



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computational Statistics & Data Analysis 47 (2004) 277–295

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

www.elsevier.com/locate/csda

Fitting piecewise linear threshold autoregressive models by means of genetic algorithms[☆]

R. Baragona^{a,*}, F. Battaglia^b, D. Cucina^b

^a*Dipartimento di Sociologia e Comunicazione, Università di Roma La Sapienza, Via Salaria 113, I-00198 Roma, Italy*

^b*Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma La Sapienza, Piazzale Aldo Moro 5, I-00100 Roma, Italy*

Received 6 November 2003; accepted 8 November 2003

Abstract

A nonlinear version of the threshold autoregressive model for time series is introduced. A peculiar requirement on parameters, except possibly for the constant term, is the continuity, that seems a natural and useful assumption. This model is a special case of the general state-dependent models, where the moving-average term is dropped and a particular form for the dependence on the state is specified. Such model meets also the functional autoregressive model formulation, but the “least demanding” functional form is assumed. Further restrictive assumptions are not needed. Both identification and estimation problems will be taken into account. The proposed approach brings together the genetic algorithm, in its simplest binary form, and some basic features from spline theory. It results in a powerful flexible tool which is shown to be able to approximate a wide class of nonlinear time series models. This method is found to compare favorably with existing procedures in modeling some well-known real-time series, which often are taken as a benchmark for testing and comparing modeling procedures.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Time series; Self-exciting threshold autoregressive models; Genetic algorithms; Splines; State dependent models; Canadian lynx data; Sunspot numbers; Blowfly population data

[☆] This work was financially supported by grants from MIUR, Italy, and CNR, Italy. We wrote some Fortran programs for implementing our procedures. Our source code is freely available from the authors on request.

* Corresponding author. Tel.: +39-06-49918447; fax: +39-06-8419505.

E-mail address: roberto.baragona@uniroma1.it (R. Baragona).

1. Introduction

A wide class of nonlinear models has been developed where the state of the dynamic system determines the “regime” which drives the time series. For instance, the self-exciting threshold autoregressive (SETAR) models (see Tong, 1983, 1990, for instance) are based on the assumption that the time series is generated by several alternative linear autoregressive (AR) models according to the values assumed by some past observations. The sample space is split into regions delimited by assigned borders, or “thresholds”. Each region is associated with an AR model. Then, if the past observation is included in the i th region the i th AR model generates the next time series value.

The notion of a threshold has been extended to the case where the AR parameters are linear functions of the lagged values of the time series (see Ozaki, 1981; Tong, 1980). It seems quite natural to link this extension to the class of the functional autoregressive (FAR) models (Chen and Tsay, 1993; Cai et al., 2000), and to the even more general state-dependent models (SDM) introduced by Priestley (1988). The SDM and the FAR models, however, cope with fairly general coefficients behavior. Nonparametric methods or the extended Kalman filter have been proposed for parameters estimation. Though useful and flexible, neither provides a simple formula for representing the behavior of the estimated parameters. It is sometimes preferable to have a closed functional form available, which is easy to both estimate and interpret. The simplest function we may use is a first-order polynomial, and make it to depend on a single lagged value of the time series itself. If we allow the coefficients to follow a piecewise linear curve, then we may obtain directly a function which yields an accurate picture of the behavior of each of the AR coefficients. With limited additional computational effort, the continuity requirement may be fulfilled, as a piecewise linear continuous function is often a good approximation for most analytic functions. Furthermore, increasing the number of regimes may constitute a convenient way for approximating more complicated functions, even if these latter may not be put in closed form. Let us call piecewise linear SETAR (PLTAR) this class of models.

In order that a procedure for identifying and estimating a PLTAR model be really feasible, however, a fast and efficient algorithm is needed for locating the threshold constants. In fact, it is likely that several values for both the delay parameter and the model order have to be tried, and searching for the threshold constants has to be routinely performed. The routine that we propose is based on the genetic algorithms (GA) introduced by Holland (1975). The GAs are known to be able to provide us with a powerful optimization tool when the solution space happens to be both discrete and large, and the objective function does not fulfill the usual regularity requirements, such as continuity, differentiability and convexity (see, for instance, Michalewicz, 1996; Man et al., 1999). In addition, the GAs have been proposed by Wu and Chang (2002) for estimating SETAR models, and by Pittman and Murthy (2000) for fitting piecewise linear functions. So, our proposal seems to be adequately motivated for dealing with PLTAR models by means of GAs. However, we have to design a special procedure to take the PLTAR model features into account properly. We distinguish between the identification step and the estimation step. The former is a combinatorial-like problem

which involves only discrete numbers, that is the delay parameter, the AR order, the number and location of the threshold constants. All these are called the “structural parameters”. Unless the size of the problem is small, the candidate solutions cannot be enumerated and checked within a reasonable amount of time. On the other hand, the latter step, estimation, may be performed without too much effort, conditional on the “structural parameters” values, using standard statistical methods.

The paper is organized as follows. In Section 2 the PLTAR model is presented in some detail. In Section 3 basic spline theory is used to arrange the estimation task so that the continuity requirement is fulfilled. In Section 4 the GA for finding appropriate structural parameters is explained. In Section 5 results are reported concerning the identification and estimation of PLTAR model for some artificial time series. In Section 6 three well-known real-data sets are considered: the Canadian lynx data, the sunspot numbers and the blowfly population data. In Section 7 conclusions are drawn.

2. The PLTAR model

A SDM for the time series $\{y_t\}$, integer t , takes the form (Priestley, 1988)

$$y_t = \sum_{j=1}^p \phi_j(z_{t-1})y_{t-j} + e_t - \sum_{j=1}^q \theta_j(z_{t-1})e_{t-j}, \tag{1}$$

where the array

$$z_{t-1} = (e_{t-q}, \dots, e_{t-1}, y_{t-p}, \dots, y_{t-1})$$

is called the state vector at the time $t - 1$.

Many useful models are obtained by assuming some hypotheses on the functional form of (1). For instance, a special case of the SDM is a SETAR model where it is assumed that the functions $\theta_j(\cdot)$'s are all zero, the state vector is $z_{t-1} = y_{t-d}$ for some integer d (the delay parameter), and the functions $\phi_j(\cdot)$'s may be written as

$$\phi_j(y_{t-d}) = c_j^{(i)} \quad \text{if } r_{i-1} < y_{t-d} \leq r_i, \quad i = 1, \dots, k \tag{2}$$

for $j = 1, \dots, p$. The integer p has to be pre-specified properly. The disjoint intervals (r_{i-1}, r_i) partition the real axis, as we assume that $r_0 = -\infty$ and $r_k = +\infty$.

The delay parameter d , the number of regimes k , the threshold constants r_i and the AR model order p are called the “structural parameters”. Once such parameters are determined, the AR coefficients $c_j^{(i)}$ may be estimated by using standard least squares.

We consider in this paper the PLTAR model where the terms $c_j^{(i)}$'s are themselves linear functions of the state y_{t-d} , so that equalities (2), for $j = 1, \dots, p$, generalize to

$$\phi_j(y_{t-d}) = \alpha_j^{(i)} + \beta_j^{(i)} y_{t-d} \quad \text{if } r_{i-1} < y_{t-d} \leq r_i, \quad i = 1, \dots, k. \tag{3}$$

The coefficients $\phi_j(\cdot)$'s are assumed continuous functions of the state y_{t-d} . This assumption seems of interest, because a piecewise linear function may satisfactorily approximate most analytic functions. Moreover, as Priestley (1988, p. 102) observes, “in practice the threshold model is essentially a device for describing a continuous

nonlinear relationship by a step-function approximation”; on the other hand “we can always approximate a step function by a continuous function with large (but finite) gradients”.

The piecewise linearity requirement may be easily incorporated in the model by means of a simple re-parameterization similar to that used for spline functions. Let us define

$$S_1(u) = u, \quad S_i(u) = \begin{cases} 0 & \text{if } u \leq r_{i-1}, \\ u - r_{i-1} & \text{if } u > r_{i-1}, \end{cases} \quad i = 2, \dots, k.$$

Then we may write

$$\phi_j(y_{t-d}) = \lambda_j + \sum_{i=1}^k v_j^{(i)} S_i(y_{t-d}), \quad (4)$$

where the correspondence with the previously used parameters $\alpha_j^{(i)}$'s and $\beta_j^{(i)}$'s in (3) is easily obtained

$$\alpha_j^{(i)} = \lambda_j - \sum_{s=1}^i v_j^{(s)} r_{s-1}, \quad \beta_j^{(i)} = \sum_{s=1}^i v_j^{(s)}$$

and, by recalling (4), the complete model may be written as

$$y_t = \sum_{j=1}^p \lambda_j y_{t-j} + \sum_{j=1}^p \sum_{i=1}^k v_j^{(i)} S_i(y_{t-d}) y_{t-j} + e_t \quad (5)$$

and is characterized by a linear ordinary AR part and a nonlinear AR scheme.

If nonzero mean series have to be modeled, a constant term may be added in (5). Such a constant cannot depend linearly on y_{t-d} (since in that case it would be indistinguishable from an ordinary AR parameter at lag d) but may be possibly chosen different at each regime (thus, the mean depends on the regime). The complete model is

$$y_t = c(y_{t-d}) + \sum_{j=1}^p \lambda_j y_{t-j} + \sum_{j=1}^p \sum_{i=1}^k v_j^{(i)} S_i(y_{t-d}) y_{t-j} + e_t, \quad (6)$$

$$c(y_{t-d}) = c_i \quad \text{if } r_{i-1} < y_{t-d} \leq r_i, \quad i = 1, \dots, k.$$

3. Estimating the PLTAR model

We suppose that a series $\{y_1, y_2, \dots, y_n\}$ is observed. Given the “structural parameters”, that is the delay parameter d , number of regimes k , threshold constants r_i , $i = 0, 1, \dots, k$, and AR order p , the autoregressive parameters of model (6) may be estimated by ordinary least squares, by minimizing, with respect to λ_j , $v_j^{(i)}$ and the constant terms c_i 's the sum of squares

$$\text{SSQ} = \sum_{t=m+1}^n \left\{ y_t - c(y_{t-d}) - \sum_{j=1}^p \lambda_j y_{t-j} - \sum_{j=1}^p \sum_{i=1}^k v_j^{(i)} S_i(y_{t-d}) y_{t-j} \right\}^2. \quad (7)$$

The residual sum of squares (SSQ) involves all the observations starting from $m + 1$, where $m = \max(p, d)$, to n , and contains $k + p + pk$ unknown parameters, or $p(k + 1)$ if constants are not included in the model. A major difference with the SETAR model is that minimization cannot be done separately, for observations belonging to each regime, because SSQ (7) does not split into terms which include only some parameters subset.

4. Identifying the PLTAR model

The main interest here is in proposing an effective method for the identification task, that is determining the “structural parameters”. We are faced with the problem of simultaneously finding, according to some optimization criterion, the number of regimes, the threshold constants, the delay parameter, and the autoregressive order. We noted already that this is a kind of combinatorial problem that, unless its dimension is lowest, typically requires some heuristic algorithm to be employed. We propose the GA approach which allows us to try several proposal solutions and seek for the one that maximizes a properly chosen optimization criterion. Details about the effectiveness of the GA as optimization tool may be found in [Jennison and Sheehan \(1995\)](#).

We developed our GA along the following guidelines.

- (i) The algorithm is replicated over a reasonable range of values $d = d_1, \dots, d_D$. Therefore, we shall build the best model for any choice of d , and the choice (the comparison) will be left to the analyst.
- (ii) The GA is given the task of finding both the number of regimes k and the threshold constants r_1, \dots, r_{k-1} .
- (iii) The common AR model order p is determined by using the AIC criterion.

The AIC (see [Akaike, 1977](#)) is a well-known criterion which may serve as the basis of the order selection procedure. Its application to order selection for nonlinear models is discussed in [Ozaki and Oda \(1978\)](#) and [Ozaki \(1982\)](#). Though it is documented in the literature that the AIC is not consistent in choosing the right order of the model, in the present context we may use it because we have only to compare models in the presence of a finite number of observations. Alternative automatic classification criteria may be employed as well (see, for instance, [Fuller, 1996](#), p. 437–439, and references therein).

We let $r_0 = \min\{y_1, \dots, y_n\}$, $r_k = \max\{y_1, \dots, y_n\}$, and the other threshold values will be selected inside the set of the time series values. This does not imply any loss of generality, because the inequality $y_{t-d} \leq r_i$, say, may well be replaced by $y_{t-d} \leq y_{t_j}$, where y_{t_j} is the greatest observations less than or equal to r_i . The steps of the present GA are structured according to the procedure presented in [Goldberg \(1989\)](#) as simple GA. A maximum number of regimes is pre-specified, K say. Let $\{y_{t_1}, y_{t_2}, \dots, y_{t_n}\}$ be the set of the time series values arranged in nondecreasing order. A tentative solution is represented by a binary array of length n . Each bit corresponds to a time point t_j , and y_{t_j} is the observed value of the series at time t_j . Let $\mathbf{x} = (x_1, \dots, x_n)$ be such array. If $x_j = 1$, then y_{t_j} is a threshold constant, while $x_j = 0$ otherwise. Since $y_{t_1} = r_0$

and $y_{t_n} = r_k$, we have $x_1 = x_n = 1$, and the solution is essentially characterized by the string $(x_2, x_3, \dots, x_{n-1})$. The number of regimes is given by $x_2 + x_3 + \dots + x_{n-1} + 1$: a string is not admissible if the latter sum exceeds K , or some regimes contain too few observations. A pre-specified number of admissible strings, s say, are generated at random, and form the set of the initial tentative solutions. For each given string \mathbf{x} , and for each p in a range $[1, P]$, the parameters are estimated according to the previous section, and the AIC is computed:

$$\text{AIC}(p) = (n - m) \log \hat{\sigma}_p^2 + 2m,$$

where $m = k + p + pk$ if the constant is included, whilst $m = p(k + 1)$ otherwise, and $\hat{\sigma}_p^2$ is the minimum value attained in (7) divided by $n - m$. If

$$\text{AIC}(p^*) = \min_{1 \leq p \leq P} \text{AIC}(p)$$

we select order p^* .

The evaluation of a string within the GA framework is done by means of a positive real-valued function called “fitness function” which measures goodness of solution. Let us define our fitness function

$$f(\mathbf{x}) = \exp(-\text{AIC}(p^*)/C), \quad (8)$$

where C is a problem-dependent constant which is introduced to prevent the occurrence of overflow in the computation and to scale fitness suitably.

The set of initial solutions is manipulated by means of the so-called “evolutionary operators”. Many have been proposed, but, according to the simple GA procedure, we consider only selection, crossover and mutation. These three operators modify the solutions and produce a new set with the aim of increasing the average fitness. The procedure is iterative, and in each step a new set of solutions is generated from the previous one. Any set of solutions is called “population”, and the iterative procedure mimics the evolution of a biological population through a sequence of generations in such a way that more and more tentative solutions (the “individuals” in the populations) approach the global optimum. The adoption of the so-called “elitist strategy” is recommended as well. Such device consists in retaining in a special location the best string of each generation, so that it cannot be destroyed by the mutation or crossover operators, and it survives to the next population certainly. The procedure stops as soon as the maximum pre-specified number of generations, N say, is attained, or some stopping criterion is met.

The three evolutionary operators are as follows.

Selection: For s times, a string is drawn from the current population with probability proportional to its fitness. The new strings replace the population. This device is often referred to as “roulette wheel rule”.

Crossover: We adopt the simplest form, that is the single point crossover. We consider $[s/2]$ string pairs, where the two strings in each pair are chosen at random, and operate a crossover for each of them with pre-specified probability p_c . If crossover is to be performed, a positive integer, called the “cutting point”, is chosen uniformly randomly in the range from 2 to $n - 2$. Let ℓ be the cutting point, and the pair

of strings be

$$\begin{aligned} \mathbf{x}_a &= (x_1^a, \dots, x_{\ell}^a, x_{\ell+1}^a, \dots, x_n^a), \\ \mathbf{x}_b &= (x_1^b, \dots, x_{\ell}^b, x_{\ell+1}^b, \dots, x_n^b). \end{aligned}$$

Two new strings are generated by exchanging the bits on the right side of the cutting point between the two current strings. The new strings are

$$\begin{aligned} \mathbf{x}_a &= (x_1^a, \dots, x_{\ell}^a, x_{\ell+1}^b, \dots, x_n^b), \\ \mathbf{x}_b &= (x_1^b, \dots, x_{\ell}^b, x_{\ell+1}^a, \dots, x_n^a). \end{aligned}$$

and are taken to replace the old ones.

Mutation: Any bit $\{x_j, j=2, \dots, n-1\}$ of any string is allowed to flip with probability p_m , usually quite small.

The three operators are designed for different tasks. Selection makes the best fit individuals to spread in the next generations. Crossover combines promising solutions, as they come from the selection step, to put together blocks that are themselves parts of “high quality” solutions. Mutation maintains diversity in the population, and possibly recovers bits that would be impossible to create by means of the other two operators.

We stress that a similar procedure may be applied for identification of the SETAR models. Only a simple modification is needed concerned with the fitness function (8), by substituting to (7) the appropriate residual sum of squares of the SETAR model. In comparison to Wu and Chang (2002), our procedure is designed for obtaining the best model for any choice of the delay parameter d and number of regimes k , and allows selection of the values of the threshold constants r_i in the entire set of time series values.

5. Simulation results

Our simulation study is aimed at checking two properties of our model building procedure. The first one is its ability to identify and estimate the parameters of artificial time series that are generated by a PLTAR model. The second one is the ability of our procedure to yield an accurate description of artificial time series that are generated by some other nonlinear models such as SETAR and EXPAR.

Let, for instance, the time series $\{y_t\}$ be generated according to the PLTAR model

$$y_t = \begin{cases} -0.8y_{t-1} + e_t & \text{if } y_{t-1} \leq -1, \\ (0.6 + 1.4y_{t-1})y_{t-1} + e_t & \text{if } -1 < y_{t-1} \leq 0, \\ (0.6 - 1.4y_{t-1})y_{t-1} + e_t & \text{if } 0 < y_{t-1} \leq 1, \\ -0.8y_{t-1} + e_t & \text{if } y_{t-1} > 1, \end{cases} \quad (9)$$

where $\{e_t\}$ is a Gaussian white noise with zero mean and unit variance. The delay parameter is $d=1$, the number of regimes is $k=4$ and the AR order is $p=1$. There are

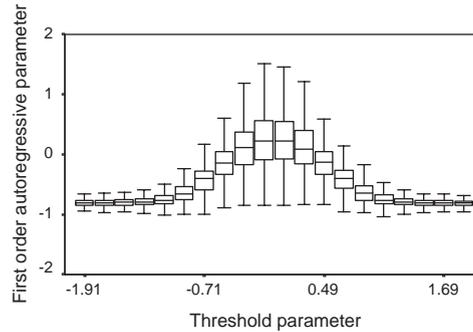


Fig. 1. Boxplot of the distributions of the estimated parameter for time series generated by a PLTAR: 1000 series of 500 observations.

no constant terms. Models similar to (9) have been proposed by Ozaki (1981, 1982) and Priestley (1988) amongst others. The routines RANDOM and PPND16 (Applied Statistics Algorithms) have been used for generating a stretch of 2000 Gaussian standard random numbers $\{e_t\}$. Then, model (9) has been used to provide us with the time series $\{y_t\}$. We have retained the last 500 observations and discarded the initial 1500. Discarding so many values is needed for eliminating the transient effect of the starting values in threshold type models, as noted in Chan and Cheung (1994), for instance. This procedure has been replicated 1000 times, so that we have 1000 series with 500 observations each. Also, 1000 series with 1000 observations each have been generated as described before. In the GA a different generator was used for providing uniform random numbers in the interval (0,1), that is the UNIFORM routine. We made a rather standard choice for the GA parameters, that is the population size was taken $s=30$, the crossover and mutation probabilities were 0.8 and 0.001, respectively, and the number of generations was 1000. For the parameters choice in optimization problems see, for instance, De Jong (1975), Mitchell (1996), p. 175, Haupt and Haupt (1998), p. 113, Chatterjee et al. (1996) and Chatterjee and Laudato (1997). The number of regimes was allowed to vary from 1 to 5, and the AR order from 1 to 5 as well. All delay parameter values were tried from 1 to $d_D = 5$. In each regime, at least 40 observations were required. The constant C we considered appropriate to compute the fitness function (8) was 100.

The results obtained for the time series generated by the model (9) may be summarized in Figs. 1 and 2 for 500 observations and in Figs. 3 and 4 for 1000 observations. In both cases the estimated parameter curve reproduces the true parameter curve fairly well on the average. Boxplots show that the standard error of the estimates is small when the autoregressive parameter is relatively large, while it increases when the parameter is smaller in modulus, or more variable with respect to y_{t-1} . In the latter case, estimates may be biased (towards zero). Nevertheless, the estimation procedure based on the GA is able to yield model estimates such that the predicted (one-step-ahead) time series, over 1000 replications, are close to the artificial time series.

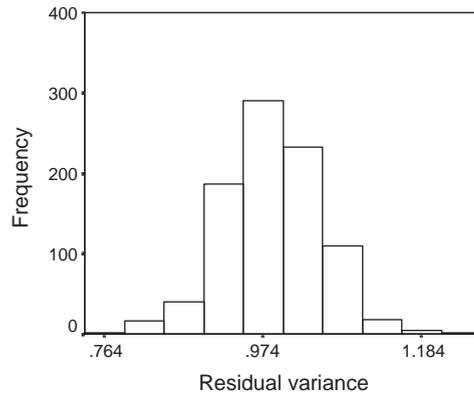


Fig. 2. Histogram of the residual variance for time series generated by a PLTAR: 1000 series of 500 observations.

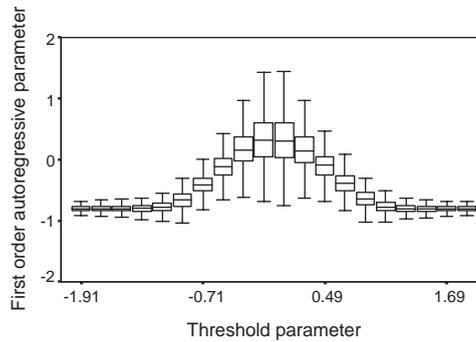


Fig. 3. Boxplot of the distributions of the estimated parameter for time series generated by a PLTAR: 1000 series of 1000 observations.

In the second simulation experiment we consider one of the many threshold models studied in Chan and Cheung (1994):

$$y_t = \begin{cases} 0.9y_{t-1} + e_t & \text{if } y_{t-1} \leq 0, \\ -0.1y_{t-1} + e_t & \text{if } y_{t-1} > 0, \end{cases}$$

where $\{e_t\}$ has zero mean and $\sigma_e^2 = 1$. For each of 1000 replications we generated 2000 observations, and retained the last 500. Also, 1000 artificial time series of 1000 observations were generated. We estimated on these artificial series a SETAR model according to the procedure reported in Tong (1990), a SETAR model by using the GA and a PLTAR model by means of our GA-based procedure. Distributions of the parameter estimated curve yielded by each of the three algorithms, along with the respective residual variance distribution, are reported for 500 observations in Figs. 5–10. In all cases the adherence is good for values of the state y_{t-1} away

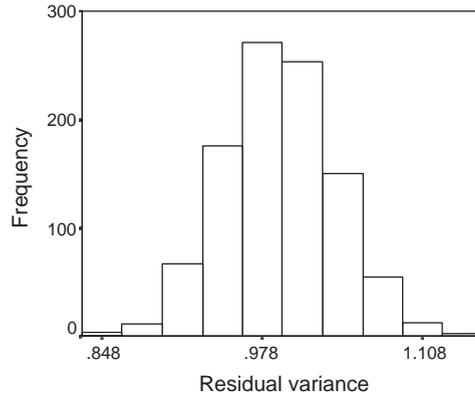


Fig. 4. Histogram of the residual variance for time series generated by a PLTAR: 1000 series of 1000 observations.

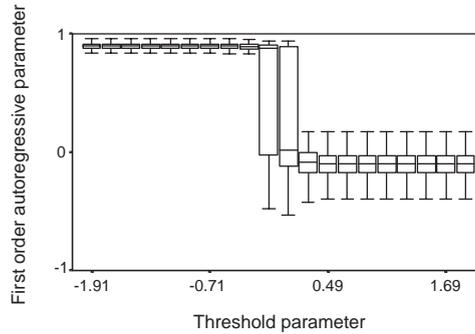


Fig. 5. Boxplot of the distributions of the estimated parameter (SETAR model, first method): 1000 series of 500 observations.

from zero, since bias is negligible, while slight differences arise if y_{t-1} is near zero. The residual variance with the GA-based method is however less variable and more centered on the true value 1. Furthermore, it may be seen that the step in the parameter is adequately approximated by the PLTAR model. Similar results were obtained for 1000 observations.

The third simulation experiment is concerned with estimating a PLTAR model for the time series generated by the exponential autoregressive (EXPAR) model (Haggan and Ozaki, 1981)

$$y_t = \{1.95 + 0.23 \exp(-y_{t-1}^2)\}y_{t-1} + \{-0.96 - 0.24 \exp(-y_{t-1}^2)\}y_{t-2} + e_t, \quad (10)$$

where $\{e_t\}$ is a Gaussian white noise with zero mean and unit variance. For 1000 replications we generated 2000 observations and discarded the first 1500. This way 1000 series of 500 observations were available. Also for 1000 replications series of 1000 observations were generated. We adapted to such artificial time series a PLTAR

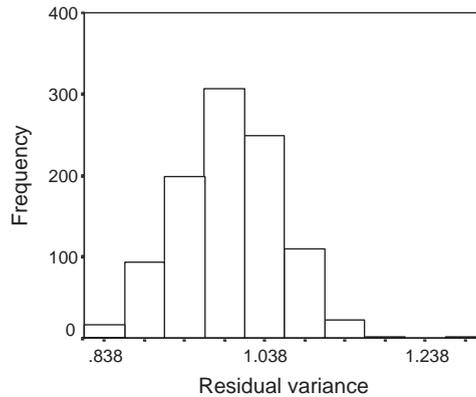


Fig. 6. Histogram of the residual variance (SETAR model, first method): 1000 series of 500 observations.

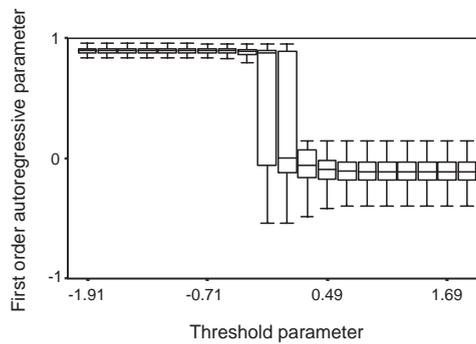


Fig. 7. Boxplot of the distributions of the estimated parameter (SETAR model, GA method): 1000 series of 500 observations.

model and used our procedure based on the GA. The estimated PLTAR model was found able to represent fairly well the artificial time series $\{y_t\}$ though this latter was generated by a different model. For 500 observations, the average residual variance was 0.93 (standard error of the estimate 0.07) by the PLTAR model and 1.48 (standard error 0.83) by estimating the correctly identified EXPAR model (10). For 1000 observations we obtained similar results.

We considered as well the model fitted in Ozaki (1982) to the Canadian lynx data and used in Cai et al. (2000) for simulation purpose. We generated 1000 series of both 500 and 1000 observations from the model

$$y_t = a_1(y_{t-1})y_{t-1} + a_2(y_{t-1})y_{t-2} + e_t, \tag{11}$$

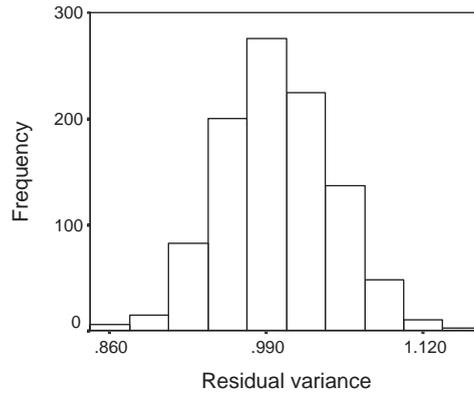


Fig. 8. Histogram of the residual variance (SETAR model, GA method): 1000 series of 500 observations.

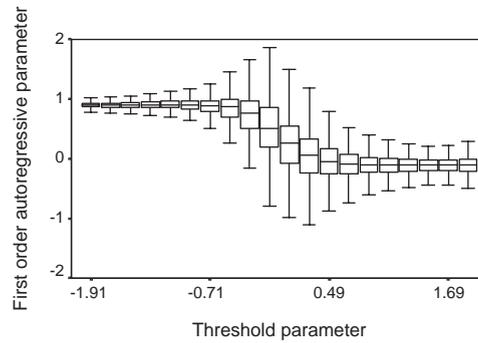


Fig. 9. Boxplot of the distributions of the estimated parameter (PLTAR model): 1000 series of 500 observations.

where

$$a_1(y_{t-1}) = 0.138 + (0.316 + 0.982y_{t-1}) \exp(-3.89y_{t-1}^2),$$

$$a_2(y_{t-1}) = -0.437 - (0.659 + 1.26y_{t-1}) \exp(-3.89y_{t-1}^2)$$

and $\{e_t\}$ is a stretch of independent identically distributed Gaussian variates with mean zero and variance 1. A PLTAR model was fitted to these artificial time series by applying our GA procedure. The average residual variance was 0.95 (0.07) for 500 observations. Fitting to the data the correct model (11) by least squares, we obtained the average residual variance equal to 0.99 (0.07). This means that the rather complicated parameters of model (11) are well approximated by the piecewise linear continuous functions. Similar results we obtained for 1000 observations.

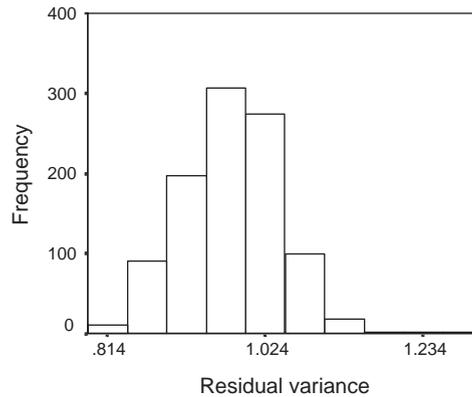


Fig. 10. Histogram of the residual variance (PLTAR model): 1000 series of 500 observations.

6. Applications to real time series

We analyze three well-known data sets, the Canadian lynx data, the sunspot numbers and the blowfly population data. These are listed in [Tong \(1990\)](#) and downloadable from the web sites ([RWC](#); [TSDL](#)).

The annual records of the number of lynx trapped in the Mckenzie River district of North-west Canada from 1821 to 1934 are known as Canadian lynx data, and include 114 observations. This time series has been extensively studied (see [Tong, 1990](#)). Data are usually transformed as \log_{10} (number recorded as trapped in year $1820+t$), $t = 1, \dots, 114$. The following SETAR(2;5,2) model ([Tong, 1980](#); [Tong and Dabas, 1990](#))

$$y_t = \begin{cases} 0.768 + 1.064y_{t-1} - 0.200y_{t-2} + 0.164y_{t-3} - 0.428y_{t-4} \\ \quad + 0.1817y_{t-5} + e_t & \text{if } y_{t-2} \leq 3.05, \\ 2.254 + 1.474y_{t-1} - 1.202y_{t-2} + e_t & \text{if } y_{t-2} > 3.05, \end{cases}$$

seems able to reproduce quite well the time series behavior. Parameter estimates were computed using the first 100 observations. The last 14 observations were set apart to check the model forecasting adequacy.

For this model the residual variance is 0.0415, the AIC value -268 , and the one-step-ahead forecast average square error (ASE) on years 1921–1934 is 0.0136. We also estimated a SETAR model for this data set using our GA procedure and $d=2$, obtaining slightly better results (residual variance 0.040, AIC = -271 , and ASE = 0.0101).

Our PLTAR models were estimated by using the first 100 observations. The delay parameter d has been considered from 1 to 6. For each value of d , the GA procedure was applied with the AR order varying from 1 to 10, the number of regimes from 1 to 3, and, in each regime, at least 20 observations were required. In [Table 1](#) the results are reported. It may be seen that the minimum AIC is obtained for delay parameter 3, 2 regimes and AR order 4, whilst the threshold constant equals 3.111. The least

Table 1
Fitting a PLTAR model to the Canadian lynx data

d	AIC	Regimes	Thresholds	AR order	σ_e^2	ASE (21–34)
1	–260.63	3	2.94; 3.411	2	0.0433	0.0224
2	–273.78	2	3.31	4	0.0350	0.0121
3	–274.91	2	3.111	4	0.0345	0.0225
4	–268.15	2	2.769	2	0.0425	0.0269
5	–256.16	3	2.894; 3.433	2	0.0455	0.0345
6	–260.82	3	3.111; 3.433	2	0.0432	0.0153

ASE, however, was found by assuming the delay parameter equal to 2, and, in this case, the threshold constant is assumed 3.31. By comparing the rows in Table 1 which correspond to $d=2$ and 3, we may judge the two models fairly equivalent, with a slight preference for the choice $d=2$ as it seems to ensure better forecasting performance. In Tong (1990), p. 377, motivation is provided for choosing either $d=2$ or 3.

The chosen model is the following:

$$y_t = \begin{cases} 0.3230 + (1.2106 - 0.1037y_{t-2})y_{t-1} + (-0.7704 + 0.4257y_{t-2})y_{t-2} \\ \quad + (2.0506 - 0.8498y_{t-2})y_{t-3} + (-1.1132 + 0.3726y_{t-2})y_{t-4} \\ \quad + e_t \quad \text{if } y_{t-2} \leq 3.3101, \\ 0.0484 + (-9.8474 + 3.2370y_{t-2})y_{t-1} + (25.4669 - 7.5009y_{t-2})y_{t-2} \\ \quad + (-24.9680 + 7.3127y_{t-2})y_{t-3} + (10.1568 - 3.0322y_{t-2})y_{t-4} \\ \quad + e_t \quad \text{if } y_{t-2} > 3.3101. \end{cases}$$

The Ljung–Box test statistics is 19.95, and the McLeod–Li one is 15.29. As the lags are 15 the χ^2 quantile at 0.95 level is about 25, therefore we do not reject the null hypothesis that residuals are uncorrelated and have linear structure.

Let us now consider the yearly recorded sunspot events. The time series starts from 1700 and is regularly updated (RWC). In Tong and Lim (1980) a SETAR model was proposed for the sunspot numbers (1700–1920). Let us write down this model as given in Tong (1990), p. 425, where some minor corrections were incorporated

$$y_t = \begin{cases} 11.97 + 1.71y_{t-1} - 1.26y_{t-2} + 0.236y_{t-3} + e_t & \text{if } y_{t-3} \leq 36.6, \\ 7.84 + 0.73y_{t-1} - 0.04y_{t-2} - 0.20y_{t-3} + 0.16y_{t-4} \\ \quad - 0.22y_{t-5} - 0.02y_{t-6} + 0.15y_{t-7} - 0.24y_{t-8} \\ \quad + 0.31y_{t-9} - 0.37y_{t-10} + 0.38y_{t-11} + e_t & \text{if } y_{t-3} > 36.6. \end{cases} \quad (12)$$

Model (12) is a SETAR(2;3,11) with residual variance 153.71, AIC=1084.33, and ASE (1921–1955) equal to 153.88. By employing our procedure based on GA, the best SETAR model for these data turned out to be a SETAR(4;2,3,3,1), with delay

Table 2
Fitting a PLTAR model to the sunspot numbers

d	AIC	Regimes	Thresholds	AR order	σ_e^2	ASE (21–55)
1	1105.93	4	24.1; 40; 57.1	2	173.74	245.6
2	1058.93	4	16; 35.6; 52.2	3	132.27	183.3
3	1065.97	3	20.9; 38.5	3	142.14	147.7
4	1086.74	4	9.6; 20.9; 38.5	4	144.03	217.0
5	1119.77	3	27; 47.8	2	191.04	220.2
6	1109.17	2	66.6	3	181.59	202.6
7	1099.61	3	36; 47.8	3	166.96	149.8
8	1088.91	2	57.1	3	164.81	140.0

parameter $d = 3$, having residual variance 136.66, AIC=1064 and ASE=138.5. In Table 2 some results from the PLTAR models, estimated by varying the delay parameter from 1 to 8, are displayed. We may see that the least AIC is obtained for $d = 2$. The GA searched for the optimal number of regimes as well, in the range 1–4. In each regime, however, there had to be at least 30 observations. The maximum AR order was set to 12.

Often the sunspot numbers are transformed according to $2\{[1+(\text{sunspot numbers in year } (1699 + t)]^{1/2} - 1\}$, $t = 1, 2, \dots$. In Ghaddar and Tong (1981) a SETAR(2;11,3) model was fitted to the 280 transformed data (1700–1979), with delay parameter 8. In Tong (1990, p. 421), this model was considered with some minor corrections of the autoregressive parameter estimates. The modified version is the following SETAR(2;10,2) model

$$y_t = \begin{cases} 1.89 + 0.86y_{t-1} + 0.08y_{t-2} - 0.32y_{t-3} + 0.16y_{t-4} - 0.21y_{t-5} \\ \quad - 0.00y_{t-6} + 0.19y_{t-7} - 0.28y_{t-8} + 0.2y_{t-9} + 0.1y_{t-10} + e_t \\ \text{if } y_{t-8} \leq 11.93, \\ 4.53 + 1.41y_{t-1} - 0.78y_{t-2} + e_t \\ \text{if } y_{t-8} > 11.93. \end{cases}$$

The pooled residual variance is 3.734, AIC = 381.08, and the ASE over the years 1980–2001 is about 4.189. Several values for the delay parameter d were tried to be used with our GA procedure for estimating some PLTAR models to the transformed data (1700–1979). In Table 3 the results are displayed.

Both in the transformed and untransformed case it may be seen that the use of piecewise linearly varying parameters may provide an equivalent or slightly better fitting with more parsimonious models. From the residuals (transformed data) we may compute the Ljung–Box and McLeod–Li test statistics with 35 lags. We obtain 28.02 and 45.69, respectively, so that the model is to be accepted.

The blowfly population data are the bidaily population sizes of the blowflies obtained by Nicholson (1957). There are 350 observations available. In Tsay (1988) several

Table 3
Fitting a PLTAR model to the transformed sunspot numbers

d	AIC	regimes	thresholds	AR order	σ_e^2	ASE (80–01)
1	415.6	2	7.4234	3	4.343	4.267
2	376.11	3	5.2388; 10.1820	3	3.638	3.088
3	358.07	4	9.15; 11.40; 15.20	4	3.180	4.427
4	391.25	4	7.74; 10.82; 13.62	3	3.736	7.542
5	407.84	2	14.1493	3	4.219	3.1984
6	407.09	3	11.4015; 16.0997	3	4.084	6.495
7	385.66	3	8.5262; 13.3623	3	3.770	4.249
8	383.95	3	11.9571; 14.5892	3	3.746	3.756

Table 4
Fitting a PLTAR model to the \log_{10} -transformed blowfly data

d	AIC	regimes	thresholds	AR order	σ_e^2
1	–814.27	4	2.7412; 3.0997; 3.4333	4	0.0123
2	–692.00	1		4	0.0267
3	–703.62	2	2.7528	4	0.0239
4	–714.31	4	2.7152; 3.0378; 3.2938	4	0.0204
5	–684.37	3	2.734; 3.0878	4	0.0251
6	–693.92	3	2.9165; 3.236	4	0.0239
7	–716.67	3	2.9566; 3.2217	4	0.0213
8	–727.70	4	2.9614; 3.2529; 3.5562	3	0.0201

models were considered for this data set. A smooth threshold autoregressive (STAR) model was suggested in [Chan and Tong \(1986\)](#) for the \log_{10} -transformed series. Two STAR models were estimated for observations (1–206) and reported in [Tsay \(1988\)](#). The residual variance is 0.021 for the first version and 0.023 for the second one, and, for this latter, we may compute the AIC equal to –714.76. Though this model was judged inadequate (see [Tsay, 1988, p. 249](#)), we tried to estimate a PLTAR model on the same 206 observations by using our GA procedure. We chose the maximum AR order equal to 4, and the maximum number of regimes equal to 4. In [Table 4](#) the results for the delay parameters from 1 to 8 are displayed, and the least AIC values are attained for $d=1$ and 8. The residuals from the fit of the PLTAR model are plotted in [Fig. 11](#) for $d=1$ and in [Fig. 12](#) for $d=8$. It may be seen that both PLTAR models for delay parameters 1 and 8 are satisfactory. The residuals for $d=8$ are slightly less autocorrelated, while their variance is larger due to an isolated large value. This supports the suggestion given in [Tsay \(1988\)](#), where a similar plot is displayed for the residuals from a STAR model, that at least one outlying observation is present in the data set. The Ljung–Box statistics was computed 37.39 with 35 lags. We may conclude that there is no linear structure left in the residuals' sequence. In addition, according

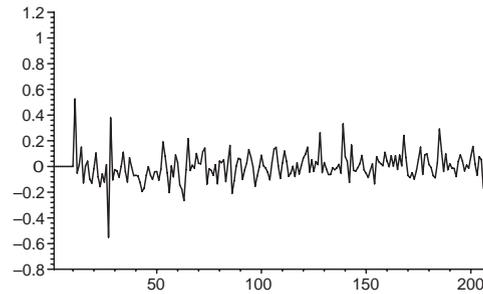


Fig. 11. Residuals from the PLTAR model fitted to the blowfly data, $d = 1$.

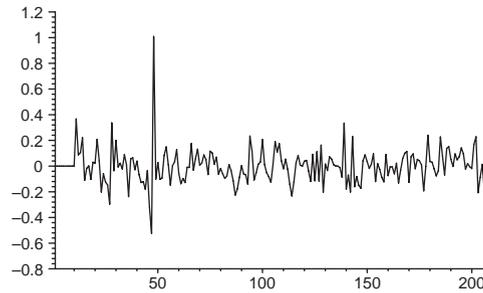


Fig. 12. Residuals from the PLTAR model fitted to the blowfly data, $d = 8$.

to the McLeod–Li test (24.94 for 35 lags), we are allowed to accept the absence of nonlinearity.

7. Conclusions

We considered a special nonlinear threshold model that we called piecewise linear (self-exciting) threshold autoregressive (PLTAR). Such model may be obtained either as a particular state-dependent model or a particular functional autoregressive model. We proposed an easy method for identifying and estimating the PLTAR model, which is essentially based on a genetic algorithm. In a simulation experiment we used the PLTAR model for approximating some different nonlinear models, obtaining satisfactory results. We also fitted the PLTAR model to some well-known real time series. Both adequacy of fitting and forecasting performance were found quite satisfactory.

References

Akaike, H., 1977. An entropy maximisation principle. In: Krishnaiah, P.R. (Ed.), Proceedings of the Symposium on Applied Statistics. North-Holland, Amsterdam, pp. 27–41.

- Applied Statistics Algorithms, web site <http://lib.stat.cmu.edu/apstat/>. The routine RANDOM is the algorithm AS183 (Wichmann, B.A., Hill, I.D., 1982. An efficient and portable pseudo-random number generator. *Applied Statistics*, 31, 188–190; correction, 33 (1984), 123). An alternative is also provided by the real function UNIFORM (L'Ecuyer, P. Efficient and portable combined random number generators, C.A.C.M., vol. 31, 742–749 & 774–, June 1988). The routine PPND16 is the algorithm AS241 (Wichura, M. J., 1988. The percentage points of the normal distribution. *Applied Statistics*, 37, 477–484).
- Cai, Z., Fan, J., Yao, Q., 2000. Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* 95, 941–956.
- Chan, W.-S., Cheung, S.-H., 1994. On robust estimation of threshold autoregressions. *J. Forecasting* 13, 37–49.
- Chan, K.S., Tong, H., 1986. On estimating thresholds in autoregressive models. *J. Time Series Anal.* 7, 179–190.
- Chatterjee, S., Laudato, M., 1997. Genetic algorithms in statistics: procedures and applications. *Comm. Statist. Theory Methods* 26 (4), 1617–1630.
- Chatterjee, S., Laudato, M., Lynch, L.A., 1996. Genetic algorithms and their statistical applications: an introduction. *Comput. Statist. Data Anal.* 22, 633–651.
- Chen, R., Tsay, R.S., 1993. Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* 88, 298–308.
- De Jong, K.A., 1975. An analysis of the behavior of a class of genetic adaptive systems. Ph.D. Thesis, Department of Computer and Communication Sciences, University of Michigan, Ann Arbor, MI.
- Fuller, W.A., 1996. *Introduction to Statistical Time Series*, (2nd Edition). Wiley, New York.
- Ghaddar, D.K., Tong, H., 1981. Data transformation and self-exciting threshold autoregression. *Appl. Statist.* 30, 238–248.
- Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
- Haggan, V., Ozaki, T., 1981. Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* 68, 189–196.
- Haupt, R.L., Haupt, S.E., 1998. *Practical Genetic Algorithms*. Wiley, New York.
- Holland, J.H., 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor (2nd Edition: The MIT Press, Cambridge, 1992).
- Jennison, C., Sheehan, N., 1995. Theoretical and empirical properties of the genetic algorithm as a numerical optimizer. *J. Comput. Graphic. Statist.* 4, 296–318.
- Man, K.F., Tang, K.S., Kwong, S., 1999. *Genetic Algorithms: Concepts and Designs*. Springer, London.
- Michalewicz, Z., 1996. *Genetic Algorithms + Data Structures = Evolution Programs*, (3rd Edition). Springer, Berlin.
- Mitchell, M., 1996. *An Introduction to Genetic Algorithms*. The MIT Press, Cambridge, MA.
- Nicholson, A.J., 1957. The self-adjustment of populations to change. *Cold Spring Harbour Symp. Quant. Biol.* 22, 153–173.
- Ozaki, T., 1981. Non-linear threshold autoregressive models for non-linear random vibrations. *J. Appl. Probab.* 18, 443–451.
- Ozaki, T., 1982. The statistical analysis of perturbed limit cycle processes using nonlinear time series models. *J. Time Series Anal.* 3, 29–41.
- Ozaki, T., Oda, H., 1978. Non-linear time series model identification by Akaike's information criterion. In: Dubuisson, B. (Ed.), *Information and Systems*. Pergamon Press, Oxford, pp. 83–91.
- Pittman, J., Murthy, C.A., 2000. Fitting optimal piecewise linear functions using genetic algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (7), 701–718.
- Priestley, M.B., 1988. *Non-linear and Non-stationary Time Series Analysis*. Academic Press, New York.
- RWC, Belgium World Data Center for the Sunspot Index. Web site is <http://sidc.oma.be>
- Tong, H., 1980. A view on non-linear time series model building. In: Anderson, O.D. (Ed.), *Time Series: Proceedings of the International Conference*, Nottingham University, March 1979. North-Holland, Amsterdam, pp. 41–56.
- Tong, H., 1983. *Threshold Models in Non-linear Time Series Analysis*. Springer, New York.
- Tong, H., 1990. *Non-linear Time Series. A Dynamical System Approach*. Oxford Science Publications, Clarendon Press, Oxford.

- Tong, H., Dabas, P., 1990. Cluster of time series models: an example. *J. Appl. Statist.* 17 (2), 187–198.
- Tong, H., Lim, K.S., 1980. Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. Ser. B* 42, 245–292.
- Tsay, R.S., 1988. Non-linear time series analysis of blowfly population. *J. Time Series Anal.* 9, 247–263.
- TSDL, Time Series Data Library maintained by Rob Hyndman and Muhammad Akram. Web site is <http://www-personal.buseco.monash.edu.au/~hyndman/TSDL>
- Wu, B., Chang, C.-L., 2002. Using genetic algorithms to parameters (d, r) estimation for threshold autoregressive models. *Comput. Statist. Data Anal.* 38, 315–330.