over this period is a large negative coefficient on the second lag in the lower regime, indicating that the US economy moves swiftly out of recession.

Tiao and Tsay (1994) find that the empirical performance of the SETAR model relative to a linear AR model is markedly improved when the comparison is made in terms of how well they forecast when the economy is in recession. The reason is easily understood. Since a clear majority of the sample data points (approximately 78%) fall in the upper regime, the linear AR(2) model will be largely determined by these points and will closely match the upper-regime SETAR model. Thus the forecast performance of the two models will be broadly similar when the economy is in the expansionary phase of the business cycle. To the extent that the data points in the lower regime are characterized by a different process, there should be gains to the SETAR model during the contractionary phase. Clements and Smith (1996) and (1997) find evidence for this effect in empirical and Monte Carlo analyses of the forecast performance of SETAR and linear models. If we do not evaluate forecasts conditional upon regimes, then the gains in the minority regime need to be sufficiently large for the SETAR to perform well on average.

Clements and Smith (1997) argue that the empirical forecasting exercise results of Tiao and Tsay (1994) may be misleading as an indication of the out-of-sample forecast performance of the SETAR model, since their specification utilizes information that would not have been known when the forecasts were made. However, rectifying this shortcoming still yields gains conditional on being in the lower regime.

To summarize: a reading of the literature suggests that SETAR model forecasts of US GNP are superior to forecasts from linear models, particularly when the forecasts are made during a recession (more precisely, when growth is negative). In Section 3.4 we review the evidence for the statistical significance of more than one regime in TAR models of US GNP.

As shown in Table 2 a SETAR(2; 2, 2) minimizes AIC for each of the four model estimation periods. The estimates for the sub- and full-sample periods are similar for a given data vintage, indicating parameter constancy, but differ markedly between vintages. The models for the more recent vintage indicate a threshold at a quarterly growth rate of 0.32%, so that there is a distinction between low growth and high growth rather than between absolute declines and increases in the level of GNP, as for the earlier vintage analysed by Tiao and Tsay (1994), Potter (1995) and Hansen (1996b). Figure 2 depicts the in-sample fit of the SETAR models for the two full-sample periods, which is similar to that of the preferred linear model (see Figure 1).

3.4. Testing for more than one regime: the SETAR model

Hansen (1996b) 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

Table 2. SETAR models.

Sample	47:2-84:4	47:2–90:4	59:2–90:4	59:2–96:2
Lower regime				
α_{0L}	−0.4996	−0.4693	0.2099	0.2528
α_{1L}	0.3976	0.3936	0.1374	0.1687
α_{2L}	−0.8676	−0.8520	−0.2345	−0.1482
σ_L	1.2844	1.2684	1.2393	1.1027
Upper regime				
α_{0U}	0.4573	0.4016	0.5530	0.5405
α_{1U}	0.3223	0.3160	0.3337	0.3338
α_{2U}	0.1541	0.1863	0.0225	0.0234
σ_U	0.9333	0.8775	0.6408	0.6126
Threshold	−0.0580	−0.0580	0.3189	0.3189
Delay	2	2	2	2
N_L	34	35	30	39
N_U	115	138	95	108
AIC	0.0882	−0.0429	−0.4773	−0.5867

is non-standard. Hansen (1996b) finds only weak evidence for rejecting the linear model in favour of the Potter (1995) SETAR model of US GNP.

The testing approach and an explanation of the test statistics are outlined in Appendix. Here we record the results of calculating those statistics for our two data samples (1947–90 and 1959–96) for two SETAR model specifications. The SETAR(2;5,5) is the fifth-order model of Potter (1995) with the third and fourth lags excluded. From Table 3, which records the p-values for the $\sup T_T$, $\text{ave}\, T_T$ and $\exp T_T$ statistics of the null of linearity, it is apparent that the fifth-order model appears to obtain more support from the data than the second-order model.

Nonetheless, the evidence for the SETAR model is weak – on any test and for either sample period the null of linearity is not rejected at the 5% level. The results for the fifth-order model and the earlier sample period are similar to those reported by Potter, (1995, Table IV, p. 115). However, Potter (1995, same table) also records Monte Carlo evidence indicating that the tests are too conservative, particularly the heteroscedasticity-robust versions, and that the power at the nominal 5% level is low. Correcting for size, he finds evidence in favour of non-linearity at the 10% level. The $\sup T_T$ and $\exp T_T$ tests of the fifth-order model on the later sample have p-values only just over 10%, so a size correction here might suggest a similar outcome.

Recent research by Diebold and Chen (1996) on the tests of structural change of Andrews (1993) and Andrews and Ploberger (1994) suggests that bootstrapping the distributions of the test statistics results in much smaller size distortions in some cases than using the asymptotic distributions. Given the similarities between the testing procedures for structural change and

Table 3. Asymptotic *p*-values of linear null versus SETAR model. The results were obtained using Bruce Hansen's Gauss code `tar.prg`.

SETAR model	(2; 2, 2)	(2; 5, 5)	(2; 2, 2)	(2; 5, 5)
	1947–90		1959–96	
	Robust LM statistics			
sup T_T	0.653	0.191	0.263	0.473
exp T_T	0.529	0.182	0.556	0.305
ave T_T	0.477	0.265	0.698	0.208
	Standard LM statistics			
sup T_T	0.094	0.054	0.860	0.125
exp T_T	0.183	0.100	0.860	0.113
ave T_T	0.322	0.278	0.855	0.176

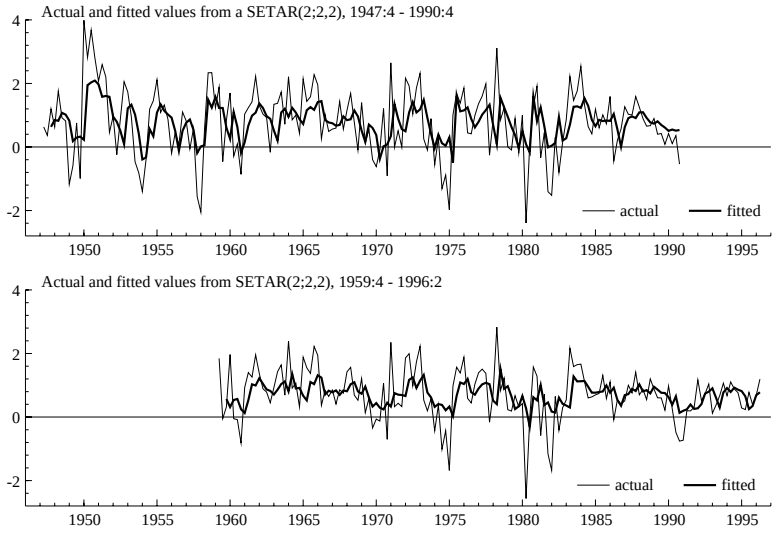

Figure 2. Full-sample SETAR model fitted values.

SETAR non-linearities, estimating the finite sample distribution via the bootstrap may also improve the size of tests in the current context, but that is beyond the scope of this paper.

4. MARKOV-SWITCHING AUTOREGRESSIVE MODELS

The Hamilton (1989) model of the US business cycle fostered a great deal of interest in the MS-AR model as an empirical vehicle for characterizing macroeconomic fluctuations, and there

have been a number of subsequent extensions and refinements.[5] Contractions and expansions are modelled as switching regimes of the stochastic process generating the growth rate of real GNP. The regimes are associated with different conditional distributions of the growth rate of real GNP, where, for example, the mean is positive in the first regime ('expansion') and negative in the second regime ('contraction').

The Hamilton (1989) model of the US business cycle fits a fourth-order autoregression ($p = 4$) to the quarterly percentage change in US real GNP from 1953 to 1984:

$$\Delta y_t - \mu(s_t) = \alpha_1\{\Delta y_{t-1} - \mu(s_{t-1})\} + \cdots + \alpha_4\{\Delta y_{t-p} - \mu(s_{t-4})\} + \epsilon_t, \tag{4}$$

where $\epsilon_t \sim NID(0, \sigma^2)$ and the conditional mean $\mu(s_t)$ switches between two states ($M = 2$):

$$\mu(s_t) = \begin{cases} \mu_1 > 0 & \text{if } s_t = 1 \text{ ('expansion' or 'boom'),} \\ \mu_2 < 0 & \text{if } s_t = 2 \text{ ('contraction' or 'recession').} \end{cases}$$

The variance of the disturbance term, σ^2, is assumed to be the same in both regimes.

The general idea behind the class of regime-switching models is that the parameters of an AR process depend upon an *unobservable* regime variable $s_t \in \{1, \ldots, M\}$, which represents the probability of being in a particular state of the world. In this case, only the mean switches, hence the model is referred to as a MS*M*ean(2)-AR(4), to denote two regimes ($M = 2$) and four lags ($p = 4$). In more general specifications the AR parameters and the variance could depend on the state s_t of the Markov chain. A major advantage of the MS-AR model is its flexibility in modelling time series subject to regime shifts.

A complete description of the MS-AR model requires the formulation of a mechanism that governs the evolution of the stochastic and unobservable regimes on which the parameters of the autoregression depend. Once a law has been specified for the states s_t, the evolution of regimes can be inferred from the data. In MS-AR models the regime-generating process is an ergodic Markov chain with a finite number of states $s_t = 1, \ldots, M$ (in (4), $M = 2$) defined by the transition probabilities

$$p_{ij} = \Pr(s_{t+1} = j | s_t = i), \quad \sum_{j=1}^{M} p_{ij} = 1 \qquad \forall i, j \in \{1, \ldots, M\}. \tag{5}$$

There is evidence that in some instances the assumption of fixed transition probabilities p_{ij} should be relaxed, and models with time-varying and duration-dependent transition probabilities have been considered (see, for example, Diebold *et al.* (1993), Diebold *et al.* (1994), Filardo (1994), Lahiri and Wang (1994), and Durland and McCurdy (1994)). The former are modelled as logistic functions (to bound the probabilities between 0 and 1) of economic variables. When applied to modelling US GNP, the latter indicate that the probability of transition out of recession is increasing in the duration of the recession (see Filardo (1994)). We do not consider these extensions here.

4.1. Estimation

The maximization of the likelihood function of an MS-AR model entails an iterative estimation technique to obtain estimates of the parameters of the autoregression and the transition prob-

[5] See *inter alia* Albert and Chib (1993), Diebold *et al.* (1994), Ghysels (1994), Goodwin (1993), Hamilton (1994), Kähler and Marnet (1994), Kim (1994), Krolzig and Lütkepohl (1995), Krolzig (1997c), Lam (1990), McCulloch and Tsay (1994), Phillips (1991) and Sensier (1996).

abilities governing the Markov chain of the unobserved states. Denote this parameter vector by λ, so that for the MSM(2)-AR(4) model $\lambda = (\mu_s, \alpha_1, \ldots, \alpha_4, \sigma^2, p_{11}, p_{22})$. λ is chosen to maximize the likelihood for given observations $Y_T = (y_T', \ldots, y_{1-p}')'$.

Maximum likelihood (ML) estimation of the model is based on an implementation of the expectation maximization (EM) algorithm proposed by Hamilton (1990) for this class of model – an overview on alternative numerical techniques for the ML estimation of AR(M)-MS(p) models is given in Krolzig (1997c). The EM algorithm introduced by Dempster *et al*. (1977) is designed for a general class of models where the observed time series depends on some unobservable stochastic variables – for MS-AR models these are the regime variables s_t. Each iteration of the EM algorithm consists of two steps. The *expectation* step involves a pass through the filtering and smoothing algorithms, using the estimated parameter vector $\lambda^{(j-1)}$ of the last maximization step in place of the unknown true parameter vector. This delivers an estimate of the smoothed probabilities $\Pr(S|Y, \lambda^{(j-1)})$ of the unobserved states s_t (where S records the history of the Markov chain). In the *maximization* step, an estimate of the parameter vector λ is derived as a solution $\tilde{\lambda}$ of the first-order conditions associated with the likelihood function, where the conditional regime probabilities $\Pr(S|Y, \lambda)$ are replaced with the smoothed probabilities $\Pr(S|Y, \lambda^{(j-1)})$ derived in the last expectation step. Equipped with the new parameter vector λ the filtered and smoothed probabilities are updated in the next expectation step, and so on, guaranteeing an increase in the value of the likelihood function at each step.

Regimes constructed in this way are an important instrument for interpreting business cycles using MS-AR models. They constitute an optimal inference on the latent state of the economy, whereby probabilities are assigned to the unobserved regimes 'expansion' and 'contraction' conditional on the available information set.

4.2. Forecasting

The derivation of an optimal predictor can often be quite complicated in empirical work for non-linear time-series models. As highlighted in Krolzig (1997a), an attractive feature of MS-AR models is the ease with which forecasts can be obtained.

For the mean-squared prediction error criterion, the optimal predictor $\widehat{\Delta y}_{t+h|t}$ is given by the conditional mean:

$$\widehat{\Delta y}_{t+h|t} = E(\Delta y_{t+h}|Y_t) \tag{6}$$

for a given information set Y_t. In contrast to linear models, the MSE optimal predictor $\widehat{\Delta y}_{t+h|t}$ does not usually have the property of being a linear predictor. For MS-AR models with regime-invariant AR parameters, however, the conditional mean can easily be derived analytically, unlike for many non-linear models (see the discussion of forecasting the SETAR model, Section 3).

In MSM(M)-AR(p) models with a regime-dependent mean $\mu(s_t)$, the conditional expectation $\widehat{\Delta y}_{T+h|T}$, is given by

$$E(\Delta y_{T+h}|Y_T) = \sum_{s_{T+h}=1}^{M} \cdots \sum_{s_{T+h-p}=1}^{M} E(\Delta y_{T+h}|Y_T, s_{T+h}, \ldots, s_{T+h-p}) \times \Pr(s_{T+h}, \ldots, s_{T+h-p}|Y_T),$$

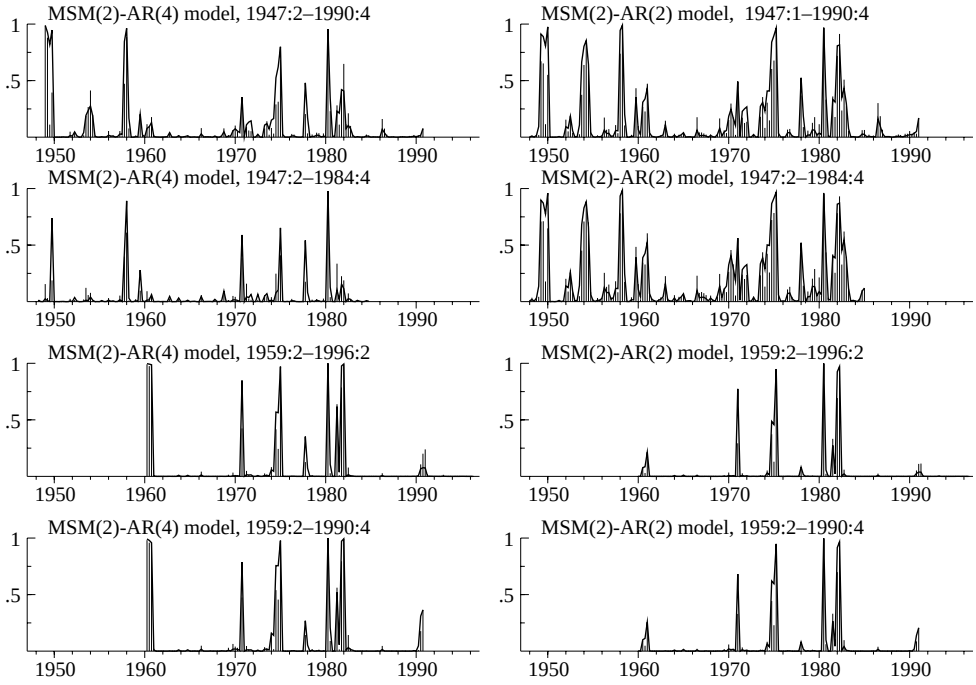

Figure 3. MSM(2)-AR models' smoothed and filtered probabilities of the lower regime, L.

where Y_T is the full-sample information.

For the particular case used by Hamilton – MSM(2)-AR(4) – this amounts to,

$$E(\Delta y_{T+h}|Y_T) = \sum_{s_{T+h}=1}^{2} \cdots \sum_{s_{T+h-4}=1}^{2} \left[\mu(s_{T+h}) + \sum_{k=1}^{4} \alpha_k \{\Delta y_{T+h-k} - \mu(s_{T+h-k})\} \right] \Pr(s_{T+h}, \ldots, s_{T+h-4}|Y_T) \tag{7}$$

which leads to the recursion,

$$\widehat{\Delta y}_{T+h|T} = \hat{\mu}_{T+h|T} + \sum_{k=1}^{4} \alpha_k (\widehat{\Delta y}_{T+h-k|T} - \hat{\mu}_{T+h-k|T}) \tag{8}$$

with initial values $\widehat{\Delta y}_{T+h|T} = \Delta y_{T+h}$ for $h \leq 0$ and where the predicted mean is given by,

$$\hat{\mu}_{T+h|T} = \sum_{j=1}^{2} \mu_j \Pr(s_{T+h} = j|Y_T).$$

The predicted regime probabilities:

$$\Pr(s_{T+h} = j|Y_T) = \sum_{i=1}^{2} \Pr(s_{T+h} = j|s_T = i) \Pr(s_T = i|Y_T)$$

only depend on the transition probabilities $\Pr(s_t = j|s_{t-1} = i) = p_{ij}$, $i, j = 1, 2$ and the filtered regime probabilities $\Pr(s_T = i|Y_T)$.

More generally, the optimal predictor of the MS-AR model is linear in the last p observations and the last regime inference, but there exists no purely linear representation of the optimal predictor in the information set. This is discussed in more detail in Krolzig (1997c, Ch. 4). Note, however, that the optimal forecasting rule becomes linear in the limit as the regimes become completely unpredictable, defined by, $\Pr(s_t|s_{t-1}) = \Pr(s_t)$ for $s_t, s_{t-1} = 1, 2$, since then $\hat{\mu}_{T+h} = \bar{\mu}$, the unconditional mean of y_t.

4.3. MS-AR models of US GNP

Hamilton's MSM(2)-AR(4) Model. Figure 3 presents the time paths of smoothed full-sample probabilities (line) and filtered probabilities (bars) for the contractionary regime (L) for the MSM(2)-AR(4) model, for the two data vintages we consider. These are the probabilities of being in a recession at time t.[6] Figure 3 along with Table 4 suggests that the statistical characterization of the US business cycle afforded by the MSM(2)-AR(4) model is inadequate. The MSM(2)-AR(4) model estimated over both our sample periods (1947–90 and 1959–96, and sub-samples thereof) does not exhibit business-cycle features. Our findings are in line with Hess and Iwata (1995), but in contrast to those of Hamilton (1989) (sample period: 1953–84) and Krolzig (1997b) (sample period: 1960–91), suggesting that the business-cycle interpretation of the model becomes rather laboured when the estimation period includes the end of World War II, the Korean War, and the most recent (nineties) data. The average duration of regime L is only a little over 1 for the sub-sample of the earlier data vintage, with a probability of staying in regime L (p_{LL}) of only 15%. Thus for this period, at least, the MSM(2)-AR(4) model often attributes single, isolated observations to regime L, and is more a 'model of outliers' than a business-cycle model. For the other three periods the duration of regime L is never much over two periods.

In line with our results with the SETAR model, we find that the contractionary regime L picks up the negative shocks in 1970, 1974/75, 1980–82. Nevertheless, the parameter estimates in Table 4 are reasonably constant between the sub-sample and full-sample periods for both data vintages. The structural stability of the MSM(2)-AR(4) model suggests its potential relevance for forecasting US GNP growth.

Table 6 reports the results of estimating the MSM(2)-AR(p) for $p = 1, \ldots, 7$, for the 1947–90 sample period. On the basis of minimizing AIC, the model with four lags appears over-parameterized, since $p = 2$ is optimal. Nevertheless, the two models offer a similar characterization of the business cycle, and little turns on whether the third and fourth lag terms are included; see Figure 3. An informal check on the stability of the MSM(2)-AR(2) model is provided by the 'closeness' of the estimates for the full and sub-sample periods recorded in Table 5. The small differences between the MSM(2)-AR(2) models in Tables 5 and 6 are due to different numbers of initial values (5 in the former, and 7 in the latter).

[6]The filtered regime probabilities $\Pr(s_t = m|Y_t) = \Pr(s_t = m|y_t, y_{t-1}, \ldots, y_0)$ are based on information up to time t, while the smoothed probabilities $\Pr(s_t = m|Y_T) = \Pr(s_t = m|y_T, \ldots, y_{t+1}, y_t, y_{t-1}, \ldots, y_0)$ are calculated from full-sample information, employing observations known only after period t. Each constitutes optimal inference on the state of nature given the information set.

Table 4. MSM(2)-AR(4) models.

Sample	47:2–84:4	47:2–90:4	59:2–90:4	59:2–96:2
Mean μ_L	−1.2708	−0.9919	−0.9991	−1.1335
Mean μ_H	0.8862	0.9607	0.8961	0.8388
α_1	0.3720	0.3210	0.2953	0.3445
α_2	0.2296	0.2629	0.1070	0.1240
α_3	−0.0966	−0.0458	−0.1005	−0.0762
α_4	−0.1927	−0.0975	0.0468	0.0299
σ^2	0.8313	0.7307	0.4182	0.3767
Trans.prob p_{LL}	0.0510	0.6163	0.5597	0.5135
Trans.prob p_{HH}	0.9589	0.9616	0.9551	0.9635
Uncond.prob.L	0.0415	0.0909	0.0925	0.0683
Uncond.prob.H	0.9585	0.9091	0.9075	0.9317
Duration L	1.05	2.61	2.27	2.06
Duration H	24.32	26.07	22.28	28.05
Observations	147	168	123	145
Loglikelihood	−210.27	−232.52	−142.69	−158.38

MSIH(3)-AR(4) Model. We found that an adequate 'business-cycle' model of US GNP (in the sense of generating regime durations consonant with estimates based on the NBER chronology, for example) required the introduction of a third regime and a regime-dependent error variance:

$$y_t = \mu(s_t) + \sum_{k=1}^{4} \alpha_k y_{t-k} + \epsilon_t, \tag{9}$$

where $\epsilon_t \sim NID(\sigma^2(s_t))$ and $s_t \in \{1, 2, 3\}$ is generated by a Markov chain. The specification has a shifting intercept term (MSI*ntercept*, rather than MSM*ean-adjusted*) and in the following will be denoted as an MSIH(3)-AR(4) model (where the H flags the heteroskedastic error term). The lag order is four.

Figure 4 and Table 7 summarize the business-cycle characteristics of this model. The figure depicts the filtered and smoothed probabilities of the 'high growth' regime H and the contractionary regime L (the middle regime M probabilities are not shown). The expansion and contraction episodes produced by the three-regime model correspond fairly closely to the NBER classifications of business-cycle turning points. In contrast to the two-regime model, all three regimes are reasonably persistent. We include this model in the empirical forecast comparison.

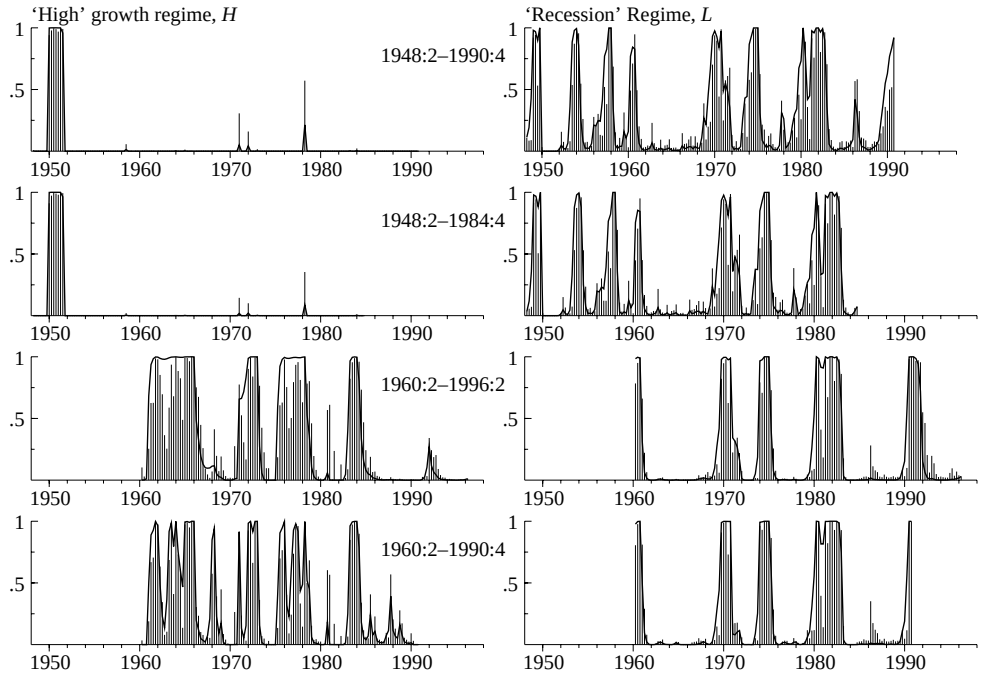

Figure 4. MSIH(3)-AR(4) model smoothed and filtered probabilities of the 'extreme' regimes, H, L.

4.4. Testing for more than 1 regime: the MS-AR model

Results of formal tests of the MS-AR model against AR models have been at best mixed. Hansen (1992) and (1996a) is unable to reject an AR(4) in favour of Hamilton's model (on the Hamilton data) using a standardized LR test designed to deliver (asymptotically) valid inference. Conventional testing approaches are not applicable due to the presence of unidentified nuisance parameters under the null of linearity (that is, the transition probabilities) and because the scores associated with parameters of interest under the alternative may be identically zero under the null. Since Hansen's approach delivers only a bound on the asymptotic distribution of the standardized LR test, the test may be conservative, tending to be under-sized in practice and of low power.[7]

Table 8 reports the p-values of the standardized LR test of a linear AR model against the MS-AR model for our two sample sizes, and, given the results on selecting the lag order of the model, for second-order models as well as the more popular fourth-order model. For comparison, the first two rows of the table record the tests of the Hamilton model for 1952–84, and for $p = 2$ as well as $p = 4$ (the latter approximately reproduces part of Hansen (1996a, Table III, p. 196)). In no case do we reject the null of 1-state at even the 20% level.

Hansen (1992) warns against using Monte Carlo based tests of the Hamilton model (as in for example, Lam (1990) and Cecchetti *et al.* (1990)) because the empirical null distribution

[7]Hansen argues that this is not in fact the case, based on Monte Carlo calculations of the finite sample size and power of the standardized LR test.

Table 5. MSM(2)-AR(2) models.

Sample	47:2–84:4	47:2–90:4	59:2–90:4	59:2–96:2
Mean μ_L	−0.5923	−0.9159	−1.1557	−1.2562
Mean μ_H	1.1310	1.002	0.8529	0.8100
α_1	0.2717	0.2943	0.2921	0.3302
α_2	0.2017	0.2228	0.0362	0.0519
σ^2	0.7402	0.72173	0.4704	0.4213
Trans.prob p_{LL}	0.6287	0.69325	0.3826	0.3302
Trans.prob p_{HH}	0.9117	0.9568	0.9603	0.9681
Uncond.prob.L	0.1921	0.1233	0.0604	0.0455
Uncond.prob.H	0.8079	0.8767	0.9396	0.9545
Duration L	2.69	3.26	1.62	1.49
Duration H	11.33	23.17	25.18	31.30
Observations	149	170	125	147
Loglikelihood	−215.55	−236.75	−147.94	−164.00

obtained by repeated fitting of the MS-AR model to an AR process is likely to be a lower bound for the true distribution, since the likelihood of the MS-AR model may frequently attain only a local maximum. Consequently inferences may be too liberal.

Despite the somewhat mixed evidence for non-linearities in US GNP from the formal testing procedures reviewed here, and in Section 3.4 for the SETAR model, the success of the non-linear models in characterizing important aspects of the business cycle, and the forecast successes claimed for such models in earlier studies, motivates the systematic appraisal of the forecast performance of such models.

5. COMPARISON OF EMPIRICAL FORECAST ACCURACY

The empirical forecast accuracy comparison is based on series of 'rolling' forecasts. For example, for the 1947–90 data vintage, the models are estimated once and for all on the sub-sample 1947:2–1984:4 (less observations are lost at the beginning of the sample from taking lags). A sequence of 1–16-step ahead forecasts are then generated (as described in Sections 2–4 using 1984:4 as the forecast origin. The forecast origin is then rolled forward one period to 1985:1, and another sequence of 1–16-step ahead forecasts is generated. The procedure is then repeated until we have 24×1-step forecasts, down to 8×16-step forecasts. This enables root-mean-squared forecast errors (RMSEs) to be calculated for each forecast horizon. For long horizons the smaller number of forecasts mean that the RMSE calculations are less reliable.

For the 1959–96 data sample the estimation period is 1959:2–1990:4, so that 22 observations are held back for out-of-sample forecasting.

Table 6. MSM(2)-AR(p) models for 47:2–90:4.

AR order p	0	1	2	3	4	5	6	7
Mean μ_L	−0.2011	−1.0804	−1.1246	−1.0657	−0.9919	−1.0209	−1.1154	−1.3190
Mean μ_H	1.1961	0.9660	0.9634	0.9573	0.9607	0.9635	0.9432	0.9432
α_1		0.3938	0.2893	0.3117	0.3210	0.3377	0.3364	0.3217
α_2			0.2271	0.2338	0.2629	0.2584	0.2623	0.3253
α_3				−0.0538	−0.0458	−0.0654	−0.0534	−0.0229
α_4					−0.0975	−0.1133	−0.1235	−0.0955
α_5						0.0583	0.0385	−0.1045
α_6							0.0496	−0.0378
α_7								0.2011
σ^2	0.7702	0.7352	0.7327	0.7404	0.7307	0.7248	0.7366	0.6792
Trans.prob p_{LL}	0.7747	0.5514	0.6493	0.6403	0.6163	0.6229	0.6291	0.5977
Trans.prob p_{HH}	0.9039	0.9539	0.9647	0.9646	0.9616	0.9614	0.9671	0.9636
Uncond.prob.L	0.2992	0.0933	0.0915	0.0896	0.0909	0.0929	0.0815	0.0829
Uncond.prob.H	0.7008	0.9067	0.9085	0.9105	0.9091	0.9071	0.9185	0.9171
Duration L	4.44	2.23	2.85	2.78	2.61	2.65	2.70	2.49
Duration H	10.39	21.67	28.33	28.26	26.07	25.90	30.40	27.51
Obs. in L	50.41	16.54	16.81	16.42	16.47	16.84	15.02	15.14
Obs. in H	117.59	151.46	151.19	151.58	151.53	151.16	152.99	152.86
AIC criterion	495.21	484.22	480.53	482.28	483.04	484.70	486.33	485.26
Loglikelihood	−242.61	−236.11	−233.27	−233.14	−232.52	−232.35	−232.17	−230.63

The results of the exercise are illustrated in Figures 5 and 6. On the earlier data vintage (1947–90) the 2-regime MS-AR models (the $p = 2, 4$ models are visually indistinguishable) are to be preferred, and record gains of up to 5% on horizons up to 6. The SETAR and MS(3)-AR(4) are as good as the AR model for these horizons, except at 1- and 2-steps ahead the SETAR is better. Clearly, the MS-AR 3-regime model's ability to better characterize the business-cycle features of the data (relative to the original Hamilton 2-regime model) does not directly translate into an improved forecast performance.

For the later data vintage, the AR and 2-state MS models are very similar, but the former has the edge at short horizons, while the SETAR and 3-state MS models are rather inaccurate at short horizon. Notice, however the much smaller vertical scale of the graph, reflecting the greater tranquility of the nineties.

The lesson we draw from these empirical comparisons is that the non-linear models do not always forecast better. This is perhaps unsurprising given that it is now reasonably well understood that for non-linear models 'how well we can predict depends on where we are' (Tong, 1995b, p. 410). Neither forecast period contains negative output growth of the severity witnessed historically, and it is precisely these episodes which should favour the non-linear models.

Table 7. MSIH(3)-AR(4) models.

Sample	47:2–84:4	47:2–90:4	59:2–90:4	59:2–96:2
Mean μ_H	3.0677	2.9844	1.6230	1.4435
Mean μ_M	1.2833	1.1911	0.8171	0.8659
Mean μ_L	−0.0894	−0.0251	−0.0953	−0.0625
α_1	0.0455	0.0680	−0.0467	0.0130
α_2	0.0762	0.0877	−0.0198	−0.0228
α_3	−0.1463	−0.1522	−0.0955	−0.1283
α_4	−0.1627	−0.1456	−0.0153	−0.0559
σ_H^2	0.1149	0.1478	0.3245	0.4050
σ_M^2	0.5117	0.4683	0.1013	0.1175
σ_L^2	0.9429	0.9123	0.8055	0.7724
Trans.prob p_{HH}	0.8388	0.8164	0.7434	0.9096
Trans.prob p_{HM}	0.1612	0.1836	0.2566	0.0904
Trans.prob p_{HL}	0.0000	0.0000	0.0000	0.0000
Trans.prob p_{MH}	0.0000	0.0000	0.1320	0.0000
Trans.prob p_{MM}	0.8955	0.8981	0.7754	0.9245
Trans.prob p_{ML}	0.1045	0.1019	0.0926	0.0755
Trans.prob p_{LH}	0.0255	0.0261	0.1472	0.1305
Trans.prob p_{LM}	0.1937	0.1800	0.0000	0.0216
Trans.prob p_{LL}	0.7808	0.7938	0.8528	0.8479
Uncond.prob.H	0.0486	0.0450	0.3495	0.3240
Uncond.prob.M	0.6442	0.6392	0.3993	0.4517
Uncond.prob.L	0.3072	0.3158	0.2512	0.2243
Duration H	6.2041	5.4477	3.8968	11.0656
Duration M	9.5659	9.8162	4.4524	13.2391
Duration L	4.5612	4.8502	6.7918	6.5751
Observations	147	171	123	145
Loglikelihood	−201.06	−226.10	−132.38	−145.65

6. MONTE CARLO STUDY

The Monte Carlo study allows us to evaluate the costs (in terms of forecast performance, as measured by RMSE) to using the 'wrong' non-linear model (or a linear model as an approximation to a non-linear model) when we abstract from the vagaries of the models only poorly representing the DGP, or of the non-linearities present in the past not persisting in the future.

Table 8. Standardized LR statistics for MS-AR model. See Hansen (1996a) for details of the test statistic, such as the definition of M. The results were obtained using Bruce Hansen's Gauss code `markovm.prg` with the 'Grid 3' option of Hansen (1996a).

			p-value					
	p	LR test	M					
			0	1	2	3	4	5
1952–84	4	1.546	0.713	0.713	0.622	0.658	0.650	0.652
1952–84	2	2.305	0.311	0.311	0.295	0.295	0.259	0.243
1947–90	4	1.255	0.856	0.817	0.803	0.795	0.768	0.745
1947–90	2	2.153	0.368	0.362	0.352	0.337	0.336	0.327
1959–96	4	2.410	0.254	0.237	0.264	0.260	0.244	0.246
1959–96	2	2.152	0.364	0.386	0.370	0.388	0.371	0.375

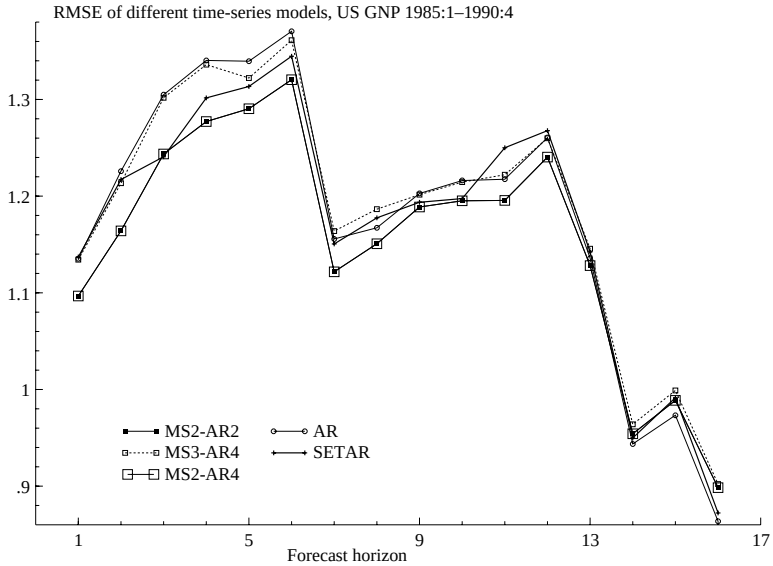

Figure 5. Empirical forecasting performance of the SETAR, MS-AR and AR models: 85:1–90:4.

By simulating the future to mimic the past, the 'non-linear features' occur in the forecast period, even if the non-linear structure captured in the empirical model was primarily due to 'outliers' and unhelpful for improving empirical forecasts.

Since the DGP is taken in turn to be each of the non-linear empirical models estimated over the full-sample of the earlier data vintage (1947–90), we are open to the charge of the specificity of our results to the particular design. While fully exploring the parameter space that might be of interest would require a very extensive set of simulations, with computational requirements

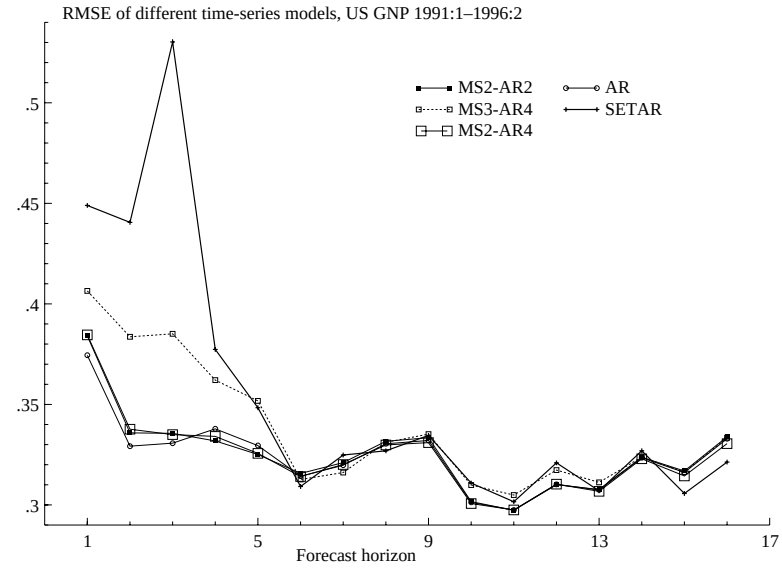

Figure 6. Empirical forecasting performance of the SETAR, MS-AR and AR models: 91:1–96:2.

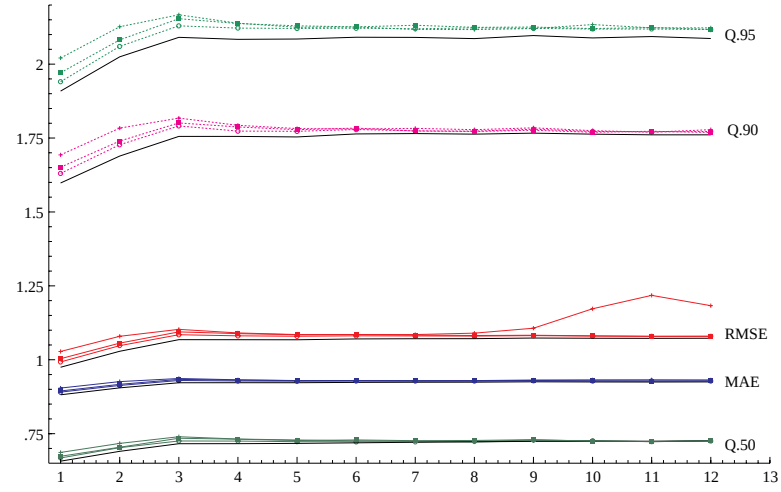

Figure 7. Monte Carlo. Forecast errors when the DGP is an AR(3) process.

that might be prohibitive given some of the necessary calculations (such as estimating the MS-AR, and forecasting the SETAR model), we have selected a few interesting departures from the estimated non-linear models to explore, and we elaborate upon these in Section 7. It seems preferable to use empirical models as the DGP, rather than an artificial DGP whose relevance for actual economics data may be questionable. It may of course be the case that although non-linearities are a feature of the DGP they are not large enough to yield much of an improvement to forecasting (see Diebold and Nason (1990)).

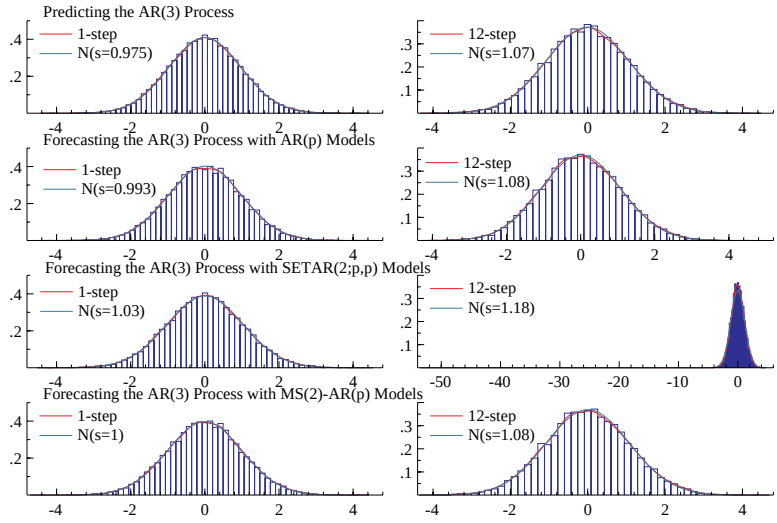

Figure 8. Monte Carlo. Forecast error density when the DGP is an AR(3) process.

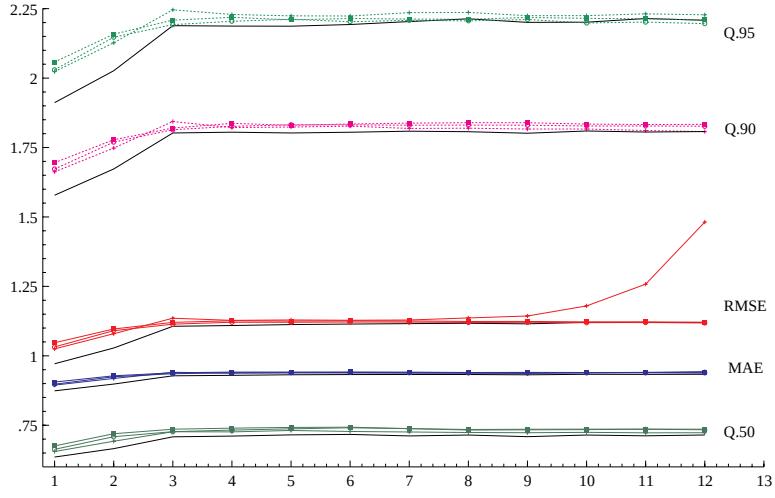

Figure 9. Monte Carlo. Forecast errors when the DGP is a SETAR(2; 2, 2) process.

Results are presented for the AR(3) DGP, the SETAR(2; 2, 2) DGP, and the MSM(2)-AR(2) DGP, respectively. For each, there are two figures. The first of each pair reports RMSEs and mean absolute error (MAE) measures, as well as selected quantiles of the distribution of absolute forecast errors. In each case, the legend is as follows: the solid line refers to the DGP, and the other lines are as in Figures 5 and 6 for the empirical analysis, i.e., the circles are the AR, the solid boxes the MS-AR, and the pluses the SETAR. The second figure in each pair plots the estimated forecast error densities with super-imposed Gaussian densities, for selected horizons. The MS-AR and SETAR are restricted to be two-regime models, but

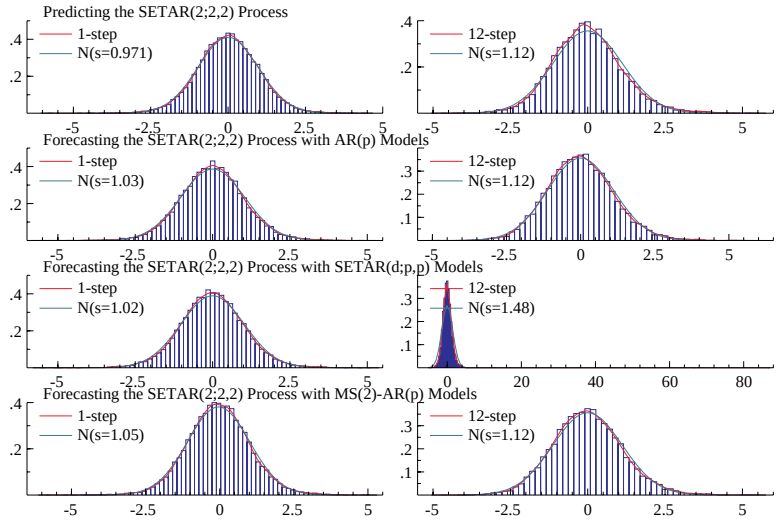

Figure 10. Monte Carlo. Forecast error density when the DGP is a SETAR(2; 2, 2) process.

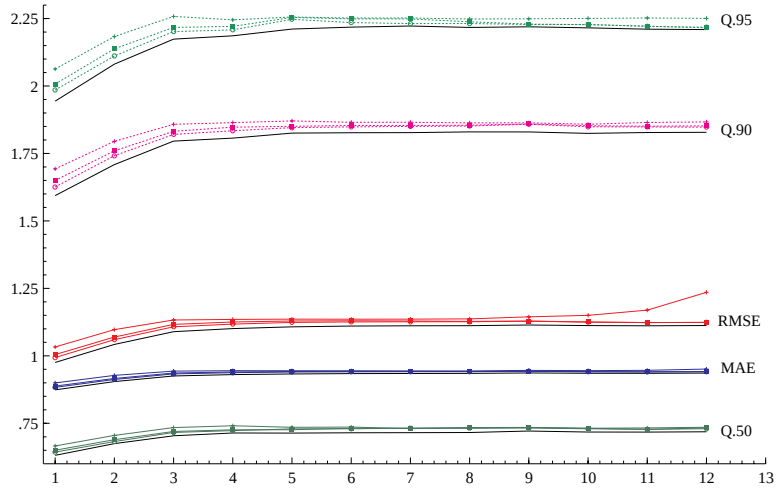

Figure 11. Monte Carlo. Forecast errors when the DGP is an MS(2)-AR(2) process.

are otherwise unrestricted, so that p (and d, r for the SETAR) are chosen on each iteration of the Monte Carlo to minimize AIC. As in the empirical exercise, the SETAR forecasts are calculated by Monte Carlo using 500 iterations. On a small number of iterations the SETAR model forecasts explode for both the SETAR and MS-AR DGPs. This only affects longer horizons, and we report RMSEs for the full 1,000 iterations, but also consider quantiles of the probability distribution of the absolute forecast errors that exclude the errant ones (for the SETAR DGP, for example, three of the 1,000 estimated models produced explosive forecasts).

Consider first the AR DGP, and Figures 7 and 8. The impact on forecast accuracy of

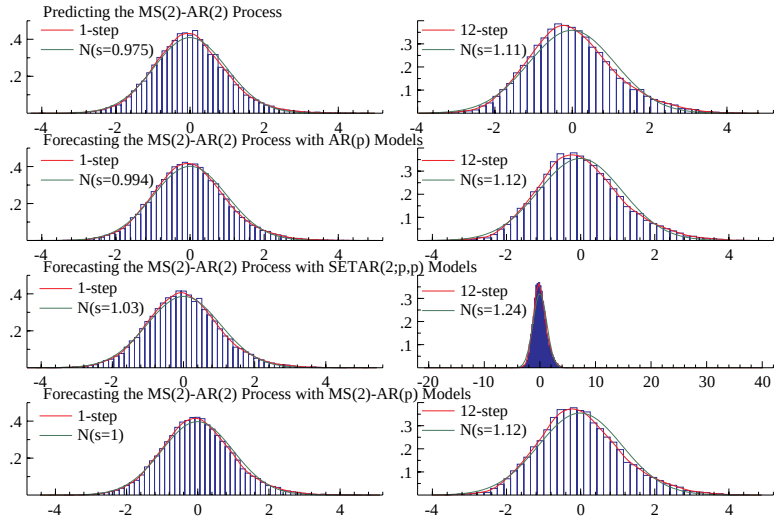

Figure 12. Monte Carlo. Forecast error density when the DGP is an MS(2)-AR(2) process.

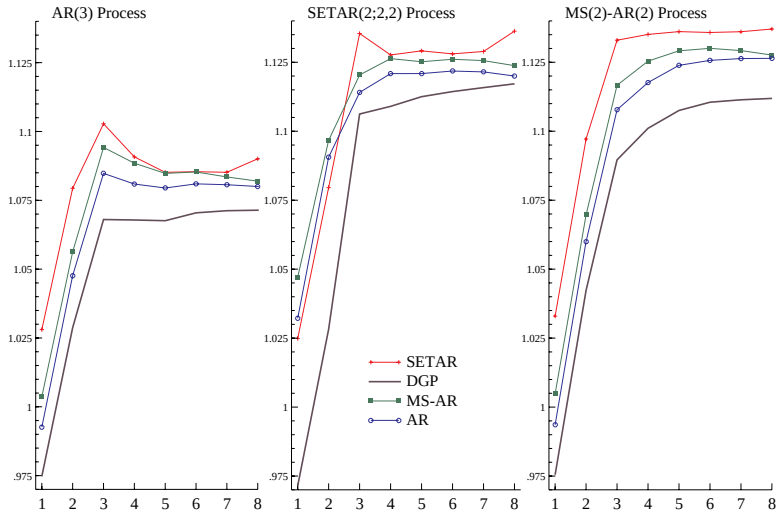

Figure 13. Monte Carlo comparison of the models on RMSE.

selecting the order of the AR model and estimating its parameters is given by the heights of the lines with circles above the solid lines. There appears to be little additional cost in terms of forecast accuracy to using one of the non-linear models. The SETAR method of generating forecasts tends to produce a slightly more dispersed forecast density – see the 95th percentile at short horizons, as well as a few explosive forecasts at longer horizons (see the RMSE, and the density in Figure 8).

Next, consider the SETAR DGP and Figures 9 and 10. First, the SETAR model is only

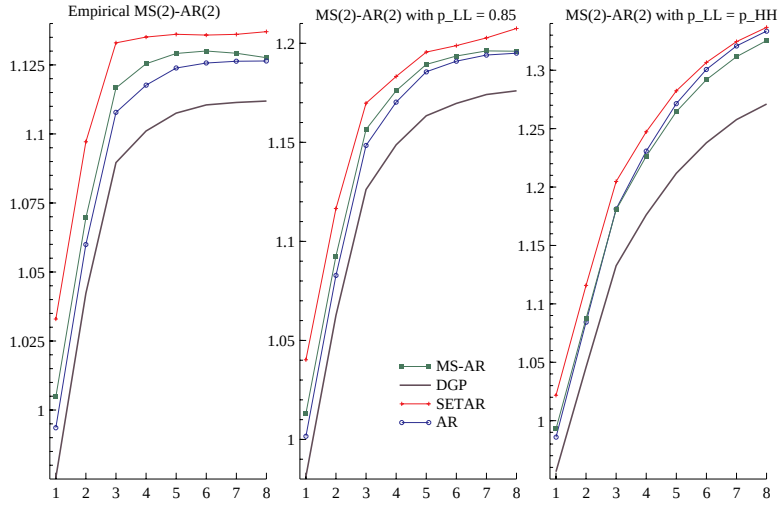

Figure 14. Post-simulation analysis I. RMSE when the DGP is an MS(2)-AR(2).

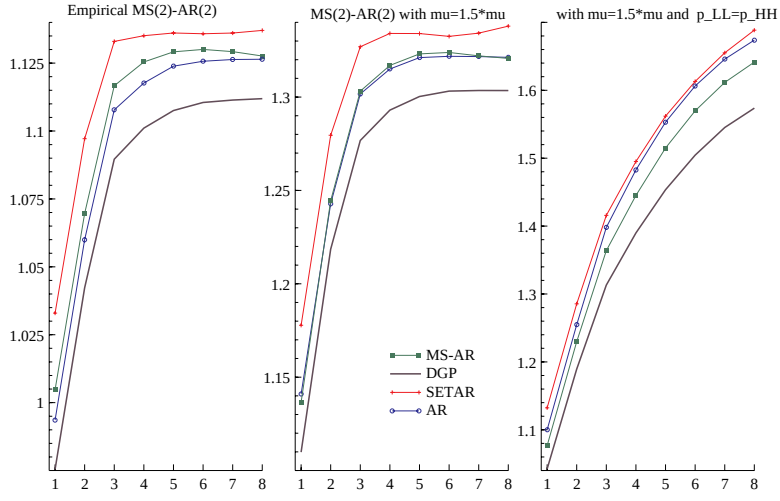

Figure 15. Post-simulation analysis II. RMSE when the DGP is an MS(2)-AR(2).

better than the AR model at 1- and 2-step horizons on RMSE. This is most clearly seen from Figure 13. The maximum cost to using the wrong non-linear model (the MS-AR model) occurs at these horizons, and is less than 2%. Thereafter, the AR outperforms both non-linear models.

Surprisingly, when the MS-AR is the DGP (Figures 11 and 12) the AR model is best at short horizons on RMSE. The cost of using the SETAR is generally greater than when the roles are reversed. The explanation has little to do with the initialization of the EM algorithm. Using the true parameter values as starting points for the EM algorithm instead of the data-

based initial values $\alpha_1^{(0)} = \cdots = \alpha_4^{(0)} = 0$, $p_{11}^{(0)} = p_{22}^{(0)} = 0.75$, $\sigma^{2(0)} = T^{-1}\sum_t(\Delta y_t - \overline{\Delta y})^2$, $\mu_1^{(0)} = \Delta y_{[0.25]}$, $\mu_2^{(0)} = \Delta y_{[0.75]}$, which might be fairly far away from the true parameter vector, has little effect on the forecasting performance of the estimated MS-AR model.

Nevertheless, the failure to improve on the forecast performance of linear models is the finding that warrants further investigation, and we take this up in Section 7, where we isolate the factors on which the performance of the MS-AR model turns.

7. POST-SIMULATION ANALYSIS

The Monte Carlo results employing empirical business-cycle models for the DGP have shown that linear models are relatively robust forecasting devices even when the DGP is non-linear. In this section we derive some theoretical explanations for this surprising outcome, and modify the Monte Carlo to illustrate. We focus on the MS-AR model since it allows an explicit analytical expression for the optimal predictor.

For the sake of simplicity consider an MSM(2)-AR(1):

$$\Delta y_t - \mu(s_t) = \alpha\{\Delta y_{t-1} - \mu(s_{t-1})\} + \epsilon_t, \tag{10}$$

which can be rewritten as the sum of two independent processes:

$$\Delta y_t - \mu_y = \mu_t + z_t,$$

where μ_y is the unconditional mean of Δy_t, such that $E(\mu_t) = E(z_t) = 0$. While the process z_t is Gaussian:

$$z_t = \alpha z_{t-1} + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2),$$

the other component, μ_t, represents the contribution of the Markov chain:

$$\mu_t = (\mu_2 - \mu_1)\zeta_t$$

where $\zeta_t = 1 - \Pr(s_t = 2)$ if $s_t = 2$ and $-\Pr(s_t = 2)$ otherwise. $\Pr(s_t = 2) = p_{12}/(p_{12} + p_{21})$ is the unconditional probability of regime 2. Invoking the unrestricted VAR(1) representation of a Markov chain (see Krolzig (1997c, p. 40)):

$$\zeta_t = (p_{11} + p_{22} - 1)\zeta_{t-1} + v_t,$$

then predictions of the hidden Markov chain are given by

$$\hat{\zeta}_{t+h|t} = (p_{11} + p_{22} - 1)^h \hat{\zeta}_{t|t}$$

where $\hat{\zeta}_{t|t} = \mathsf{E}(\zeta_t|Y_t) = \Pr(s_t = 2|Y_t) - \Pr(s_t = 2)$ is the filtered probability $\Pr(s_t = 2|Y_t)$ of being in regime 2 corrected for the unconditional one.

Thus the conditional mean of Δy_{t+h} is given by

$$\begin{aligned}\widehat{\Delta y}_{t+h|t} - \mu_y &= \hat{\mu}_{t+h|t} + \hat{z}_{t+h|t} \\ &= (\mu_2 - \mu_1)(p_{11} + p_{22} - 1)^h \hat{\zeta}_{t|t} + \alpha^h\{\Delta y_t - \mu_y - (\mu_2 - \mu_1)\hat{\zeta}_{t|t}\} \\ &= \alpha^h(\Delta y_t - \mu_y) + (\mu_2 - \mu_1)\{(p_{11} + p_{22} - 1)^h - \alpha^h\}\hat{\zeta}_{t|t}. \end{aligned} \tag{11}$$

The first term in (11) is the optimal prediction rule for a linear model (cf. (2)), and the contribution of the Markov regime-switching structure is given by the term multiplied by $\hat{\zeta}_{t|t}$, where $\hat{\zeta}_{t|t}$ contains the information about the most recent regime at the time the forecast is made. Thus the contribution of the non-linear part of (11) to the overall forecast depends on both the magnitude of the regime shifts, $|\mu_2 - \mu_1|$, and on the persistence of regime shifts $p_{11} + p_{22} - 1$ relative to the persistence of the Gaussian process, given by α.

In the empirical DGP, $p_{11} + p_{22} - 1 = 0.65$, and the largest root of the AR polynomial is 0.64, so that the second reason explains the success of the linear AR model in forecasting the MS-AR process. Since the predictive power of detected regime shifts is extremely small, $p_{11} + p_{22} - 1 \approx \alpha$ in (11), the conditional expectation collapses to a linear prediction rule.

In Figures 14 and 15 we explore the potential for outperforming linear forecasts by simulating variants of the empirical MS-AR process, where the directions of change are motivated by the discussion surrounding (11). Figure 14 gives the results for an increased persistence of recessions (p_{LL}). The left graph replicates the RMSEs for the empirical MS-AR process, for ease of comparison, and the middle and right graphs report results for increasing values of p_{LL}. Figure 15 depicts a 50% increase in the difference of the regime-dependent means $\mu_H - \mu_L$ (middle graph), and the right graph couples this with a higher persistence of recessions ($p_{LL} = p_{HH} = 0.9568$). Compared with the results for the empirical DGP given in the left graphs, we see improvements in the relative forecasting performance of MS-AR model (the effects are more dramatic if the lag length is constrained to 2 as in the DGP). The performance of the SETAR improves relative to the linear AR model, but is still dominated by it at all horizons.

8. CONCLUSION

In this paper we have evaluated the forecast performance of two popular non-linear extensions of the Box and Jenkins (1970) time-series modelling tradition, applied to modelling the growth rate of US GNP. By allowing for changes in regime in the process generating the time series, the models are proposed as contenders to the constant-parameter, linear time-series models of the earlier tradition. The SETAR and MS-AR models differ in how they model the movement between regimes, and thus the changes in the parameter values of the difference equations that govern the series. The SETAR model moves between regimes depending on the past realizations of the process. For the MS-AR model the movements between regimes are unrelated to the past realizations of the process, and result from the unfolding of an unobserved stochastic process, modelled as a Markov chain.

While both the MS-AR and SETAR models are superior to linear models in capturing certain features of the business cycle, their superiority from a forecasting perspective is less convincing. In the empirical forecast accuracy comparison exercise one of our sample periods suggests allowing for non-linearity may be beneficial, while the other suggests the opposite. The outcome is sensitive to the extent to which the future is characterized by 'non-linear features'.

The simulation-based comparison controls for certain factors, and further analysis uncovered characteristics of the models that contribute to a more favourable forecast performance. Our study indicates that the linear AR model is a relatively robust forecasting device, even when the data are generated by one of the non-linear models.

As noted in the introduction, the non-linear models would undoubtedly fare better if the evaluation exercises were made conditional upon the regime. Moreover, it is not clear that lag-order selection by AIC is optimal for non-linear models, especially from a forecasting perspective.

ACKNOWLEDGEMENTS

Financial support from the UK Economic and Social Research Council under Grant L116251015 is gratefully acknowledged by both authors. Helpful comments were received from conference participants at EC^2, Florence, the Royal Economic Society Annual Meeting, University of Staffordshire, Stoke-on-Trent, the 5th Conference CEMAPRE, Lisbon, the 1997 Meeting of the Society of Economic Control, Oxford, the Econometric Society European Meeting, Toulouse, and seminar audiences at Nuffield College, Warwick and the European University Institute. We are also grateful to Giampiero Gallo and Grayham Mizon for comments. All the computations reported in this paper were carried out in Ox: see Doornik (1996).

A. APPENDIX

In this appendix we outline the Hansen (1996b) test statistics reported in Section 3.4.

We begin by writing (3) as:

$$x_t = x'_{t-1}\beta_1 + x'_{t-1}(\gamma)\beta_2 + \epsilon_t \tag{12}$$

where $x'_{t-1}(\gamma) = \mathtt{I}_d(r)x'_{t-1}$, $x'_{t-1} = [1 \;\; x_{t-1} \;\; \ldots \;\; x_{t-p}]$, $\mathtt{I}_d(r) = 1$ when $x_{t-d} \leq r$, and 0 otherwise, and $\beta'_i = [\alpha_0^{\{i\}} \;\; \ldots \;\; \alpha_p^{\{i\}}]$, $i = 1, 2$. Denote estimates under H_0 by a $\tilde{}$, so that $\tilde{\beta}_1 = (\sum x_{t-1}x'_{t-1})^{-1}(\sum x_{t-1}x_t)$, for example. Under H_1, $\hat{\beta} = (\hat{\beta}_1 : \hat{\beta}_2)$ are estimated conditional on γ, denoted $\hat{\beta}(\gamma)$, and γ is then estimated as

$$\hat{\gamma} = \arg\min \hat{\sigma}_\epsilon^2 = \arg\min\Big(\sum\{x_t - x'_{t-1}\hat{\beta}_1 - x'_{t-1}(\gamma)\hat{\beta}_2\}^2\Big)$$

and $\hat{\beta} = \hat{\beta}(\hat{\gamma})$. The scores are defined by $s_t(\gamma) = \mathrm{x}_{t-1}(\gamma)\epsilon_t$, where $\mathrm{x}_{t-1}(\gamma) = [x'_{t-1} : x'_{t-1}(\gamma)]'$, and $\tilde{s}_t(\gamma) = \mathrm{x}_{t-1}(\gamma)\tilde{\epsilon}_t$ and $\hat{s}_t(\gamma) = \mathrm{x}_{t-1}(\gamma)\hat{\epsilon}_t(\gamma)$ give the sample estimates under H_0 and H_1.

A heteroskedastic robust LM test takes the form

$$T_T(\gamma) = T\hat{\beta}'_2(\gamma)\{R'\hat{V}_T^*(\gamma)R\}^{-1}\hat{\beta}_2(\gamma)$$

where $R = (0_{p+1}I_{p+1})'$, $\hat{V}_T^*(\gamma) = M_T(\gamma)^{-1}\hat{V}_T(\gamma)M_T(\gamma)^{-1}$, $\hat{V}_T = T^{-1}\sum \tilde{s}_t(\gamma)\tilde{s}_t(\gamma)'$, and $M_T(\gamma) = T^{-1}\sum \mathrm{x}_{t-1}(\gamma)\,\mathrm{x}_{t-1}(\gamma)'$. Under homoscedastic errors, the LM statistic simplifies to

$$T_T(\gamma) = T\tilde{\sigma}_\epsilon^2\hat{\beta}'_2(\gamma)(R'M_T(\gamma)^{-1}R)^{-1}\hat{\beta}_2(\gamma)$$

where $\tilde{\sigma}_\epsilon^2 = (T - (p+1))^{-1}\sum_{t=1}^T \tilde{\epsilon}_t^2$, the estimated error variance under the null.

$T_T(\gamma)$ has an approximate χ^2 distribution with $p+1$ degrees of freedom under the null when γ is known. When γ is unknown, Hansen (1996b) shows how asymptotic p-values can be simulated for functions of $T_T(\gamma)$, denoted $g_T(T_T(\gamma))$, that map from Γ to R. The supremum, $\sup T_T \equiv \sup_{\gamma\in\Gamma} T_T(\gamma)$ was considered by Davies (1977) and (1987) as a way of testing H_0, and ave T_T (the average of T_T

over all values of Γ) and $\exp T_T$ (ln of the average of $\exp\{T_T(\gamma)/2\}$ over $\gamma \in \Gamma$, have recently been considered by Andrews and Ploberger (1994).

Whatever the functional, p-values can be obtained by simulating the conditional distribution function of the functional, denoted $\tilde{F}_T$, where by conditional is meant conditional on the data $[x_t : \mathrm{x}_{t-1}(\gamma)]$. For $j = 1, \ldots, J$ ($J = 1000$, say), generate a sample of $NID(0, 1)$ variables, $\{v_{tj}\}_{t=1}^{T}$. Then calculate,

$$S_T^j(\gamma) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathrm{x}_{t-1}(\gamma)\tilde{\epsilon}_t v_{tj} \tag{13}$$

and then,

$$T_T^j = S_T^j(\gamma)' M_T(\gamma)^{-1} R(R'\hat{V}_T^*(\gamma)R)^{-1} R' M_T(\gamma)^{-1} S_T^j(\gamma). \tag{14}$$

Finally, calculate $g_T^j = g(T_T^j)$. The p-value is then the percentage of the random sample $(g_T^1, \ldots, g_T^J)$ that exceeds the actual test statistic g_T, $\tilde{p}_T = J^{-1} \sum_{j=1}^{J} 1(g_T^j \geq g_T)$.

REFERENCES

Akaike, A. (1973). Information theory and an extension of the maximum likelihood principle, In B. N. Petrov and F. L. Saki, (Eds.) *2nd Int. Symp. Inf. Theory*. Budapest.

Albert, J. and S. Chib (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bus. Econ. Stat. 11*, 1–16.

Al-Qassam, M. S. and J. A. Lane (1989). Forecasting exponential autoregressive models of order 1. *J. Time Ser. Anal. 10*, 95–113.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*, 821–56.

Andrews, D. W. K. and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica 62*, 1383–414.

Box, G. E. P. and G. M. Jenkins (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

Cecchetti, S. G., P. Lam and N. C. Mark (1990). Mean reversion in equilibrium asset prices. *Am. Econ. Rev. 80*, 398–418.

Clements, M. P. and D. F. Hendry (1996). Intercept corrections and structural change. *J. Appl. Econometrics 11*, 475–94.

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

Clements, M. P. and J. Smith (1996). A Monte Carlo study of the forecasting performance of empirical SETAR models. Warwick Economic Research Papers No. 464, Department of Economics, University of Warwick.

Clements, M. P. and J. Smith (1997). The performance of alternative forecasting methods for SETAR models. *Int. J. Forecasting 13*, 463–75.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 64*, 247–54.

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 P. De Bruin (1997). On SETAR forecasting. *Stat. Prob. Lett.* Forthcoming.

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

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B, 39*, 1–38.

Diebold, F. X. and C. Chen (1996). Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *J. Econometrics 70*, 221–41.

Diebold, F. X., J. H. Lee and G. C. Weinbach (1994). Regime switching with time-varying transition probabilities. In C. Hargreaves (Ed.), *Non-stationary Time-series Analyses and Cointegration*, pp. 283–302. Oxford: Oxford University Press.

Diebold, F. X. and J. A. Nason (1990). Nonparametric exchange rate prediction. *J. Int. Econom., 28*, 315–32.

Diebold, F. X., G. D. Rudebusch and D. E. Sichel (1993). Further evidence on business cycle duration dependence. In J. Stock and M. Watson (Eds.), *Business Cycles, Indicators, and Forecasting*, pp. 255–80. Chicago: University of Chicago Press and NBER.

Doornik, J. A. (1996). *Object-Oriented Matrix Programming using Ox*. London: International Thomson Business Press and Oxford: `http://www.nuff.ox.ac.uk/Users/Doornik/`.

Durland, J. M. and T. H. McCurdy (1994). Duration dependent transitions in a Markov model of U.S. GNP growth. *J. Bus. Econ. Stat. 12*, 279–288.

Filardo, A. J. (1994). Business-cycle phases and their transitional dynamics. *J. Bus. Econ. Stat. 12*, 299–308.

Fildes, R. and S. Makridakis (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *Int. Stat. Rev. 63*, 289–308.

Ghysels, E. (1994). On the periodic structure of the business cycle. *J. Bus. Econ. Stat. 12*, 289–298.

Goodwin, T. H. (1993). Business-cycle analysis with a Markov-switching model. *J. Bus. Econ. Stat., 11*, 331–9.

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

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–84.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *J. Econometrics 45*, 39–70.

Hansen, B. E. (1992). The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *J. Appl. Econometrics 7*, S61–82.

Hansen, B. E. (1996a). Erratum: The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *J. Appl. Econometrics 11*, 195–8.

Hansen, B. E. (1996b). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica 64*, 413–30.

Hess, G. D. and S. Iwata (1995). *Measuring Business Cycle Features*: University of Kansas, Research Papers in Theoretical and Applied Economics No. 1995–6.

Kähler, J. and V. Marnet (1994). Markov-switching models for exchange rate dynamics and the pricing of foreign-currency options. In J. Kähler and P. Kugler (Eds.), *Econometric Analysis of Financial Markets*: Heidelberg: Physica Verlag.

Kim, C. J. (1994). Dynamic linear models with Markov-switching. *J. Econometrics 60*, 1–22.

Koop, G., M. H. Pesaran and S. M. Potter (1996). Impulse response analysis in nonlinear multivariate models. *J. Econometrics 74*, 119–47.

Krolzig, H.-M. (1997a). Forecasting Markov-switching vector autoregressive processes. Mimeo, Institute of Economics and Statistics, University of Oxford.

Krolzig, H.-M. (1997b). International business cycles: Regime shifts in the stochastic process of economic growth. Applied Economics Discussion Paper 194, University of Oxford.

Krolzig, H.-M. (1997c). *Markov Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*: Lecture Notes in Economics and Mathematical Systems, 454. Berlin: Springer-Verlag.

Krolzig, H.-M. and H. Lütkepohl (1995). Konjunkturanalyse mit Markov–Regimewechselmodellen. In K. H. Oppenländer (Ed.), *Konjunkturindikatoren. Fakten, Analysen, Verwendung*, pp. 177–96. Oldenbourg: München Wien.

Lahiri, K. and J. G. Wang (1994). Predicting cyclical turning points with leading index in a Markov switching model. *J. Forecasting 13*, 245–63.

Lam, P.-S. (1990). The Hamilton model with a general autoregressive component. Estimation and comparison with other models of economic time series. *J. Monetary Econ. 26*, 409–32.

McCulloch, R. E. and R. S. Tsay (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *J. Time Ser. Anal. 15*, 235–50.

Phillips, K. (1991). A two-country model of stochastic output with changes in regime. *J. Int. Econ. 31*, 121–42.

Potter, S. (1995). A nonlinear approach to U.S. GNP. *J. Appl. Econometrics 10*, 109–25.

Sensier, M. (1996). *Investigating Business Cycle Asymmetries in the UK*: University of Sheffield, Ph.D. Thesis.

Teräsvirta, T. and H. M. Anderson (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *J. Appl. Econometrics 7*, 119–39.

Tiao, G. C. and R. S. Tsay (1994). Some advances in non-linear and adaptive modelling in time-series. *J. Forecasting 13*, 109–31.

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

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

Tong, H. (1995a). *Non-linear Time Series. A Dynamical System Approach*. Oxford: Clarendon Press. 2nd edition.

Tong, H. (1995b). A personal overview of non-linear time series analysis from a chaos perspective. *Scandinavian J. Stat. 22*, 399–445.

Tong, H. and K. S. Lim (1980). Threshold autoregression, limit cycles and cyclical data. *J. Royal Stat. Soc. B42*, 245–92.