Advances in Management & Applied Economics, Vol. 11, No. 6, 2021, 117-138
ISSN: 1792-7544 (print version), 1792-7552(online)
https://doi.org/10.47260/amae/1167
Scientific Press International Limited

Evaluating Forecasts from State-Dependent Autoregressive Models for US GDP Growth Rate. Comparison with Alternative Approaches

Fabio Gobbi[1]

Abstract

The aim of the paper is to compare the forecasting performance of a class of state-dependent autoregressive (SDAR) models for univariate time series with two alternative families of nonlinear models, such as the SETAR and the GARCH models. The study is conducted on US GDP growth rate using quarterly data. Two methods of forecast comparison are employed. The first method consists in evaluation the average performance by using two measures such as the root mean square error (RMSE) and the mean absolute error (MAE) over different forecast horizons, while the second method make use of one of the most used statistical test to compare the accuracy of two forecast methods such as the Diebold-Mariano test.

JEL classification numbers: C22, E37, F47.
Keywords: Nonlinear models for time series, GDP growth rate, Forecasting accuracy.

[1] Department of Economics and Statistics, University of Siena, Italy.

Article Info: *Received:* October 1, 2021. *Revised*: October 20, 2021.
Published online: October 22, 2021.

1. Introduction

In this paper we propose a class of state-dependent autoregressive models (SDAR) to study nonlinearities in economic time series as the quarterly US GDP growth rate. The aim is to compare the predictive ability of SDAR models with respect to linear autoregressive (AR) time series models and two leading classes of nonlinear models such as the self-exciting threshold autoregressive (SETAR) model and the generalized autoregressive conditional heteroskedasticity (GARCH) model, that have already been proposed for US GDP. The problem we address is whether SDAR models offer a much improved forecast performance.

The class of SDAR models is a generalized version of a first-order autoregressive process where the autoregressive coefficient depends on the first lagged state variable whose equation is:

$$y_t = \alpha + \varphi(y_{t-1};\gamma)y_{t-1} + \varepsilon_t \tag{1.1}$$

where $\varphi(\cdot;\cdot)$ is a *specified* function satisfying some assumptions and depending on a set of parameters, γ. The error term ε_t is independent of y_{t-1} with zero mean and volatility σ. SDAR models are closely related to the functional-coefficient autoregressive (FAR) models introduced by Chen and Tsay (1993) where p autoregressive coefficients are given by measurable functions depending on $k<p$ lagged values of y_t. Within this framework, Cai, Fan and Yao (2000) adopt local linear regression techniques to estimate functional coefficient regression models for times series data while Chen and Liu (2001) study nonparametric estimation and hypothesis testing procedures for the same model. In Cherubini and Gobbi (2013) SDAR models are derived as a special case of a more general convolution-based autoregressive processes in which the error term is not independent of the lagged value of the state variable (see also Cherubini, Gobbi and Mulinacci, 2016). More recently, Gobbi and Mulinacci (2020) define the class of SDAR models more rigorously establishing their main statistical properties, such as stationarity and ergodicity, and determine the asymptotic behaviour of the quasi-maximum likelihood (QML) estimator of the parameters. In particular, the authors show that if $|\varphi(y)| \leq \delta < 1$ uniformly in y, the process in (1.1) is geometrically ergodic and strictly stationary. In the same paper, the authors compare the forecast performance of two specifications of SDAR models with SETAR models for time series of weekly realized volatilities extracted from three different European financial indexes, showing that SDAR models ensure a gain in the accuracy for two cases on three, at least for short and medium forecast horizons. Furthermore, Gobbi (2020) documents, through a Monte Carlo experiment, that nonlinearity in time series generated from a SDAR model strictly depends on the functional form of persistence function ψ and on the value of parameters.

A class of alternative nonlinear models we consider in this paper is the self-exciting threshold autoregressive (SETAR) models, which were first proposed and studied by Tong (1978, 1986 and 1995) and Tong and Lim (1980). In SETAR models the variable y_t is a linear autoregression within a regime but may move among regimes depending on the value taken by a lag of y_t itself. A number of authors have estimated SETAR models of US GDP. Tiao and Tsay (1994) consider a two regime SETAR model, Potter (1995) estimates a SETAR(2,5,5) but with the third and fourth regimes restricted to zero in both regimes. Both papers use time series from 1947 to 1990. A key feature of SETAR models for US GDP over this period is a large and negative coefficient on the second lag in the lower regime, indicating that US economy moves rapidly out of recession periods. Moreover, Tiao and Tsay (1994) find that the forecast performance of the SETAR model relative to a linear AR model is improved when the comparison is made when the economy is in recession (i.e., the lower regime is activated). Clements and Smith (1997) implement a Monte Carlo simulation to show that there is an significant effect of the regimes on the forecast accuracy. In particular, the authors find that the gain in the lower regime need to be sufficiently large for the SETAR to perform well on average.

The second alternative class of nonlinear models we use is represented by the generalized autoregressive conditional heteroscedasticity (GARCH) models developed by Bollerslev (1986) as an extension of ARCH models introduced by Engle (1982). GARCH models are nonlinear in variance since their crucial feature is the heteroskadasticity which assumes that volatility is not constant over time. Since the US GDP growth rate involves long-run phenomena, structural changes in volatility can occur with high probability. Kim and Nelson (1999), McConnel and Perez-Quiros (2000), Blanchard and Simon (2001) among others document a structural change in volatility of US GDP growth rate.

Hamori (2000) presents evidence that GARCH(1,1) structure is reasonable for US GDP. However, Bollerslev, Chou and Kroner (1992) notice that while GARCH effect is highly significant with daily and weekly financial data, it tends to be much milder in less frequently sample time series such as quarterly US GDP growth rate. In this paper, we estimate an AR-GARCH model in which the GARCH structure in the variance equation is combined with an autoregressive structure of the mean equation.

Our aim is to measure the forecasting accuracy for the US GDP growth rate of four different classes of nonlinear models mentioned above, SDAR, SETAR and AR-GARCH, using the linear AR as a benchmark. The evaluation of the forecast accuracy of different models adopted is conducted according to two different criteria. We first evaluate the average performance using the root mean square error (RMSE) and the mean absolute error (MAE) over different forecast horizons, from 1 to 8 quarters ahead. The second criteria is provided by the Diebold-Mariano test (DM), introduced and implemented by Diebold and Mariano (1995), to compare the forecast accuracy of two forecast methods. We use a modified version of the test proposed by Harvey, Leybourne and Newbold (1997) particularly adapted for small samples. We will show that whereas the first criteria highlights a higher performance of SDAR models with respect to the alternatives analyzed, the same conclusion is not completely confirmed by the second criteria.

The paper is organized as follows. Section 2 describes the data set used in the empirical analysis. Section 3 briefly introduces the models adopted. Section 4 reports and discusses the estimation results. In section 5 we present the forecast accuracy comparison among the models. Section 6 concludes.

2. Preliminary Data Analysis

The empirical data analysis has been carried out on the quarterly US GDP growth rate. The observation period goes from 1950.Q2 until 2017.Q3 (270 observations) and is depicted in figure 1. The series appears mean-stationary while the variance features the volatility clustering phenomenon with periods with high volatility followed by periods of low volatility. Furthermore, volatility is higher in the first part of the time series (indicatively until the 1980s).

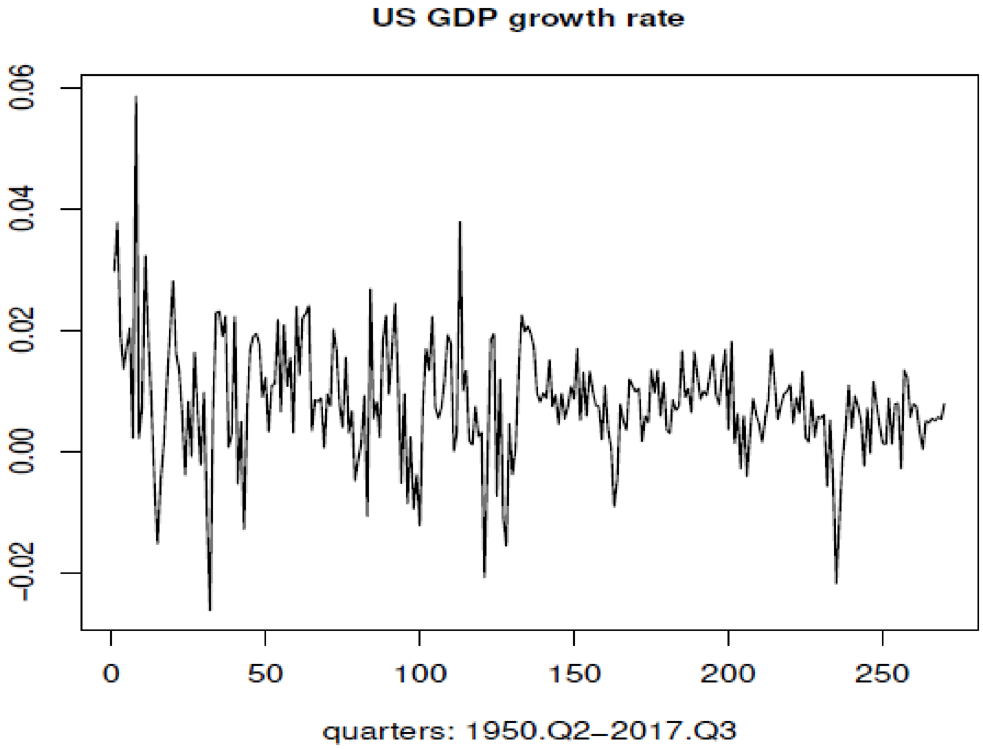

Figure 1: Quarterly US GDP growth rate

2.1 Descriptive Statistics of the US GDP Growth Rate

Table 1 reports the summary of the descriptive statistics of the US GDP growth rate. The series is characterized by excess kurtosis and positive asymmetry. The asymmetry characterized by positive skewness means that in the sample period a greater probability exists of large increases in GDP growth than larger decreases while the kurtosis exhibits leptokurticity with fat tails highlighting that extreme changes can occur more frequently. The Jarque-Bera test (Jarque and Bera, 1980 and 1987) strongly rejects the normality hypothesis. Furthermore, the Ljung-Box test (Ljung and Box, 1978) indicates autocorrelation (up to 20 lags) in the time series. The McLeod-Li test (McLeod and Li, 1983) suggests a time-varying variance structure leading to the rejection of the null of no ARCH components up to 20 lags.

Table 1: Descriptive statistics for US GDP growth rate

N. of obs.	270
Mean	0.00802
Median	0.00768
Maximum	0.05866
Minimum	-0.02630
Std. dev.	0.00969
Skewness	0.33119
Kurtosis	3.43276
Jarque-Bera (p-value)	0.00000
Ljung-Box (p-value)	0.00005
McLeod-Li (p-value)	0.00485

2.2 Linearity Tests

Table 2 reports the p-values of four different linearity tests performed on the full sample and on the last ten years of observations. For each test we consider different lag structures (lag=1,2,3). We employ four different linearity tests intensively used in the literature: the TNN test, the WNN test, the Tlrt test and the Tsay test. In the Terasvirta Neural Network test (TNN test), introduced in Terasvirta, Lin and Granger, (1993), and in the White Neural Network test (WNN test), discussed in Lee, White and Granger (1993), the null is the hypotheses of linearity in mean. The Tlrt test carry out the likelihood ratio test for threshold nonlinearity and was implemented by Chan (1990). The null hypothesis is that the fitted model to the time series is an AR model with a specified lag structure and the alternative is that the fitted model is a threshold autoregressive model with the same lag structure for each regime. Finally, the Tsay test, which was introduced and implemented in Tsay (1986) is a test for quadratic nonlinearity in a time series in which the null hypothesis is a normal AR process.

The results show that there is no strong evidence of nonlinearity in the full series, since in a number of cases tests lead to the acceptance of linearity. However, for at least one lag all tests reject the null. In particular, the TNN test highlights low p-values (less than 10\%) regardless of the lag structure assumed. On the other hand, the Tlrt test and the Tsay test reject the null of linearity only for a lag structure equal to 3, reflecting a weakness of the hypothesis of quadratic and threshold autoregressive nonlinearities.

In order to realize whether the nonlinearity structure strengthens or not in more recent period, we conduct the same linearity tests in a portion of the sample corresponding to the last 10 years of observations. Unfortunately, table 2 shows a weakening of the nonlinearity to the point that only in one case (TNN test with lag = 3) the hypothesis is rejected. It is possible that this result can have consequences from the point of view of the forecasting evaluation as argued in Granger and Terasvirta (1993) and in Terasvirta and Anderson (1992). Indeed, in that papers the authors suggest that the superior in-sample performance of nonlinear models will only be matched out-of-sample if the nonlinear features also characterize the later period of observation. Furthermore, even Ljung-Box and McLeod-Li tests provide p-values significantly high (0,8672 and 0.9917) indicating that this last portion of the time series is free from autocorrelation and heteroskedasticity.

Table 2: Linearity tests. p-values for different lag structures and different portions of the observed time series.

US GDP: full sample 1950.Q2-2017.Q3			
	lag=1	**lag=2**	**lag=3**
TNN test	**0.0534**	**0.0762**	**0.0029**
WNN test	**0.0512**	0.4591	0.4793
Tlrt test	0.3312	0.3112	**0.0507**
Tsay test	0.1005	0.2152	**0.0003**
US GDP: last 10 years 2008.Q1-2017.Q3			
	lag=1	**lag=2**	**lag=3**
TNN test	0.4152	0.1143	**0.0001**
WNN test	0.4161	0.3836	0.3826
Tlrt test	0.2191	0.3304	0.1046
Tsay test	0.2633	0.6185	0.3397

Note: The cases highlighted in bold lead to the rejection of the null

3. The Models

With regard to introducing the models just proposed in the introduction, we briefly present their representation referring for more details to the cited literature.

The benchmark model is the standard linear autoregressive model of order *p* (AR(p)) which has the following equation:

$$\begin{cases} y_t = \alpha + \sum_{i=1}^{p} \theta_i y_{t-i} + \varepsilon_t, \quad \mathrm{t} \geq 2 \\ \qquad \varepsilon_t \sim IID\ N(0,\sigma) \end{cases} \tag{3.2}$$

where ε_t is independent of the lagged variables $y_{t-1}, \ldots., y_{t-p}$. The vector of parameters is $\vartheta = (\alpha, \theta_1, \ldots, \theta_p, \sigma)$. The reader interested to linear autoregressive models can consult among others Hamilton (1994) and Brockwell and Davis (1991). To compare forecasting accuracy of US GDP growth rate we will specify and estimate five alternative nonlinear models within three different classes of models: SETAR, GARCH and SDAR. Below we briefly outline their representation.

- Self-exciting threshold autoregressive (SETAR) models were first proposed in Tong (1978, 1983), Tong and Lim (1980) and discussed in detail in Tong (1995). SETAR models considered in this paper assume that a variable y_t is a linear autoregression within a regime, but may move between regimes depending on the value assumed by the first lag y_{t-1}. We estimate two SETAR models, the first with two regimes and the second with three regimes. We denote SETAR(2, p_1, p_2) the model with two regimes whose specification is

$$\begin{cases} y_t = \alpha_1 + \sum_{i=1}^{p_1} \theta_{1,i} y_{t-i} + \varepsilon_{1,t}, \quad y_{t-1} \leq v \\ y_t = \alpha_2 + \sum_{i=1}^{p_2} \theta_{2,i} y_{t-i} + \varepsilon_{2,t}, \quad y_{t-1} \geq v \end{cases} \tag{3.3}$$

where v is the threshold variable, p_1 and p_2 are the orders of the linear AR within each regime, $\varepsilon_{j,t} \sim IID\ N(0, \sigma_j)$, *j*=1,2.
Furthermore $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$are independent for all *t*.
The vector of parameters is $\vartheta = (\alpha_1, \alpha_2, \theta_{1,1}, \ldots, \theta_{1,p_1}, \theta_{2,1}, \ldots, \theta_{2,p_2}, \sigma_1, \sigma_2)$.
SETAR models with three regimes, denoted by SETAR(3, p_1, p_2, p_3) are defined as:

$$\begin{cases} y_t = \alpha_1 + \sum_{i=1}^{p_1} \theta_{1,i} y_{t-i} + \varepsilon_{1,t}, \quad y_{t-1} \leq v_1 \\ y_t = \alpha_2 + \sum_{i=1}^{p_2} \theta_{2,i} y_{t-i} + \varepsilon_{2,t}, \quad v_1 < y_{t-1} \leq v_2 \\ y_t = \alpha_3 + \sum_{i=1}^{p_3} \theta_{3,i} y_{t-i} + \varepsilon_{3,t}, \quad y_{t-1} > v_2 \end{cases} \tag{3.4}$$

where v_1 and v_2 are two threshold variables, p_1, p_2, p_3 are the orders of the linear AR within each regime $\varepsilon_{j,t} \sim IID\ N\left(0, \sigma_j\right)$, *j*=1,2,3. The vector of parameters is $\vartheta = (\alpha_1, \alpha_2, \alpha_2, \theta_{1,1}, \ldots, \theta_{1,p_1}, \theta_{2,1}, \ldots, \theta_{2,p_2}, \theta_{3,1}, \ldots, \theta_{3,p_3}, \sigma_1, \sigma_2, \sigma_3)$.

- GARCH models were proposed in Bollerslev (1986) as a generalization of ARCH model introduced in Engle (1982). In this paper we consider an AR(p) component in place of a constant mean for the equation of the variable y_t in light of the preliminary analysis carried out in the previous section on the time series of US GDP growth rate.

Therefore, our specification of the model is the following:

$$\begin{cases} y_t = \alpha + \sum_{i=1}^{p} \theta_i y_{t-i} + \varepsilon_t, \quad t \geq 1 \\ \varepsilon_t | F_{t-1} \sim IID\ N(0, h_t) \\ h_t^2 = \omega_0 + \omega_1 y_{t-1}^2 + \omega_2 h_{t-1}^2 \end{cases} \tag{3.5}$$

where $F_{t-1}$is the information set which includes the lagged values of the variable y_{t-1}, $y_{t-2}, \ldots$. and the conditional variance has a GARCH(1,1) specification. The vector of parameters is $\vartheta = \left(\alpha, \theta_1, \ldots, \theta_{p,}, \omega_0, \omega_1, \omega_2\right)$.

- State-dependent autoregressive (SDAR) models have recently been discussed in Gobbi and Mulinacci (2020) where two specifications of them have been proposed. The models are characterized by an autoregressive coefficient which is a function of the first lagged variable y_{t-1} and their equation is of the form (1.1). The *persistence* function $\varphi(y_{t-1}; \gamma)$ depends on a set of parameters γ and must satisfy a number of assumptions in order to guarantee that the resulting process $(y_t)_{t\geq 1}$is stationary and ergodic, as shown in Gobbi and Mulinacci (2020). The choice of the function ψ completely determines the SDAR model. In this paper we consider two specifications of the model, denoted by SDAR1 and SDAR2. Both satisfy the required assumptions as shown in Gobbi and Mulinacci (2020). The first SDAR1 model is defined as

$$\begin{cases} y_t = \alpha + e^{-(\gamma_0 + \gamma_1 y_{t-1}^{2r})} y_{t-1} + \varepsilon_t, \ \ \mathrm{t} \geq 2 \\ \varepsilon_t \sim IID\ N(0, \sigma) \end{cases} \tag{3.6}$$

Where γ_0, $\gamma_1 > 0$. The error term $\varepsilon_t$is independent of y_{t-1} for all *t*. The vector of parameters is $\gamma = (\alpha, \gamma_0, \gamma_1, r, \sigma)$. Remark that this specification is a generalization of EXPAR models introduced by Haggan and Ozaki (1981). Some insights about the persistence function $e^{-(\gamma_0 + \gamma_1 y_{t-1}^{2r})}$ are needed. We can notice that it is decreasing with y_{t-1} and always within (0,1). Moreover, once fixed γ_0 and γ_1, its maximum is $e^{-\gamma_0}$ assumed for $y_{t-1} = 0$. From the point of view of economic interpretation, this means that the persistence induced by the autoregressive coefficient tends to be higher when the (quarterly) GDP growth rate is low, whereas it decreases when the rates are bigger both positive (strong expansion) or negative (recession). Since the SDAR model accounts only the first lag of the variable, we have chosen a function that cannot assume negative values for negative levels of y_{t-1}, which would not be adequate in the event of recessions because they would represent an excessive reactivity of the GDP growth rate which is not verified in the data. The second SADR2 model is:

$$\begin{cases} y_t = \alpha + \frac{1}{\gamma_0 + \gamma_1 y_{t-1}^{2r}} y_{t-1} + \varepsilon_t, \ \ \mathrm{t} \geq 2 \\ \varepsilon_t \sim IID\ N(0, \sigma) \end{cases} \tag{3.7}$$

where $\gamma_0 > 1$ and $\gamma_1 > 0$, whereas the statistical properties of ε_t and the vector of parameters are the same of the SDAR1 model. For this specification the same considerations about the persistence function apply. We can only observe that the maximum is $\frac{1}{\gamma_0}$.

4. Estimation

The empirical results relative to the parameter estimates of the models presented above are reported in tables 7-11 in the appendix. We use in-sample observations $(\boldsymbol{y_1}, \dots, \boldsymbol{y_n})$ from 1950.Q2 to 2017.Q3.

Within each class of models (AR, SETAR, AR-GARCH and SDAR) we select the best one following the AIC criterion. The AR lag order p is selected by fixing a maximum lag length. As reported in table 7 the AR(2) is the optimal linear model. The same procedure is used for SETAR models once the number of regimes is fixed. For two regimes, *d*=2, the lag orders $\boldsymbol{p_1}$and $\boldsymbol{p_2}$ may assume values from 1 to 5 and the selected model is that for which the pair $(\boldsymbol{p_1}, \boldsymbol{p_2})$ minimizes the AIC. A SETAR(2,1,1) is the selected model with two regimes (table 8). For three regimes, *d*=3, we have three autoregressive equations for each regime and the lag orders $\boldsymbol{p_1}, \boldsymbol{p_2}$ and $\boldsymbol{p_3}$ vary from 1 to 5. The selected model is that for which the vector $(\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3})$ minimizes the AIC. A SETAR(3,3,3,1) is the selected model with three regimes (table 9). We note that both autoregressive coefficients in the SETAR(2,1,1) model are positive but the persistence in more high in the upper regime than in the lower. On the other hand, in the SETAR(3,3,3,1) model the third autoregressive coefficient in the lower regime is negative denoting that economy reacts to recession rather quickly.

As for AR-GARCH models, in line with the number of lags estimated in the linear model, we select an AR(2)-GARCH(1,1) model after checking that a higher order in the AR component or in the conditional variance structure produced a lower AIC. The parameter estimates of this model are reported in table 10. Finally table 11 shows the results of estimating SDAR models for both specifications adopted, SDAR1 and SDAR2, depending of the chosen persistence function \$\psi\$. The estimation technique is the quasi-maximum likelihood (QML). In Gobbi and Mulinacci (2020) asymptotic properties of the QML estimator $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{0}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{1}}, \hat{\mathbf{r}}, \hat{\boldsymbol{\sigma}})$ of the vector of parameters$\boldsymbol{\vartheta}$are established. As the table clearly indicated, all parameters are highly significant. The nonlinear term, $\boldsymbol{\gamma_1}$, is widely higher than 1 for both specifications, indicating that the nonlinearity in time series is detected by both models. Some considerations about the estimated persistence functions $\hat{\boldsymbol{\varphi}}_{\mathbf{1}}(\mathbf{y_{t-1}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}_{\mathbf{0}}, \hat{\boldsymbol{\gamma}}_{\mathbf{1}}, \hat{\mathbf{r}}) = \mathbf{e}^{-(\hat{\gamma}_0 + \hat{\gamma}_1 y_{t-1}^{2\hat{r}})}$ and $\hat{\boldsymbol{\varphi}}_{\mathbf{2}}(\mathbf{y_{t-1}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}_{\mathbf{0}}, \hat{\boldsymbol{\gamma}}_{\mathbf{1}}, \hat{\mathbf{r}}) = \frac{\mathbf{1}}{\hat{\gamma}_0 + \hat{\gamma}_1 y_{t-1}^{2\hat{r}}}$ are needed. Figures 2 and 3 depict the dynamics of the US growth rate overlapped to the dynamics of estimated persistence functions $\hat{\boldsymbol{\varphi}}_{\mathbf{1}}$ (figure 2) and $\hat{\boldsymbol{\varphi}}_{\mathbf{2}}$ (figure 3).

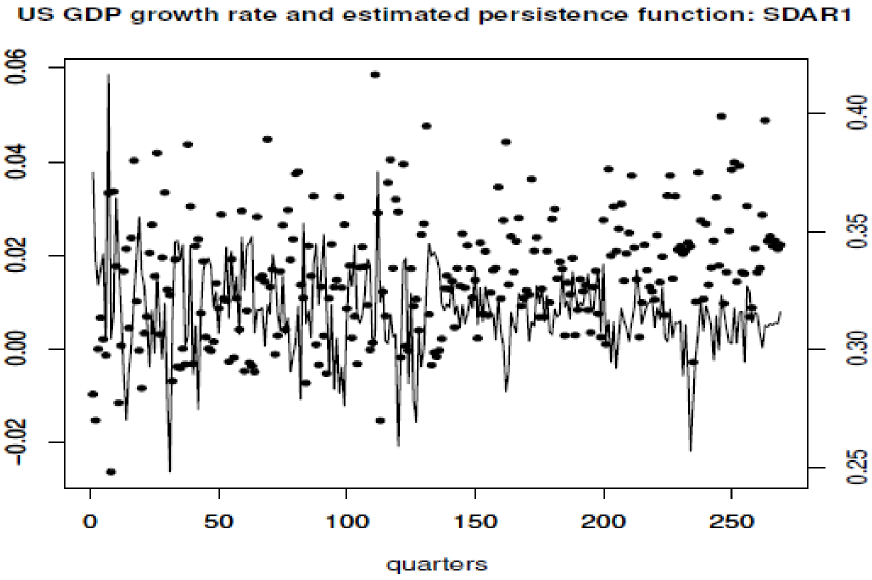

Figure 2: SDAR1 model. US GDP growth rate (line, left vertical axis) and estimated persistence function $e^{-(\hat{\gamma}_0+\hat{\gamma}_1 y_{t-1}^{2\hat{r}})}$ (points, right vertical axis)

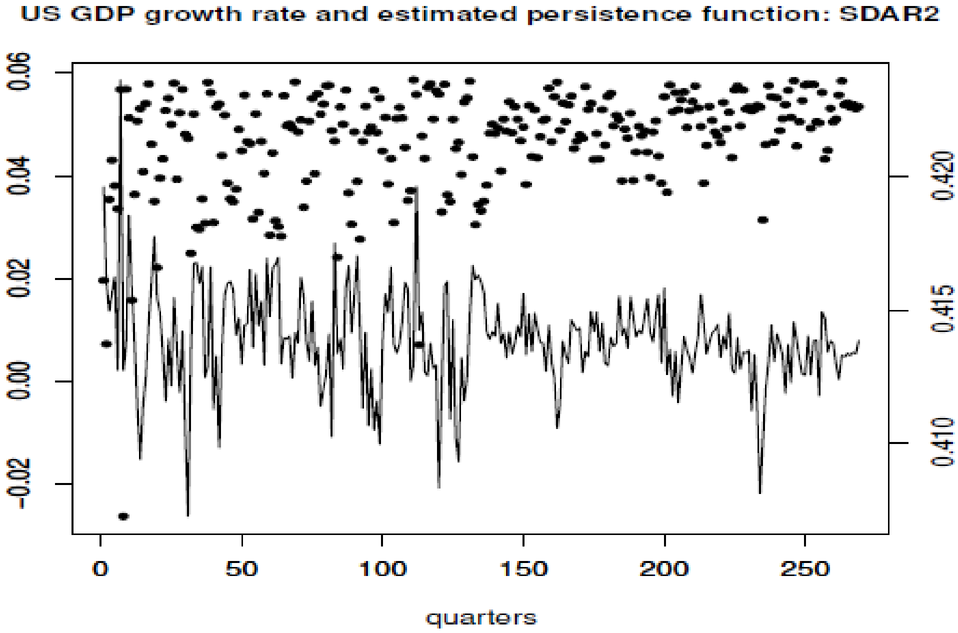

Figure 3: SDAR2 model. US GDP growth rate (line, left vertical axis) and estimated persistence function $\frac{1}{\hat{\gamma}_0+\hat{\gamma}_1 y_{t-1}^{2\hat{r}}}$ (points, right vertical axis)

We note that $\hat{\varphi}_1$ assumes on average lower values than $\hat{\varphi}_2$ which takes on a much narrower range of values. This means that the persistence estimated by the SDAR1 model is lower than estimated by the SDAR2 model. Indeed, the average value of $\hat{\varphi}_1$on the whole sample is 0.3311 whereas for the SDAR2 model the average value of $\hat{\varphi}_2$ is 0.4214. Furthermore, both functions are slightly increasing, in the sense that, on average, the persistence appears higher in the last period of the observed time series. This is confirmed for the SDAR1 model if we consider the average values of $\hat{\varphi}_1$and $\hat{\varphi}_2$ over time. In the first half of the sample (until the mid 80's) the average is 0.3252 while in the second half the average is 0.3368. On the other hand, for the SDRA2 model the dynamics is more volatile in the first half of the sample than in the second half where the value of the time-varying autoregressive coefficient rarely drops below 0.42. This behaviour appears negatively correlated with the dynamics of the conditional volatility estimated by the AR(2)-GARCH(1,1) model as reported in figure 4. It is clear that the volatility of the US GDP growth rate is significant higher in the first period of the sample with respect to the second period.

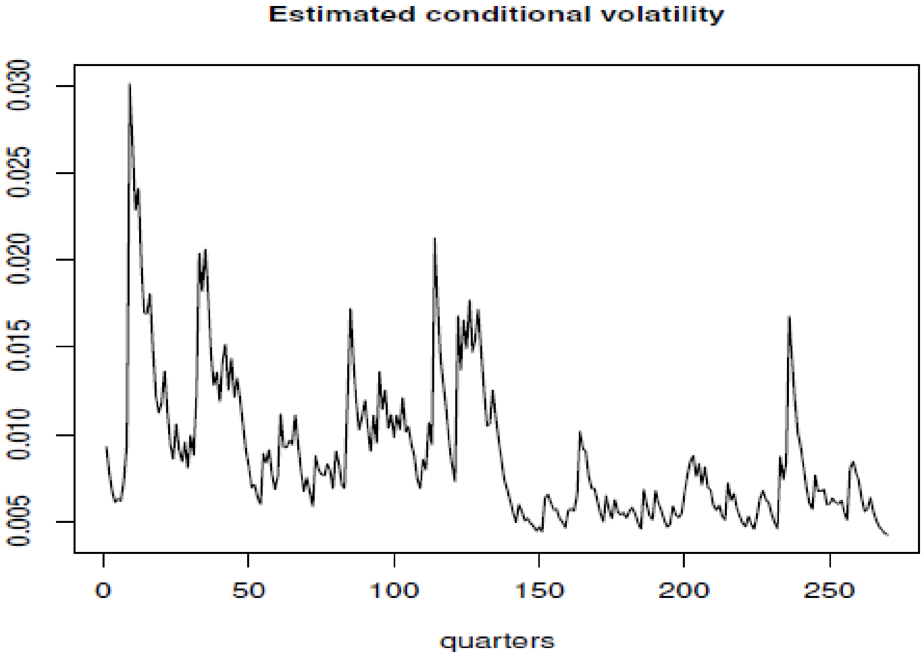

Figure 4: Estimated conditional volatility from the AR(2)-GARCH(1,1) model

To evaluate if the proposed models are well specified we consider the residuals diagnostics in table 3, which reports p-values associated to Ljung-Box and McLeod-Li tests, both up to 20 lags. The results reflect a good specification for all models under considerations since the hypothesis of absence serial autocorrelation of residuals and of squared residuals can be accepted. A separate consideration deserves SDAR models proposed in this paper. As expected, they seem more suitable for modeling the autocorrelation characterizing the time series of US GDP growth rate rather than the heteroskedasticity. Indeed, whereas the Ljung-Box test provides a strong evidence in favor of the absent of serial autocorrelation in the residuals, the p-value of the McLeod-Li test is rather low both for SDAR1 and SDAR2. In particular, in the case of the SDAR2 model an explanation can be found in the fact the estimated persistence function $\hat{\boldsymbol{\varphi}}_{\mathbf{2}}$ takes values in a much narrower range than $\hat{\boldsymbol{\varphi}}_{\mathbf{1}}$ highlighting a less adjustment to the time series dynamics. However, this seems to depend mainly on the nature of the data itself, since in the case of realized volatility the ML test result was more strong (Gobbi and Mulinacci, 2020).

Table 3: Model diagnostics.

	AR(2)	**SETAR(2,1,1)**	**SETAR(3,3,3,1)**	**AR(2)-GARCH(1,1)**	**SDAR1**	**SDAR2**
Res. autocorr.	0.8948	0.8238	0.9386	0.3815	0.4493	0.2039
Sq. Res. autocorr.	0.4138	0.6331	0.9792	0.9890	0.1294	0.0657

Note: Res. autocorr. and Sq. Res. autocorr. reports p-values of Ljung-Box and McLeod-Li tests of serial autocorrelation of fitted residuals and squared fitted residuals

5. Forecasting

We assess the forecast performance of each estimated model relative to linear AR(2) by means of Monte Carlo simulation. For SETAR models, Clements and Smith (1997) compare a number of alternative methods of obtaining multi-period forecasts and conclude that Monte Carlo method perform reasonably well. Forecasts from AR-GARCH model are obtained recursively from the variance equation. SDAR models generate forecasts by Monte Carlo simulation as proposed in Gobbi and Mulinacci (2020).

We use 8 values out-of-sample of US GDP growth rate from 2017.Q4 to 2019.Q3, denoted by $(\boldsymbol{y}_{\boldsymbol{n+h}})_{\boldsymbol{h=1,\ldots,H}}$where *n* is the last in-sample observation and *H*=1,...,8 is the forecast horizon. Let $(\tilde{\boldsymbol{y}}_{\boldsymbol{n+h}})_{\boldsymbol{h=1,\ldots,H}}$be the forecast values of the state variable generated from the models under consideration. Therefore, the forecast errors are given by$\boldsymbol{e}_{\boldsymbol{n+h}} = \boldsymbol{y}_{\boldsymbol{n+h}} - \tilde{\boldsymbol{y}}_{\boldsymbol{n+h}}$, with *h*=1,...,*H*. To compare the average accuracy of the forecasts we use two measures: the root mean square error (RMSE) and the mean absolute error (MAE). The RMSE is defined as $\sqrt{\frac{1}{H}\sum_{h=1}^{H} e_{n+h}^{2}}$whereas the MAE is given by $\frac{1}{H}\sum_{h=1}^{H} |e_{n+h}|$. Tables 4 and 5 and figures 5 and 6 summarize the results in terms of relative efficiencies for a forecast horizon from 1 to 8 quarters ahead. The relative efficiency (RE) is obtained as the ratio of the RMSE (or MAE) of the model under consideration and the RMSE (MAE) of the model used as benchmark, i.e., the linear AR(2). A value of RE greater or equal than unity indicates that the benchmark model provides more accurate forecast than the alternative nonlinear model. Form table 4 and figure 5 we deduce that, if the RE is measured in terms of the RMSE, only SDAR models offer a better performance than the linear AR(2) at least for the first 4 quarters. After this horizon the accuracy seems to be equivalent even if the linear AR(2) is slightly higher. On the other hand, the remaining alternative models tend to be worse but the SETAR(3,3,3,1) model is the only one to improve significantly over time until it become superior than the linear AR(2) for the last two forecast horizons. As regards SETAR models, Tiao and Tsay (1994) find that the forecasts obtained with this class of models are markedly superior than those obtained with the linear AR models if we only consider forecasts which are made when the economy is in the lower regime reflecting the ability of the SETAR models to capture the movements out of recession. In our observed time series the percentage of data belonging to the lower regime is of 36%, and this partly explains why on average the forecasts obtained by the SETAR model are lower.

The same considerations are strengthened if we consider the RE in terms of the MAD, as in table 5 and figure 6. In this case, both SDAR models provide a prediction with an accuracy higher than the benchmark for each forecast horizon and in the first four quarters (basically over the course of a year) the gain in the accuracy is considerable. Based on this measure we can conclude that there is an evidence that SDAR models has superior predictive ability compared to alternative models analyzed in this paper. These findings are not surprising if we consider the preliminary results on the sample. As shown in table 2, the evidence of nonlinearity is not strong, and in particular, this is confirmed and strengthened in the last ten years. The Tlrt test seems to exclude the presence of a threshold autoregressive structure in the last portion of the sample regardless of the lag considered. This can explain the relatively worse performance of SETAR models than the alternatives. SDAR models appear less conditioned by the kind and strength of the nonlinearity in the data.

Table 4: Relative efficiency of the forecasting accuracy measure RMSE with AR(2) model as benchmark.

Number of quarters ahead								
H	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SETAR(2,1,1)	2.7701	2.1335	2.4871	2.4472	1.2505	1.2512	1.2012	1.1271
SETAR(3,3,3,1)	2.0609	1.4182	2.1243	2.2162	1.0267	1.0716	0.9568	0.8736
AR(2)-GARCH(1,1)	1.0055	1.1190	1.0387	1.0954	1.0916	1.0971	1.1167	1.1297
SDAR1	0.4418	0.9250	0.8955	0.9125	1.0091	1.0071	1.0156	1.0041
SDAR2	0.6304	0.9509	0.9161	0.9671	1.0277	0.9973	1.0138	1.0313

Note: A value of the ratio lesser than 1 indicates that the nonlinear model ensures more accuracy than the AR(2) model

Table 5: Relative efficiency of the forecasting accuracy measure MAE with AR(2) model as benchmark.

Number of quarters ahead								
H	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SETAR(2,1,1)	2.7701	2.0271	2.4872	2.1588	1.4820	1.4833	1.3718	1.2552
SETAR(3,3,3,1)	2.0609	1.1984	1.8001	2.0992	1.2228	1.4214	1.1549	0.9791
AR(2)-GARCH(1,1)	1.0055	1.1044	0.9929	1.0945	1.0925	1.1457	1.1561	1.1614
SDAR1	0.4418	0.8623	0.8587	0.8365	0.9523	0.9630	0.9789	0.9924
SDAR2	0.6304	0.8938	0.8636	0.9346	0.9946	0.9719	0.9987	1.0166

Note: A value of the ratio lesser than 1 indicates that the nonlinear model ensures more accuracy than the AR(2) model

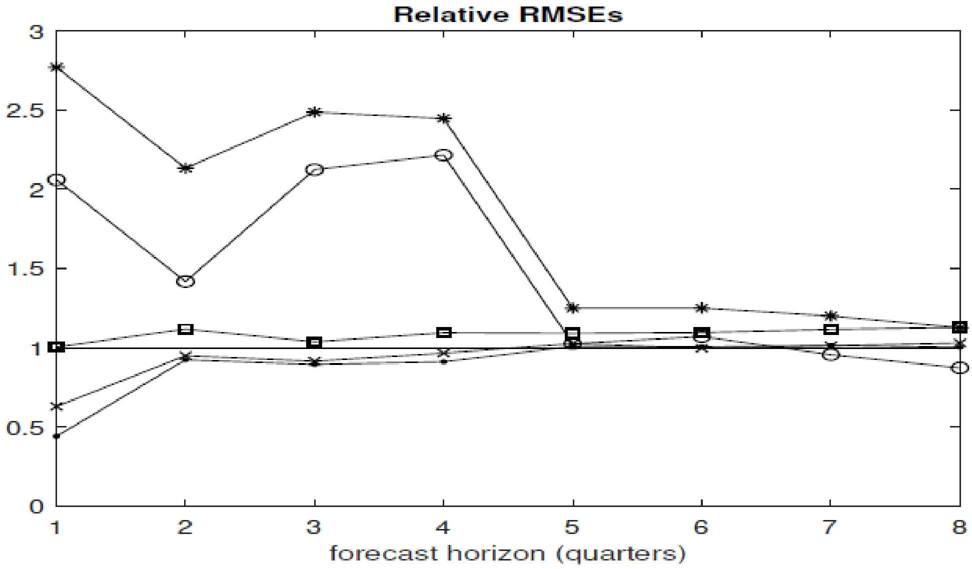

Figure 5: Relative RMSEs for each nonlinear model expressed in terms of that for AR(2). Legend: "star" for SETAR(2,1,1), "circle" for SETAR(3,3,3,1), "square" for AR(2)-GARCH(1,1), "point" for SDAR1 and "x" for SDAR2

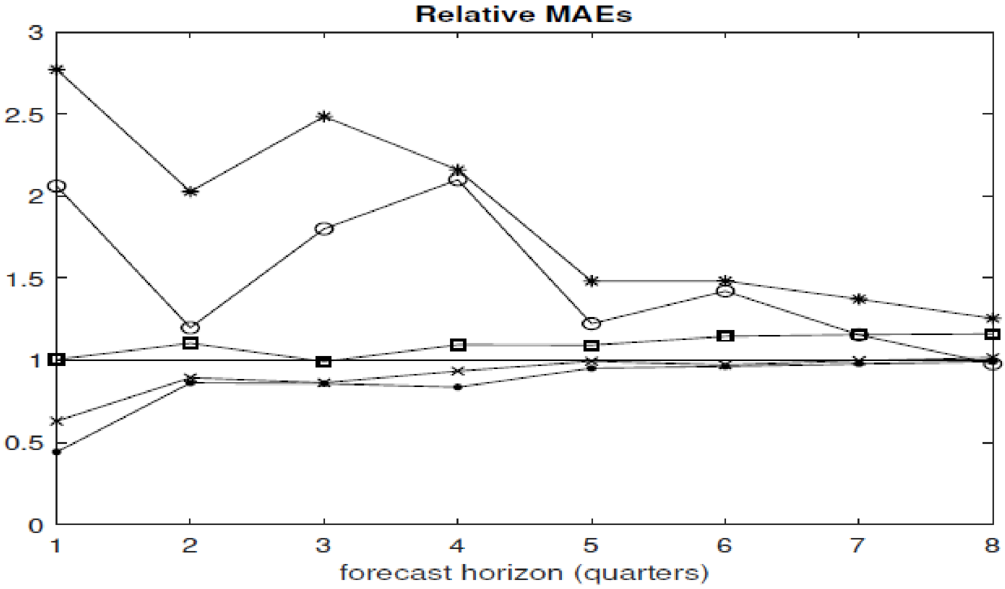

Figure 6: Relative MAEs for each nonlinear model expressed in terms of that for AR(2). Legend: "star" for SETAR(2,1,1), "circle" for SETAR(3,3,3,1), "square" for AR(2)-GARCH(1,1), "point" for SDAR1 and "x" for SDAR2

To further investigate the problem of forecasting accuracy of the models under study, we use the Diebold-Mariano (DM) test proposed by Diebold and Mariano (1995) and modified by Harvey, Leybourne and Newbold (1997) that corrects for the tendency of the test statistics to be over-sized. The null hypothesis of equal forecast accuracy between two methods is tested using a loss function $\boldsymbol{g(e^{i}_{n+h})}$, where $\boldsymbol{e^{i}_{n+h}}$ is the *h*-step ahead forecast error generated from model *i*, with *h*=1,...,*H*, and $\boldsymbol{g(x) = x^{2}}$. The loss differential for competitor forecasts *i* and *k* is defined as $\boldsymbol{d_{n+h} = g(e^{i}_{n+h})}$- $\boldsymbol{g(e^{k}_{n+h})}$, so that equal forecast accuracy entails $\boldsymbol{E[d_{n+h}] = 0}$. Diebold and Mariano (1995) derive the asymptotic distribution of the sample mean loss differential $\boldsymbol{\bar{d}}$. We perform the DM test considering the forecasts generated by the linear AR(2) model as the reference method (method 1) and the forecasts obtained from one of the alternative nonlinear model under study as the method 2. Then, the null is the hypothesis that method 1 and method 2 provide equal accuracy. The alternative can be specified in three different ways: "two-sided", method 1 and method 2 have different levels of accuracy, "greater", method 2 is more accurate than method 1, "less", method 2 is less accurate than method 1. With this construction in mind, we can evaluate the p-values of the test reported in table 6 for the first 4 forecast horizons. Differently than the case of relative efficiencies, the DM test does not provide a clear indication in favor of SDAR models. More in general, there is no evidence in favor of any nonlinear model under study. Indeed, all p-values are sufficiently high to induce to accept the null regardless of the alternative. This suggests that for the DM test both SDAR models are equivalent to the linear AR(2) as methods of forecasting. The same consideration holds for both SETAR models, whereas for the AR(2)-GARCH(1,1) model in two cases we reject the null when the alternative is of type "less", denoting that method 1 provided by the benchmark is preferred to method 2 provided by the AR(2)-GARCH(1,1).

In conclusion, we can assert that there is no tendency for SDAR models to be more accurate than the linear AR(2) model, even if it is the case to remark that the small number of observations available may influence the results of the test. On the other hand, Diebold (2015) argues that DM test is not intended for comparing models but is intended for comparing forecasts.

Table 6: p-values of the modified version of the Diebold-Mariano test proposed by Harvey, Leybourne and Newbold (1997)

Number of quarters ahead		**H=1**	**H=2**	**H=3**	**H=4**
	Alternative				
SETAR(2,1,1)	"two-sided" "greater" "lesser"	0.5342 0.7329 0.2671	0.6184 0.6908 0.3092	0.6954 0.6523 0.3477	0.7530 0.6248 0.3752
SETAR(3,3,3,1)	"two-sided" "greater" "lesser"	0.6808 0.3404 0.6596	0.6661 0.3333 0.6667	0.6859 0.3429 0.6572	0.8202 0.4101 0.5899
AR(2)-GARCH(1,1)	"two-sided" "greater" "lesser"	0.0527 0.9737 0.0262	0.0726 0.9635 0.0364	0.1337 0.9331 0.0668	0.2911 0.8541 0.1458
SDAR1	"two-sided" "greater" "lesser"	0.8862 0.5816 0.4181	0.7810 0.6075 0.3905	0.8691 0.5647 0.4316	0.9056 0.5472 0.4528
SDAR2	"two-sided" "greater" "lesser"	0.2788 0.8606 0.1394	0.2161 0.8919 0.1081	0.3206 0.8397 0.1641	0.5275 0.7363 0.2637

The null hypothesis is that the two methods of forecasting (method 1 and method 2) have the same forecast accuracy. Method 1 is the AR(2) model whereas method 2 is the contender nonlinear model. The p-values listed in column refers to three different alternatives: "two.sided", the alternative hypothesis is that AR(2) and method 2 have different levels of accuracy, "greater", the alternative hypothesis is that method 2 is more accurate than AR(2), "less", the alternative hypothesis is that method 2 is less accurate than AR(2)

6. Concluding Remarks and Future Investigations

With regard to measuring the forecast accuracy for the quarterly US GDP growth rate, we consider a state-dependent autoregressive (SDAR) model in which the autoregressive coefficient is a specified (nonlinear) function of the first lagged variable $\boldsymbol{y_{t-1}}$. The study is motivated by the conjecture that the model can yield a gain in forecasting accuracy at least for short or medium horizons as proved in Gobbi and Mulinacci (2020) for time series of weekly realized volatilities. The comparison is made with two different family of nonlinear models such as SETAR models (with 2 or 3 regimes) and AR-GARCH models. The forecast results are compared each other using a linear AR(2) model as benchmark. The evaluation of the results depends partly on the measure of forecasting accuracy adopted. Indeed, the relative RMSEs and MAEs provide a consistent evidence in favour of SDAR models compared with the competitors considered. The same evidence is not strongly confirmed by the Diebold-Mariano test, which shows results without a clear direction. In conclusion, the results encourage further investigations in order to explore the potentialities of this family of models. Three possible directions immediately come to mind. The first concerns different specifications of the persistence function in order to make more flexible the dynamics of the functional coefficient allowing to explore negative values. The second regards the lag order of SDAR models: classes of SDAR(p) models can be considered in which *p* persistence functions depend on the first *p* lags of the variable. The third direction is that to address the problem of alternative (asymmetric?) distributions of the error term.

References

[1] Blanchard, O. and Simon J. (2001). The long and large decline in U.S. output volatility, Brookings Papers and Economy Activity, 32, pp. 135-174.

[2] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroscedasticity, Journal of Econometrics, 31, pp. 307–327.

[3] Bollerslev, T., Chou, R.Y. and Kroner K.F. (1992). ARCH modelling in Finance: a review of the theory and empirical evidence, Journal of Econometrics, 52, pp. 5–39.

[4] Brockwell, P.J. and Davis, R.A. (1991). Time Series: Theory and Methods, Springer Series in Statistics, Springer-Verlag New York.

[5] Cai, Z., Fan, J. and Yao, Q. (2000). Functional-Coefficient Regressive Models for Nonlinear Time Series, J. Am. Statist. Assoc., 95(451), pp. 941-956.

[6] Chan, K. and Tsay, R. S. (1998). Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model, Biometrika, 85, pp. 413-426.

[7] Chan, K.S. (1990). Percentage points of likelihood ratio tests for threshold autoregression, Journal of Royal Statistical Society B, 53(3), pp. 691-696.

[8] Chen, R. and Liu. M.L. (2001). Functional coefficient autoregressive models: estimation and tests of hypotheses, Journal of Time Series Analysis, 22(2), pp. 151-173.

[9] Chen, R. and Tsay, R. (1993). Functional-coefficient autoregressive models, J. Am. Statist. Assoc., 88, pp. 298-308.

[10] Cherubini, U. and Gobbi, F. (2013). A Convolution-based Autoregressive Process, in F. Durante, W. Haerdle, P. Jaworski editors. Workshop on Copula in Mathematics and Quantitative Finance. Lecture Notes in Statistics-Proceedings. Springer, Berlin/Heidelberg.

[11] Cherubini, U., Gobbi, F. and Mulinacci, S. (2016). Convolution Copula Econometrics, SpringerBriefs in Statistics.

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

[13] Diebold, F.X. and Mariano, R.S. (1995). Comparing Predictive Accuracy, Journal of Business and Economic Statistics, 13, pp. 253–263.

[14] Diebold, F.X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests, Journal of Business & Economic Statistics, 33(1), DOI: 10.1080/07350015.2014.983236.

[15] Engle, R.F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica, 50, pp. 987–1008.

[16] Gobbi, F. (2020). The problem of detecting nonlinearity in time series generated by a state-dependent autoregressive model. A simulation study, to appear in Int. J. of Operational Research.

[17] Gobbi, F., Mulinacci, S. (2020). State-Dependent Autoregressive Models: Properties, Estimation and Forecasting, Available at http://arxiv.org/abs/2002.03134.

[18] Granger, C.W.J. and Terasvirta T. (1993). Modelling Nonlinear Economics Relationships, Oxford University Press, Oxford.

[19] Haggan, V. and Ozaki, T. (1981). Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model, Biometrica, 68(1), pp. 189-196.

[20] Hamilton, J.D. (1994). Time Series Analysis, Princeton: Princeton University Press.

[21] Hamori, S. (2000). Volatility of real GDP: Some evidence from the United States, the United Kingdom and Japan, Japan and the World Economy, 12, pp. 143-152.

[22] Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors, International Journal of Forecasting, 13(2), pp. 281-291.

[23] Hastie, T. and Tibshirani, R. (1990). Generalized additive models. Chapman & Hall, New York.

[24] Jarque, C. and Bera, A. (1980). Efficient tests for normality homoscedasticity and serial independence of regression residuals, Econometric Letters, 6, pp. 255–259.

[25] Jarque, C. and Bera A. (1987). A test for normality of observations and regression residuals, International Statistical Review, 55, pp.163–172.

[26] Lee, T.H., White, H. and Granger, C.W.J. (1993). Testing for neglected nonlinearity in time series models", Journal of Econometrics, 56, 269–290.

[27] Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models, Biometrika, 65, pp. 297–303.

[28] Kim, C.J. and Nelson, C.R. (1999). Has the U.S. economy become more stable? A Bayesian approach based on Markov-Switching model of the business cycle, Review of Economics and Statistics, 81, pp. 1-10.

[29] McConnel, M.M. and Perez-Quiros, G. (2000). Output fluctuations in the United States: What has changed since the early 1980s?, American Economic Review, 90, pp. 1464-1476.

[30] McLeod, A.I. and Li, W.K. (1983). Diagnostic checking ARMA time series models using squared residual autocorrelations, Journal of Time Series Analysis, 4, pp. 269-273.

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

[32] Terasvirta, T. (1994). Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models, Journal of the American Statistical Association, 89, pp. 208–218.

[33] Terasvirta, T. and Anderson, H.M. (1992). Characterizing Nonlinearities in Business Cycles with Smooth Transition Autoregressive Models, Journal of Applied Econometrics, 7, pp. 119-136.

[34] Terasvirta, T., Lin, C.F. and Granger, C.W.J. (1993). Power of the Neural Network Linearity Test, Journal of Time Series Analysis, 14, pp. 209–220.

[35] Tiao, G.C. and Tsay, R.S. (1994). Some advances in nonlinear and adaptive modelling time series, Journal of Forecasting, 13, pp. 109-131.

[36] Tong, H. (1978). On a threshold model, in Chen C.H. (ed.), Pattern Recognition and Signal Processing, 101-141. Amsterdam: Sijhoff and Noordoff.

[37] Tong, H. and Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data, Journal of the Royal Statistical Society, B 42, pp. 245-292.

[38] Tong, H. (1983). Threshold Models in Nonlinear Time Series Analysis, New York, Springer – Verlag.

[39] Tong, H. (1986). On estimating thresholds in autoregressive models, Journal of Time Series Analysis, 7, pp. 178-190.

[40] Tong, H. (1995). Nonlinear Time Series: A Dynamical System Approach, Oxford University Press.

[41] Tsay, R.S. (1986). Nonlinearity Tests for Time Series, Biometrika, 73, pp. 461-466.

Appendix: Estimation Results

In this appendix we report the parameter estimates and relative standard errors for each model we have considered.

Table 7: AR(2) model

	AR(2)	
Parameter	**Estimate**	**SE**
$\boldsymbol{\alpha}$	-0.0001744	0.000545596
$\boldsymbol{\theta_1}$	0.2424	0.05967707***
$\boldsymbol{\theta_2}$	0.1438	0.05912069***
Residuals variance	0.00008006	
AIC	-1765.6	

Table 8: SETAR(2,1,1) model

	SETAR(2,1,1)	
Parameter	**Estimate**	**SE**
$\boldsymbol{\alpha_1}$	0.00385287	0.00099631***
$\boldsymbol{\theta_{1,1}}$	0.16463935	0.08071829***
$\boldsymbol{\alpha_2}$	0.00643961	0.00102267***
$\boldsymbol{\theta_{2,1}}$	0.33423222	0.08150986***
Residuals variance	0.00007705	
AIC	-2547	

Table 9: SETAR(3,3,3,1) model

	SETAR(3,3,3,1)	
Parameter	**Estimate**	**SE**
$\boldsymbol{\alpha_1}$	0.00543833	0.00192818***
$\boldsymbol{\theta_{1,1}}$	0.22227239	0.13149926**
$\boldsymbol{\theta_{1,2}}$	0.04383025	0.20883351
$\boldsymbol{\theta_{1,3}}$	-0.24087024	0.13134799**
$\boldsymbol{\alpha_2}$	-0.00015534	0.00428223
$\boldsymbol{\theta_{2,1}}$	0.06276289	0.11205026
$\boldsymbol{\theta_{2,2}}$	0.25931445	1.03329799
$\boldsymbol{\theta_{2,3}}$	0.37828874	0.12097545***
$\boldsymbol{\alpha_3}$	0.00644022	0.00101855***
$\boldsymbol{\theta_{3,1}}$	0.33414492	0.08318353***
Residuals variance	0.00007303	
AIC	-2548	

Table 10: AR(2)-GARCH(1,1) model

	AR(2)-GARCH(1,1)	
Parameter	**Estimate**	**SE**
$\boldsymbol{\alpha}$	0.003830	0.000736***
$\boldsymbol{\theta_1}$	0.27930	0.06937***
$\boldsymbol{\theta_2}$	0.26041	0.06320***
$\boldsymbol{\omega_0}$	0.00000	0.000000*
$\boldsymbol{\omega_1}$	0.35453	0.09174***
$\boldsymbol{\omega_2}$	0.64672	0.08248***
Residuals variance	0.00008112	
AIC	-6.745015	

Table 11: SDAR models

	SDAR1		**SDAR2**	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**
$\boldsymbol{\alpha}$	0.00501	0.00063***	0.00457	0.00055***
$\boldsymbol{\gamma_0}$	0.66578	0.02678***	2.36048	0.01604***
$\boldsymbol{\gamma_1}$	1.73221	0.01381***	2.69493	0.07033***
$\boldsymbol{r}$	0.19363	0.00791***	0.58967	0.00117***
$\boldsymbol{\sigma}$	0.01017	0.00154***	0.00965	0.00165***
Residuals variance	0.00008461		0.00008603	
AIC	-1747.372		-1746.778	