COMISEF
Computational Optimization Methods in Statistics, Econometrics and Finance

- Marie Curie Research and Training Network funded by the EU Commission through MRTN-CT-2006-034270 -

COMISEF WORKING PAPERS SERIES

WPS-026 28/01/2010

Multi-regime models for nonlinear nonstationary time series

F. Battaglia
M. K. Protopapas

Multi–regime models for nonlinear nonstationary time series

Francesco Battaglia and Mattheos K. Protopapas

January 28, 2010

Abstract

Nonlinear nonstationary models for time series are considered, where the series is generated from an autoregressive equation whose coefficients change both according to time and the delayed values of the series itself, switching between several regimes. The transition from one regime to the next one may be discontinuous (self-exciting threshold model), smooth (smooth transition model) or continuous linear (piecewise linear threshold model). A genetic algorithm for identifying and estimating such models is proposed, and its behavior is evaluated through a simulation study and application to temperature data and a financial index.

1 Introduction

Often time series exhibit a nonlinear or nonstationary behavior and cannot be adequately fitted by the popular autoregressive moving average models. Several more complicated models have been suggested, and among them a particularly interesting and simple class is that of the multi-regime threshold models. The series is generated by several alternative linear autoregressive equations (the regimes) and the generating process switches from one to another according to the value of an indicator, that may be related to time or to another time series (the driving variable, also a possibly delayed value of the series itself). In the first case we have a nonstationary but linear model (also called structural change, general references are Bai and Perron, 1998; Lin and Teräsvirta, 1994), while in the second case the model is nonlinear but stationary, under suitable choices of the parameter values. These are usually called threshold models; a general reference is Tong (1990). However for time series with complicated dynamic structure, such models should be considered only as approximations or partial tools for describing only a part of the features of the series. We assume, as in Rissanen (2007) that there is no absolutely correct model, but only models with a better or worse fit. Moreover, especially when the fitting measurement is related to the second moments, often a confusion may arise between the two kinds of models, nonlinear stationary and linear nonstationary (see Koop and Potter, 2001; Carrasco , 2002; Dupleich Ulloa, 2005).

A possible option is considering models which are simultaneously regime changing both according to time and to a driving variable. Obviously, identification is much more complicated and this is perhaps the reason why such nonlinear, nonstationary models have been rarely addressed.

Lundberg et al. (2003) proposed a model based on a first order autoregressive structure with the parameter changing both according to time and to a driving variable alternating between two regimes only, with smooth change (in the same way as smooth transition autoregressive models of Teräsvirta, 1994). Battaglia and Protopapas (2009) extended this framework to allow also change of the autoregressive parameters in a piecewise linear fashion (piecewise linear threshold multi-regime model, Baragona et al., 2004) and proposed a genetic algorithm for building such a model, but also limited to two regimes.

The main contribution of the present paper consists in removing the limitation of two regimes. This involves non trivial complications in the genetic algorithm and requires a completely new coding. The proposed procedure enables to build models that alternate according to time, among several regimes, and inside each of them, the series follows different threshold models which may exhibit several regimes according to the level of the driving variable. The resulting models are therefore suitable also for very long time series with strongly nonlinear behavior.

In Section 2 we describe the model and the identification and estimation issues. Section 3 introduces the genetic algorithm and presents the proposed procedure in details. The results of a simulation study are summarized in Section 4, and applications to some real time series are discussed in Section 5; Section 6 concludes the paper.

2 The Model and Estimates

The original autoregressive threshold model proposed by Tong (1990), has at each t an autoregressive structure, where parameters change according to the value of another series (the driving variable). If the driving variable is given by the delayed series itself, we have a self exciting threshold model (SETAR):

$$X_t = \phi_0^{(k)} + \phi_1^{(k)} X_{t-1} + \ldots + \phi_p^{(k)} X_{t-p} + \epsilon_t \quad , \; X_{t-d} \in R_k \tag{1}$$

where $\{R_k\}$ is a partition of the real line, and d is called the delay. A generalization proposed by Teräsvirta (1998), avoiding discontinuities in the autoregressive parameters, is called smooth transition autoregressive model (STAR), since the transition from one regime to the next is driven by a continuous function (generally a logistic). If r_L denotes the number of regimes and $R_k = (l_{k-1}, l_k], k = 1\ldots, r_L$, the STAR equation may be written:

$$X_t = \sum_{k=1}^{r_L} \sum_{j=1}^{p} \phi_j^{(k)} G_{k-1}(X_{t-d}) X_{t-j} + \epsilon_t \tag{2}$$

where $G_0(x) = 1$ and $G_k(x) = [1 + \exp(-\gamma_L(x - l_k))]^{-1}$. The behavior of the autoregressive coefficients is essentially constant in each regime, with a continuous smooth change between regimes, whose speed is controlled by the constant $\gamma_L(> 0)$. The SETAR model may be interpreted as a special case of the STAR model, when γ tends to infinity.

A different proposal, where the autoregressive coefficients change linearly and continuously with the driving variable X_{t-d}, but with different slope in each regime, is the piecewise linear threshold model (Baragona et al., 2004), described by

$$X_t = \sum_{j=1}^{p}[\phi_j^{(0)} + \phi_j^{(1)} X_{t-d} + \sum_{k=2}^{r_L} \phi_j^{(k)} \max(0, X_{t-d} - l_{k-1})]X_{t-j} + \epsilon_t \quad (3)$$

Here the autoregressive coefficient behave like a linear spline across regimes. The PLTAR may be written in a similar fashion to the STAR letting $S_k(x) = \max(0, x - l_{k-1})$, $S_1(x) = x$, $S_0(x) = 1$ and

$$X_t = \sum_{k=1}^{r_L+1} \sum_{j=1}^{p} \phi_j^{(k)} S_{k-1}(X_{t-d}) X_{t-j} + \epsilon_t \quad (4)$$

Note however that here there is an additional parameter (the linear term in X_{t-d}) for each lag. Therefore the sum over k ranges from 1 to the number of regimes plus 1.

To allow for non zero and varying means, intercept terms $\phi_0^{(k)}$ may be added. For a STAR model these terms may depend on the driving variable in the same fashion as before, while for a PLTAR the linear term $\phi_0^{(1)}$ as in (3) disappears because it would be undistinguishable from $\phi_d^{(0)}$ in the case $d \leq p$:

$$\phi_0 = \sum_{k=1}^{r_L} \phi_0^{(k)} G_{k-1}(X_{t-d})$$

or

$$\phi_0 = \phi_0^{(0)} + \sum_{k=2}^{r_L} \phi_0^{(k)} S_k(X_{t-d})$$

for the STAR and the PLTAR case respectively.

In order to take into account additional nonstationarity, we allow each of the coefficients $\phi_j^{(k)}$ to depend on time also, according to a STAR or PLTAR structure. Let r_T denote the number of regimes in time, and t_j denote the thresholds so that the regimes are defined by the partition $R'_k = (t_{k-1}, t_k]$, where $1 = t_0 < t_1 < \ldots < t_{r_T} = N$ (where N is the series length).

We may allow STAR dependence on time using

$$\phi_j^{(k)} = \sum_{i=1}^{r_T} \beta_j(i,k) G'_{i-1}(t) \ , \ j = 1, \ldots, p \quad (5)$$

with $G'_0(t) = 1$, $G'_i(t) = [1 + \exp(-\gamma_T(t - t_i)]^{-1}$, $i = 1, \ldots, r_T$.

Alternatively, a time nonstationarity following a PLTAR structure may be defined as

$$\phi_j^{(k)} = \sum_{i=1}^{r_T+1} \beta_j(i,k) S'_{i-1}(t) \ , \ j = 1, \ldots, p \tag{6}$$

where $S'_i(t) = \max(0, t - t_{i-1})/N$, $S'_1(t) = t/N$, $S'_0(t) = 1$.

On combining the different types of dependence on levels of the driving variable, and time, nine different kinds of models result: Stationary, STAR or PLTAR in time, combined with linear, STAR or PLTAR in levels. Denoting $r_L^* = r_L$ for STAR models in levels, $r_L^* = r_{L+1}$ for PLTAR, and $r_L^* = 1$ for linear models, and analogously $r_T^* = r_T$ for STAR models in time, $r_T^* = r_{T+1}$ for PLTAR in time, and $r_T^* = 1$ for stationary models, the total number of parameters is $r_L^* r_T^* p$ (or $r_L^* r_T^*(p+1) - r_T$ if the intercept term is present). The model may be written in a more compact form as a state dependent model (Priestley, 1988)

$$X_t = \sum_{j=0}^{p} \phi_j(t, X_{t-d}) X_{t-j} + \epsilon_t \tag{7}$$

where the autoregressive functional coefficients $\phi_j(t, X_{t-d})$ are expressed as linear functions of the elementary parameters $\beta_j(i,k)$:

$$\phi_j(t, X_{t-d}) = \sum_{i=1}^{r_T^*} \sum_{k=1}^{r_L^*} \beta_j(i,k) c_{ik}(t, X_{t-d}) \tag{8}$$

and the coefficients c_{ik} are different for different model types, and describe the dependence on the transition functions. A simple matrix expression for the coefficients c_{ik} is derived in the Appendix. Since the final model equation is linear in the elementary parameters $\beta_j(i,k)$, they may be simply estimated by means of least squares.

Finally, to achieve slightly more parsimony, we shall allow a different order for the dependence on time or on levels, in the sense that the dependence of the parameters $\phi_j(t, X_{t-d})$ on t may be limited to a maximum lag less than p, which we call order in time; the same may be true for the dependence on levels X_{t-d}, limited to a maximum lag that we call order in levels, while p will be denoted as overall order.

Summarizing, a multi-regime model is defined by: its type in levels (linear, STAR or PLTAR); its type in time (stationary, STAR or PLTAR); the delay d of the driving variable; the orders in time and levels; the thresholds in levels l_k and in time t_k; and if needed, logistic speed coefficients γ_L and γ_T. Conditional on the values of all these structural parameters, the remaining elementary parameters $\beta_j(i,k)$ which drive the evolution of the autoregressive weights, may be estimated by least squares.

The identification of a multi-regime model requires the selection of the solution that possesses the best properties, out of a large and discrete space of elements. Such kind of problems, common in statistical applications, have been

recently addresses by means of meta-heuristic methods, and in particular genetic algorithms. In addition to several statistical identification problems, e.g. autoregressive moving average model fitting (Gaetan, 2000), outlier detection (Crawford and Wainwright, 1995), variable selection in regression (Chatterjee et al., 1996), the genetic algorithm was suggested specifically for threshold models by Wu and Chang (2002); Davis et al. (2006); Battaglia and Protopapas (2009).

A genetic algorithm for building multi-regime models is presented in the next Section. Given a time series and an identification criterion, it selects the best type of model (in time and in levels), number of regimes, orders and thresholds in time and level, and provides least squares estimates of the model coefficients.

3 Genetic Algorithms

A genetic algorithm is an optimization heuristic algorithm inspired by the process of the evolution of life (Holland, 1975). Candidate solutions are encoded as chromosomes (usually binary strings), and the algorithm evolves a set (population) of these chromosomes, using transformation operators (crossover and mutation), in steps called generations, to achieve a near optimal solution. The performance of the candidate solutions is measured by a "fitness function". In the case of a canonical genetic algorithm that includes an elitist strategy, like the one employed here, Rudolph (1997) has shown that the difference between the optimal fitness value and the best fitness reached in the n-th generation is a non-negative supermartingale converging to zero almost surely as $n \to \infty$. Readers' knowledge of genetic algorithms is assumed throughout this paper. Holland (1975) and Goldberg (1989) describe genetic algorithms in depth.

3.1 Chromosome Encoding

In the multi-regime threshold models analyzed here, the following decision variables are relevant: the type of model in time and in levels, the order in time and the order in levels, and in the case of STAR and PLTAR models, the number of regimes (to a maximum of 4) in time and in levels, and the delay of the driving variable and the thresholds in time and in levels. Finally, there are also the γ coefficients in time and/or in levels, in the case of a STAR model.

All the relevant decision variables are encoded in the chromosome in distinct bit-strings (genes). For determining the number of thresholds, the type of the model and most other variables, a mapping of the binary values of the corresponding gene to integer values is required; this is done by the transformation

$$x = \sum_{n=1}^{L} e_n 2^{n-1} + d \tag{9}$$

where x is the value of the decision variable, L is the length of the corresponding bit-field in the chromosome, e_n, the n_{th} bit of the bit-field, and d the minimum

value of the decision variable (so, as described below, for the model type $d = 0$, while in the case of the number of regimes $d = 1$). For determining the threshold values, the binary values of the corresponding genes should be transformed to real values between 0 and 1 . Thus a slightly different transformation is used:

$$x = \frac{\sum_{n=1}^{L} e_n 2^{n-1}}{2^L - 1} \tag{10}$$

The first two genes determine the model type. Since there are 3 possible types of models (STAR, PLTAR and stationary or linear) two bits are required for each of these genes. A value of one in the first gene corresponds to a STAR model in time, a value of two to a PLTAR model, and a zero or three to a stationary model. The same values apply to levels, encoded in the second gene.

The next genes encode the values of the order in time and levels. The order of the model in time ranges from 0 to 7, so 3 bits are required. Consequently, bits 5-7 of the chromosome account for the gene that encodes the order in time. The same holds true for the next bit-string, which denotes the order of the model in levels, and resides on bits 8-10. The overall order of the model is equal to the maximum of the orders in time and in levels.

The next two genes in the chromosome determine the number of regimes in time and in levels, respectively. We analyze a maximum of four regimes in this study, so two bits are required for each of these genes.

Then, there are 3 genes encoding the thresholds in time and 3 for the thresholds in delay. This is the most delicate part of the coding procedure, since most intuitive ways of coding, such as binary indicators, lead to large redundancy or illegal chromosome values. We use fixed-length chromosomes, but the number of thresholds taken into consideration when a chromosome is evaluated, depends on the number of regimes of the chromosome. If there are r_T number of regimes in time (r_L in levels), only the first $r_T - 1$ thresholds in time ($r_L - 1$ in levels) are used ("active"). Each of the threshold genes consists of 12 bits. The binary values in the "active" threshold genes are first transformed to real values u_i, $0 \leq u_i \leq 1$ by (10). These numbers determine the percentage of the remaining time series observations that are attributed to the corresponding regime. Since a minimum of m observations are assumed to be present in each regime, a second transformation is required; the equations that map the real values u_i that are extracted from the threshold genes to the values that determine the actual thresholds, depend on the number of regimes of the chromosome. So, for 2 regimes in time , the first m observations should fall on the first regime, and the last m into the second; thus a value of zero in $u_{1,t}$ corresponds to a threshold given by the m_{th} observation of the time series and a value of one corresponds to observation $N - m$ (N is the total number of observations). If there are three regimes, the last m values should fall in the third regime, and at least m preceding observations should fall into the second. A value of one in the u_{1t} corresponds to the fact that the first regime is as large as possible: in that case the first threshold corresponds to observation $N - 2m$, the second regime consists of the observations $N - 2m + 1, \ldots, N - m$, and the third of the

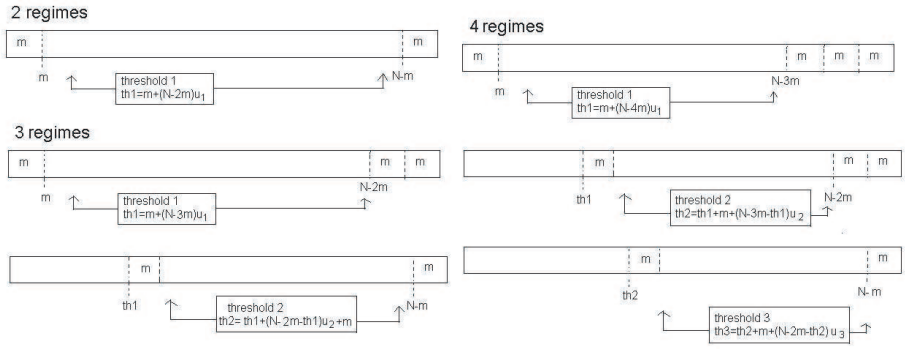

Figure 1: Threshold coding

observations $N-m+1,\ldots,N$. The whole process is described in the following equations, and illustrated in figure 1.

- Two regimes. If the gene is denoted by u_1, the threshold is $t_1 = m+(N-2m)u_1$
- Three regimes. Genes u_1 and u_2. Thresholds: $t_1 = m+(N-3m)u_1$, and $t_2 = t_1+m+(N-2m-t_1)u_2$.
- Four regimes. The genes are denoted by u_1, u_2, u_3. The thresholds are obtained from: $t_1 = m+(N-4m)u_1$; $t_2 = t_1+m+(N-3m-t_1)u_2$, and $t_3 = t_2+m+(N-2m-t_2)u_3$.

The values t_i are enough to determine the values of the thresholds in time, since the actual thresholds will be points in time (a number between 1 and N). In the case of the threshold in levels, the thresholds are the actual observations themselves. Threshold τ for example, can be chosen as a real number in-between two consecutive (in terms of magnitude) observations, $Y_k \leq \tau < Y_{k+1}$, or — something that has the same effect— as the observed value Y_k, where $\{Y_t\}$ denotes the observations arranged in increasing order. We have chosen the latter, constraining the threshold values to belong to the set of the observed values of the time series. So, to finally determine the thresholds in levels, we map the values determined by the process described in the figures, to the corresponding time series observations using the sorted set of observations. So, for $u_1 = 0$, the m_{th} lowest observation of the time series is used as the first threshold, while e. g. for 3 regimes in levels and $u_1 = 1$, the first threshold in levels is the $(N-2m)_{th}$ lowest observation of the time series. In our implementation we set $m = N/10$.

The next gene encodes the delay of the driving variable in the cases of STAR or PLTAR models. If the chromosome represents a stationary model its value is neglected. It occupies 3 bits, since we assume a maximum delay of 8.

The final two genes encode the γ values for a STAR model in time and/or in delay, and consist of $L_\gamma = 7$ bits each. First a mapping of these binary numbers to real numbers between (γ_1, γ_2), is applied:

$$\gamma = \gamma_1 + x(\gamma_2 - \gamma_1)/(2^L - 1) \tag{11}$$

Since the γ parameter controls the speed of change from 0 to 1 in the logistic function, we can select a maximum value, that makes the STAR model essentially indistinguishable from a SETAR, i.e. such that, for a sufficiently small value ϵ the logistic function has value ϵ immediately before the threshold, and value $(1-\epsilon)$ immediately after. In the time-varying case, if we assume that t_0 is the threshold in time, $G'(t_0 - 1) = \{1 + e^{\gamma}\}^{-1} = \epsilon$ and $G'(t_0 + 1) = \{1 + e^{-\gamma}\}^{-1} = 1 - \epsilon$, which amounts to $\gamma = \log\{(1-\epsilon)/\epsilon\}$. A similar argument can be used for the γ parameter in levels, applied to the ordered sequence of the observed values this time. If Y_T denotes the threshold in levels, then $G(Y_{T-1}) = \left[1 + e^{-\gamma(Y_{T-1} - Y_T)}\right]^{-1} = \epsilon$ and $G(Y_{T+1}) = \left[1 + e^{-\gamma(Y_{T+1} - Y_T)}\right]^{-1} = 1 - \epsilon$. For evaluating approximately the constraints we put $Y_{T+1} - Y_T \simeq Y_T - Y_{T-1} \simeq (Y_N - Y_1)/N = s$ so that they are solved by $\gamma = \log\{(1-\epsilon)/\epsilon\}/s$. In order to select the minimum gamma values, we shall assume that the change from ϵ to $1-\epsilon$ in the logistic function requires an interval not longer than $1/q$ of the full observation interval. Consequently in a time-varying STAR $G'\left(t_0 - \frac{N}{2q}\right) = \left[1 + \exp\left(\frac{\gamma N}{2q}\right)\right]^{-1} = \epsilon$ and $G'\left(t_0 + \frac{N}{2q}\right) = \left[1 + \exp\left(\frac{-\gamma N}{2q}\right)\right]^{-1} = 1 - \epsilon$ which leads to $\gamma = \log\{(1-\epsilon)/\epsilon\}2q/N$; the result in the case of γ in levels are similar. Consequently, the search interval for gamma in time will be $[(2q/N)\log(1-\epsilon)/\epsilon, \log(1-\epsilon)/\epsilon]$, and for gamma in levels $[2q/(Y_N - Y_1)\log(1-\epsilon)/\epsilon, N/(Y_N - Y_1)\log(1-\epsilon)/\epsilon]$. In our applications we selected $\epsilon = 0.01$ and $q = 10$.

3.2 Fitness, crossover and mutation

In order to compute the fitness, for each given chromosome the model is estimated by least squares and the residual variance estimate $\hat{\sigma}^2$ is computed. The most popular choice in time series literature is evaluating fitness through an identification criterion. This amounts to computing a penalized gaussian likelihood: if model M has p parameters, the quantities

$$IC(M) = N \log \hat{\sigma}^2_M + c(p)$$

are evaluated and the model with minimum value is selected. Choosing a linear penalization function $c(p) = cp$ corresponds to the generalized AIC criterion (Bhansali and Downham, 1977): the original Akaike's criterion is obtained for $c = 2$, while e. g. the Schwartz criterion corresponds to $c = \log N$. The value $c = 3$ was selected in Battaglia and Protopapas (2009), because in this case its behavior is equivalent to a test of linearity-stationarity against two-regimes alternatives proposed by Lundberg et al. (2003). A different choice (also adopted

for multi-regime in time models by Davis et al., 2006) is the minimum description length of Rissanen (2007), which is based on minimizing the code length that completely describes the data. The function $c(p)$ is taken equal to the code length of the fitted model, which may be difficult to estimate in some cases.

In some statistical problems addressed by means of genetic algorithms (e.g. , outlier detection in time series, see Baragona et al., 2001) the tuning constant c may be chosen so that the identification criterion $IC(M)$ becomes proportional to the posterior probability of the model in a Bayesian framework, and the value of the constant c is determined by the prior distribution.

In any case, the fitness function must be related to the identification criterion through a monotone decreasing transformation, and positiveness has to be ensured. We use a simple negative exponential transformation:

$$\text{fitness}(M) = \exp\{-IC(M)/N\} = \hat{\sigma}_M^{-2} \exp\{-c(p)/N\}.$$

The penalization term adopted here for simulations and applications is $c(p) = 3p$, essentially for comparison with results of Battaglia and Protopapas (2009).

Parent selection is "roulette wheel selection", i.e. the probability of a chromosome to be selected as a parent is proportional to its fitness.

Ultimately, the chromosome consists of $n = 89$ binary digits. Alander (1992) suggests a population size between n and $2n$, while Reeves (1993) suggests that the population size should be adequate so that the probability of any allele (bit value) to be present at each locus (bit position) is high enough, as given by the formula $n \geq 1 + \log(-l/\log p)/\log 2$ where l is the number of genes, p the probability and n the population size. With our chromosome length, a population size as large as 20 ensures a probability larger than 0.999. Thus, we use a population of 50 chromosomes.

A larger chromosome length than that used in Battaglia and Protopapas (2009) implies that a higher number of generations is required for the algorithm to perform effectively, therefore we used 700 generations. The crossover operator used is random point crossover (Goldberg, 1989) and was applied with probability one; the set of possible cutting points consists of the boundary points between genes so that genes are inherited as a whole after the crossover operation is applied. A bit-wise independent mutation operator is then applied to the children chromosomes; the probability of mutating a bit is fixed throughout the course of the algorithm, we chose 0.025.

Finally, a form of elitism is employed: the best chromosome of any generation is inherited, as it is, in the population of the next generation, replacing a random chromosome of the population.

4 A Simulation Study

An extensive simulation study was performed in order to evaluate the behavior of the proposed algorithm. In general, for each time and level regime, simulated data followed a first order autoregressive model obtained from gaussian standardized white noise, and only the autoregressive coefficient values changed

according to regimes. Though restricted to the simplest autoregressive scheme, this experiment helped to verify that the proposed procedure is able to detect nonlinearity and nonstationarity.

We considered both the simulated set analyzed in Battaglia and Protopapas (2009), consisting of series of 500 observations, replicated 100 times, generated according to various models with no more than two regimes (see Battaglia and Protopapas, 2009), and a new set of series simulated according to several models with three regimes, each with 100 replications and 600 observations.

First of all, we compare the results with these obtained in Battaglia and Protopapas (2009), where a similar genetic algorithm procedure is employed, but allowing for more than 2 regimes. The results are very similar, and the models with 3 (or more) regimes (which would be wrong since the analyzed series there had two regimes) are very rarely selected, while the detection ability of nonlinearity and nonstationarity is equivalent for the two algorithms.

Since a complete account of the results cannot be given here because of lack of space, we consider only selected cases. On varying the values of autoregressive coefficients however, the results remain essentially stable.

1. White noise. The stationary in time and linear in levels type is selected 95 times out of 100, thus confirming the appropriateness of the tuning constant.

2. Stationary linear. On a set of 100 replications of a first order autoregressive series with parameter 0.7 the genetic algorithm selected a stationary in time - linear in levels type 97% times.

3. Nonstationary with 2 regimes, linear. The analyzed series were generated according to an AR(1) with parameter 0.5 for the first 250 data, and parameter -0.7 for the last 250. The genetic algorithm selects always linear models, nonstationary 95% times (STAR 90, PLTAR 5), and two regimes in time are selected 93% times.

4. Nonstationary with 3 regimes, linear. The series have 600 observations and thresholds at 150 and 400. The structure is AR(1) and the values of the autoregressive coefficients in the three regimes are 0.5, -0.7, and 0.7 respectively. The genetic algorithm procedure, repeated on 100 replications, selected a linear nonstationary model 94 times (81 STAR and 13 PLTAR), and nonlinear nonstationary model in the remaining 6 cases. The estimated number of time regimes was 3 in 65 replications, and 4 in the other 35 cases.

5. Nonlinear, stationary. These series were generated according to a self exciting threshold mechanism with two regimes: if $X_{t-1} \leq 0$, an AR(1) equation is used with coefficient 0.5, and if $X_{t-1} > 0$, the coefficient is -0.3. Here, on 100 replications the proposed procedure selected a nonlinear stationary model 98 times (74 times STAR and 24 PLTAR) and in the remaining two cases a nonlinear nonstationary model. The estimated

number of regimes in levels was 2 with frequency 91, and 3 with frequency 9.

6. Nonstationary and nonlinear. We considered series simulated according to first order autoregressive processes where the coefficient is varied both according to time, with three regimes $1 \leq t \leq 150$, $150 < t \leq 400$, $400 < t \leq 600$, and simultaneously according to the levels, in two regimes $X_{t-1} \leq 0$ and $X_{t-1} > 0$. Nine different datasets were considered, each consisting of 100 replications and series length 600. For the first one, at the low level regime ($X_{t-1} \leq 0$) the AR parameter for the three time regimes was respectively 0.45, -0.45, and -0.15, while for the high level regime ($X_{t-1} > 0$) the corresponding figures were -0.75, 0.25, and -0.85. The other datasets were simulated by subtracting repeatedly 0.05 from each autoregressive coefficients, and maintaining the difference among regimes constant. The results are summarized in table 1. Different choices of the autoregressive coefficients do not influence largely the results. The nonlinear and nonstationary nature of the series is correctly detected nearly always. The STAR-STAR type is suggested in 60-65 replications out of 100, while in the other cases STAR-PLTAR or PLTAR-PLTAR type is selected. Also, the estimates of the number of regimes are satisfying. For time, the correct number (3) is selected with frequencies of about 70%, while in about 20 cases the estimate is 4, and seldom 2. For the levels, the correct number (2) is selected more than 90 times out of 100, and sometimes the estimate is 3.

Summarizing, the simulation confirms that the genetic algorithm procedure is able to detect nonlinear and nonstationary features in the analyzed series, and also the estimation of the number of regimes is relatively precise. It was also seen that the discrimination between smooth transition and piecewise linear threshold is not easy, and clear cut, also because often the fitting obtained with the two model types tends to be equivalent.

5 Applications

To gain some insight into the potentialities of the proposed method, we consider some example data in two subject areas: climatology and finance. In the first case, yearly temperature data are analyzed to discover structural changes, while in the second case we build a flexible nonlinear nonstationary model for a financial index and compare its forecast ability in contrast to the random walk hypothesis.

Table 1: Results for the 9 nonlinear and nonstationary datasets with two regimes in levels and three regimes in time and different selections of autoregressive parameters with 100 replications each

	dataset:	1	2	3	4	5	6	7	8	9
STAR in time - STAR in levels		53	63	59	58	58	59	63	65	63
other nonlinear, nonstationary		45	34	37	42	42	41	37	34	36
stationary or linear		2	3	4	0	0	0	0	1	1
time regimes	1	2	2	4	0	0	0	0	1	1
	2	11	5	4	5	7	8	3	8	5
	3	64	70	72	68	68	70	77	70	73
	4	23	23	20	27	25	22	20	21	21
level regimes	1	1	2	2	0	0	0	0	1	1
	2	91	94	97	98	96	94	98	95	96
	3	8	4	1	2	4	6	2	4	3

5.1 Yearly Temperature data

We take into account the series of the average yearly temperature recorded at the Brera Observatory in Milan from 1763 to 2007 [1]. The data are plotted in figure 2 (continuous line, Celsius Degrees $\times 10$). It is apparent that there is a positive trend in the last part of the series.

When trying to fit an ARIMA model, the minimum values of several identification criteria (AIC, Schwartz, Hannan and Quinn) are attained selecting an IMA (1,1) model, and the same suggestion is given by the SCA expert system (Liu and Hudak, 1992).

The estimated model is

$$X_t - X_{t-1} = 0.083 + \epsilon_t - 0.836\epsilon_{t-1} \tag{12}$$

The application of our genetic algorithm suggests as an optimal model a linear structure with nonstationary features, and precisely a PLTAR model in time with two regimes, the threshold being around 1893, and a particularly simple first order form: $X_t = \phi_0(t) + \phi_1(t)X_{t-1} + \epsilon_t$.

The estimated parameters are as follows:

$$\phi_0(t) = 144.38 - 44.68\frac{t}{N} - 91.68\frac{\max\{0, t-123\}}{N} \tag{13}$$

$$\phi_1(t) = -0.12 + 0.29\frac{t}{N} + 0.91\frac{\max\{0, t-123\}}{N} \tag{14}$$

where $N = 245$ is the series length.

In the first regime the model is moderately nonstationary with a slowly decreasing local level and a small autoregressive parameter, while in the second regime the autoregressive parameter increases considerably, and also the local

[1] source: data sets of project HISTALP - Historical Instrumental Climatological Surface Time Series of the greater Alpine region, available at http://www.zamg.ac.at/histalp

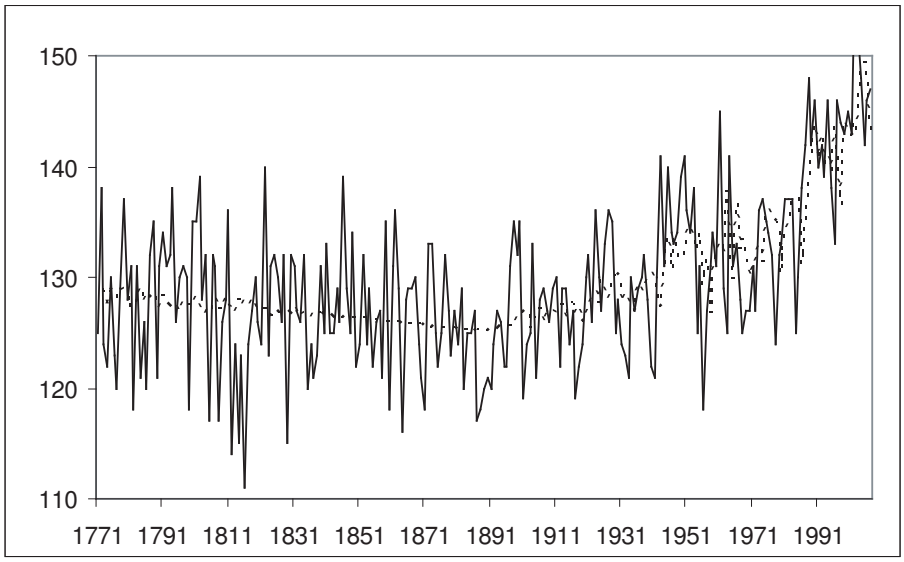

Figure 2: Average yearly temperature recorded at the Brera Observatory in Milan,1771–2007

level increases (as measured by $\frac{\phi_0(t)}{1-\phi_1(t)}$). The fitted data is presented in figure 2 (dotted line).

The residual variance is slightly less than 30 over the entire period, and also separately in each of the two regimes the fitting is better than the IMA (1,1) model.

The proposed model implies an increasing temperature and also the speed of the increase is represented as growing.

Closer details may be found in figure 3 where the last fifty years are shown, together with the fitted data according to the PLTAR model (continuous line) and the IMA(1,1) model (dotted line). The closer fitting ability of the nonstationary model is apparent. The figure also includes, for some future years, the eventual forecast function of the two models, indicating that the temperature increase rate predicted by the PLTAR model is larger than that implied by the integrated moving average.

5.2 Dow Jones Industrial Average Index

The daily closure values of the Dow Jones Industrial Average Index ($\wedge DJL$) in years 2005-2009 are analyzed here (source: Yahoo finance data sets). The data is plotted in figure 4. We used the data from January 2005 to the end of September 2009 for building the model, and the last time span (October, November and December 2009) was employed for out-of-range forecasts.

The series shows an instability in mean. The overall average is 11099 and

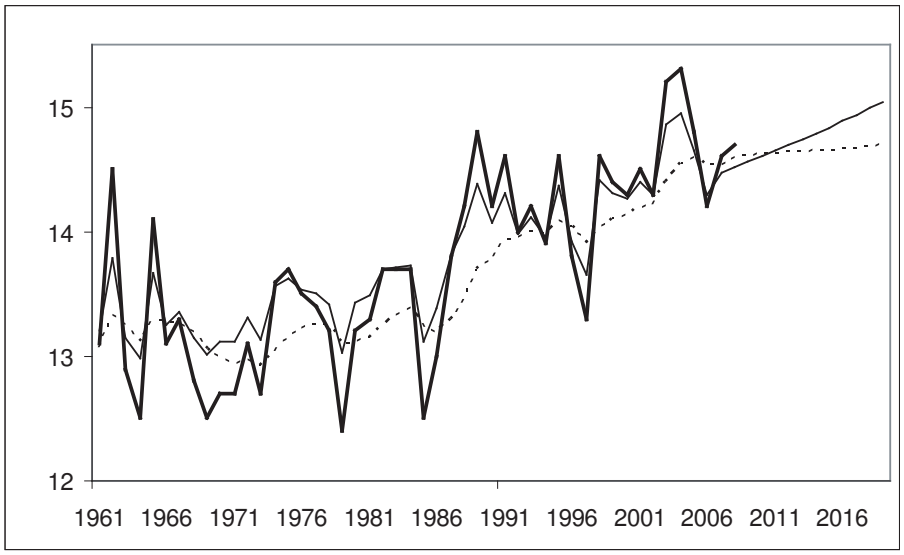

Figure 3: Yearly temperature data, last fifty years and eventual forecasts

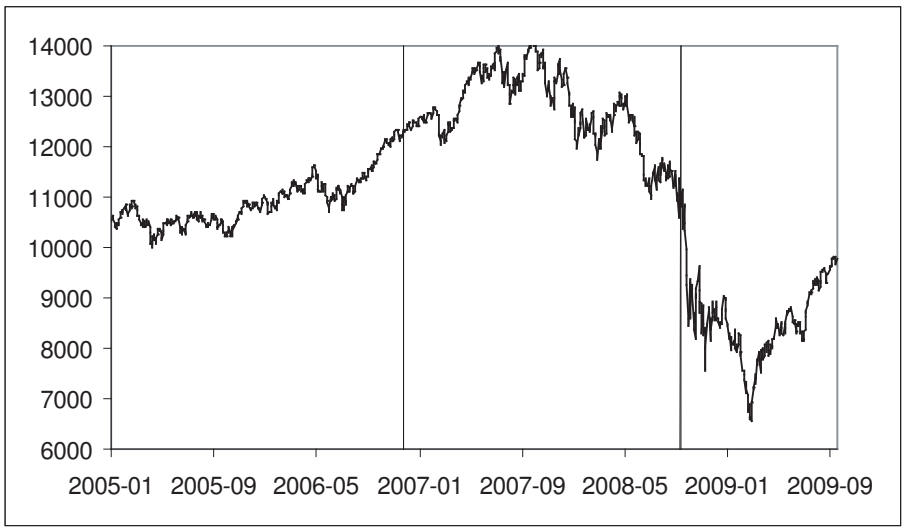

Figure 4: Daily closure values Dow Jones Industrial Average Index ($\wedge DJL$), years 2005-2009

Table 2: Number of observations in each regime

level regime:	≤ 10127	$10127 - 11043$	> 11043	total
Time regime: 1-478	4	298	175	477
479-929	0	7	445	452
930-1186	248	8	1	257
total	252	313	621	1186

the variance 2816919. The first difference series has an average -0.82 and variance 19417. A linear ARIMA analysis does not strongly contradict the random walk hypothesis that the first differences are white noise. The minimum AIC estimated model is an ARIMA (3,1,0) model:

$$\nabla x_t = -0.13\nabla x_{t-1} - 0.1\nabla x_{t-2} + 0.07\nabla x_{t-3} + \epsilon_t \tag{15}$$

but its residual variance is only very slightly smaller than that of the first difference series ($R^2 = 0.03$).

We used the genetic algorithm for building a multi-regime model with Dow Jones data; we analyzed both original data, first differences and the series of returns ($\log\{x_t/x_{t-1}\}$), but the best fitting results (in terms of R^2) were obtained for the original data [2], and these only will be reported here.

The best obtained model is nonlinear and nonstationary with a smooth transition STAR structure in both domains. As far as time is concerned, three regimes have been obtained, threshold being at about the end of November, 2006, and at about the middle of September 2008 (drawn as vertical dotted lines in figure 4). The estimated gamma coefficient in time is small, with respect to the range allowed in the estimation procedure, denoting a relatively slow transition speed between regimes in time.

For the nonlinear behavior, we also obtained three suggested regimes, with thresholds equal to 10127 and 11043. The estimated delay in the driving variable is 5, indicating essentially a weekly influence. The overall residual variance of the STAR-STAR model was 15109, and the selected orders were one in time and two in levels. It has to be noted, however, that the regimes in time and levels interact: the number of observations falling in each combination (time, levels regime) are reported in Table 2.

It may be concluded that in the first time regime the suggested model is a STAR with two level regimes, while in the second and third time span the data is essentially modelled by two different linear models. After a slight simplification obtained by eliminating some small (non significant) parameters, the suggested model is as follows:

- first time regime
  $X_t = [0.96X_{t-1} + 50 + 0.06X_{t-2} + 0.12X_{t-3}][1 - G(X_{t-5})]$

[2] Note that the qualitative difference between models on original or differenced data resides essentially in the choice of the driving variable for the regimes in levels

Table 3: Variance of the model residuals and the differenced data for each time regime

Time regime:	1–478	479–929	930–1186	total
model residuals	67.9	132.8	172.0	122.9
differenced data	69.4	135.3	218.5	139.4

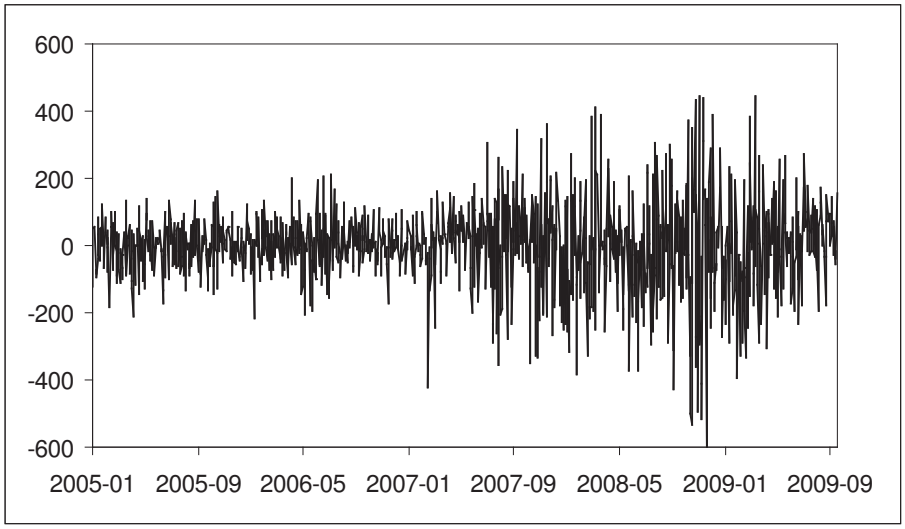

Figure 5: Model Residuals for Dow Jones Data.

$$+ [X_{t-1} - 171 - 0.1X_{t-2} + 0.13X_{t-3} - 0.06X_{t-4}]G(X_{t-5}) + \epsilon_t$$

- second time regime $X_t = 199 + 0.86X_{t-1} + 0.11X_{t-2} + \epsilon_t$

- third time regime $X_t = 256 + 0.86X_{t-1} + 0.07X_{t-2} + 0.2X_{t-3} - 0.16X_{t-4} + \epsilon_t$

with $G(x) = [1 + \exp\{-0.36(x - 10127)\}]^{-1}$.

This model is uniformly superior to the first differences series (corresponding to the random walk hypothesis) as may be seen on breaking down the residual variance by regimes, figures are reported in Table 3:

The model residuals are reported in figure 5. It may be seen that the residuals exhibit heteroskedasticity and the variance in the last regime looks considerably larger than in the previous ones; a LM-ARCH test on the entire square residual series is highly significant. However, if we consider separately the three sub-series related to different time regimes, for the first ($t = 1, 478$) and second span ($t = 479, 929$) the LM-ARCH test is not significant (p-value > 0.2),

whereas for the third sub-series it is significant (p-value 0.002). Therefore we can conclude that for the first and second regime no heteroskedasticity correction appears necessary, while the most recent data seem more heteroskedastic. We have fitted to the last regime a GARCH(1,1) model:

$$h_t = 80.1 + 0.926 h_{t-1} + 0.07 \epsilon_{t-1}^2 + \epsilon_t^2 \tag{16}$$

The standardized residuals of this model look satisfactorily stable, their asymmetry index is -0.25, and kurtosis is 3.23. The Jarque-Bera test does not reject the normality hypothesis (p-value 0.16) and also a LM-ARCH test for remaining heteroskedasticity is not significant (with p-value around 0.5).

To complete the analysis, we used the proposed genetic algorithm for building a multi-regime model on the squared residual series, but with driving variable equal to the original series. The resulting model was a STAR in time and STAR in levels, with a R^2 of about 0.30. Three regimes in time were detected, with thresholds 474 and 714: the first threshold is equal to that of the previous model, while the second one is much before, since the second regime of the raw data series starts at $t = 929$. However, the local fitness on varying the second threshold between 714 and 929 is relatively flat: the range is from 113.2 to 113.3 (for comparison, models with only two regimes in time have fitness less than 108). This suggests a rather slow, or non monotone transition from the second time regime to the third one.

As far as levels are concerned, the model has two regimes with threshold 9665. In this case also there is an interaction between time and level regimes; in the first and second regime there are observations only in the upper level regime, while for the third time period the data is nearly equally divided between the upper and lower level regime. As a result, the behavior implied by the model (here also neglecting some very small parameters) is the following:

- First time regime ($t = 1, 474$):
  $\epsilon_t^2 = 0.36 + 0.12 \epsilon_{t-1}^2 + u_t$
- Second time regime ($t = 415, 714$):
  $\epsilon_t^2 = 74.5 + 0.16 \epsilon_{t-4}^2 + 0.12 \epsilon_{t-5} + u_t$
- Third time regime ($t = 715, 1156$):
  $\epsilon_t^2 = [106.7 - 0.19 \epsilon_{t-1}^2 + 0.12 \epsilon_{t-4} + 0.17 \epsilon_{t-5} + 0.18 \epsilon_{t-6}][1 - G_\epsilon(X_{t-5})]$
  $+ [245.6 - 0.1 \epsilon_{t-1}^2] G_\epsilon(X_{t-5}) + u_t$

where $G_\epsilon(x) = [1 + \exp\{-0.88(x - 9665)\}]^{-1}$.

This model accounts for a slight heteroskedasticity in the first and second time regime, and for stronger volatility changes in the third period, which are in this case modelled as a two-regime smooth transition mechanism.

As an overall result, we may see that a multi-regime nonlinear and nonstationary model may explain dependence in the data, that cannot be accounted for by linear models, and in the case of Dow Jones there is also an indication of an evolutionary behavior involving both the levels and the volatility dynamics.

Table 4: Average observed square forecast errors, lead 1 to 5, observations from Oct 1 2009 to Dec 24 2009

lead:	1	2	3	4	5
multi-regime model	9102	15047	20320	24583	26823
random walk	9129	15834	23557	29984	34456

Finally, to check possible advantages in terms of forecast ability, we have considered the last three months of 2009 for computing out-of-sample forecasts. We computed the lead-1 to lead-5 forecasts for each day from October 1, 2009 to December 24, 2009 using the identified model, which for the third regime is a simple linear autoregressive:

$$X_t = 256 + 0.865X_{t-1} + 0.069X_{t-2} + 0.205X_{t-3} - 0.163X_{t-4} + \epsilon_t \qquad (17)$$

Table 4 reports the average square forecast errors, and those obtained by a random walk, for comparison. It may be seen that the error is similar for lead-1 forecasts, but the multi-regime model shows an increasing advantage as the horizon increases, up to a percentage reduction about 22% for lead-5 forecasts.

6 Conclusions

The findings on simulated and real time series indicate that the proposed algorithm is able to suggest multi-regime models that may reproduce nonlinear and nonstationary features, which could not be discovered with the most popular ARMA methodology. The simultaneous dependence on time and levels permits more efficient and parsimonious models. In addition, splitting the entire time span into several regimes allows to fit the most recent data more closely, providing better out-of-sample forecasts.

For heteroskedastic series, the multi-regime fitting procedure may be applied both to conditional mean and to conditional variance, helping to discover if and how a change in the level dynamics influences the volatility, and vice versa.

The space of solutions to be explored by the genetic algorithm is in our case high-dimensional; therefore a large number of generations (with respect to other implementations in similar statistical problems) is recommended. A possible alternative option would be a larger population size, but this still requires more practical experience.

Finally, an important issue is the choice of the fitness function. Choosing a fitness form inversely linked to penalized gaussian likelihood provides a flexible way of incorporating, through the choice of the tuning constant, both particular constraints on the number of regimes or the autoregressive order, and possibly prior knowledge.

Acknowledgements

This work was supported by the European Commission through Marie Curie Research and Training Network COMISEF Computational Methods in Statistics, Econometrics and Finance, and by Italian Ministry of Education through a national research grant PRIN2007 "Evolutionary Computation in Statistics".

Appendix

On writing the models in state dependent form:

$$X_t = \sum_{j=0}^{p} \phi_j(t, X_{t-d}) X_{t-j} + \epsilon_t \tag{18}$$

the autoregressive coefficients may be written in a weighted form:

$$\phi_j(t,x) = \sum_{i=1}^{r_T^*} \sum_{k=1}^{r_L^*} \beta_j(i,k) c_{ik}(t,x) \tag{19}$$

where the β's are unknown parameters and the coefficients c_{ik} are essentially given by the product of a level transition function $G_{k-1}(x)$ or $S_{k-1}(x)$, and a time transition function $G'_{i-1}(t)$ or $S'_{i-1}(t)$. We define first two indicator vectors, one for levels $\mathbf{z}_L$ and one for time $\mathbf{z}_T$:

$$\mathbf{z}_L = (z_{0L}, z_{sL}, z_{pL})' \quad \mathbf{z}_T = (z_{0T}, z_{sT}, z_{pT})'$$

depending on the model type as follows: $z_{0L} = 1$ if the model is linear, and zero otherwise, $z_{sL} = 1$ if the model is STAR in levels and zero otherwise, and $z_{pL} = 1$ if the model is PLTAR in levels and zero otherwise. We define z_T's in a similar way, relating to the model type in time . The transition functions may be put in matrix form as follows

$$v_i(t) = [0, G'_{i-1}(t), S'_{i-1}(t)]' \quad w_k(x) = [0, G_{k-1}(x), S_{k-1}(x)]'$$

with $v_1(t) = [111]'$, $v_{r_{T+1}}(t) = [0, 0, S'_{r_T}(t)]'$, $w_1(x) = [111]'$, and $w_{r_{L+1}}(x) = [0, 0, S_{r_L}(x)]'$. In this way, the correct factor for each coefficient c_{ik} and each model type is obtained from the scalar products $w_k(x)' z_L$, in levels and $v_i(t)' z_T$ in time. Rearranging in a matrix form:

$$V(t) = [v_1(t), v_2(t), \ldots, v_{r_{T+1}}(t)] \quad W(x) = [w_1(x), w_2(x), \ldots, w_{r_{T+1}}(x)]$$

the matrix $C(t,x)$ of the coefficients may be written:

$$C(t,x) = V(t)' z_T z'_L W(x) \tag{20}$$

On defining also the matrices of parameters $B_j = \{\beta_j(i,k)\}$ we finally obtain

$$\phi(t,x) = \mathrm{tr}\{B_j C(t,x)'\} = \mathrm{tr}\{B_j W(x)' z_L z'_T V(t)\}.$$

References

Alander, J. T. (1992), "On Optimal Population Size of Genetic Algorithms," in *Proc. CompEuro92*, IEEE Computer Society Press, pp. 65–70.

Bai, J., and Peron, P. (1998), "Estimating and Testing Linear Models with Multiple Structural Changes" *Econometrica*, 66, 47–78.

Baragona, R., Battaglia, F., and Calzini, C. (2001), "Genetic Algorithms for the Identification of Additive and Innovation Outliers in Time Series," *Computational Statistics & Data Analysis*, 37, 1–12.

Baragona, R., Battaglia, F., and Cucina, D. (2004), "Fitting Piecewise Linear Threshold Autoregressive Models by Means of Genetic Algorithms," *Computational Statistics & Data Analysis*, 47, 277–295.

Battaglia, F., and Protopapas, M.K. (2009), "Time-varying Multi-regime Models Fitting by Genetic Algorithms," COMISEF *Working Paper Series*, WP09, http://www.comisef.eu/files/wps009.pdf

Bhansali, R. J. and Downham, D. Y. (1977), "Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's EPF Criterion," *Biometrika*, 64, 547–551.

Carrasco, M. (2002), "Misspecified Structural Change, Thresholds and Markov-Switching Models", *Journal of Econometrics*, 109, 239–273.

Chatterjee, S., Laudato, M., and Lynch, L. A. (1996), "Genetic Algorithms and Their Statistical Applications: an Introduction," *Computational Statistics & Data Analysis*, 22, 633–651.

Crawford, K. D. and Wainwright, R. L. (1995), "Applying Genetic Algorithms to Outlier Detection," in *Proceedings of the Sixth International Conference on Genetic Algorithms*, ed. Eshelman, L. J., San Mateo, CA: Morgan Kaufmann, pp. 546–550.

Davis, R., Lee, T., and Rodriguez-Yam, G. (2006), "Structural Break Estimation for Nonstationary Time Series Models," *Journal of the American Statistical Association*, 101, 223–239.

Dupleich Ulloa, M.R. (2005), "Testing for Breaks and Threshold Effects: A Nested and Non-Nested Approach", *Technical Report, University of Cambridge*, http://www-cfap.jbs.cam.ac.uk/events/files/ulloa.pdf

Gaetan, C. (2000), "Subset ARMA Model Identification Using Genetic Algorithms," *Journal of Time Series Analysis*, 21, 559–570.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Reading, MA: Addison-Wesley.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Press.

Koop, G., Potter, S. (2001), "Are Apparent Finding of NonLinearity due to Structural Instability in Economic Time Series?", *Econometrics Journal*, 4, 37–55.

Lin, C. and Teräsvirta, T. (1994), "Testing the Constancy of Regression Parameters Against Continuous Structural Change," *Journal of Econometrics*, 62, 211–228.

Liu, L., and Hudak, G. B. (1992), *Forecasting and Time Series Analysis Using the SCA Statistical System*, Oak Brook: Scientific Computing Associates.

Lundberg, S., Teräsvirta, T., and van Dijk, D. (2003), "Time-Varying Smooth Transition Autoregressive Models," *Journal of Business and Economic Statistics*, 21, 104–121.

Priestley, M. B. (1988), *Non-linear and Non-stationary Time Series Analysis*, London: Academic Press.

Reeves, C. R. (1993), *Modern Heuristic Techniques for Combinatorial Problems*, New York: Wiley.

Rissanen, J. (2007), *Information and Complexity in Statistical Models*, Berlin: Springer.

Rudolph, G. (1997), *Convergence Properties of Evolutionary Algorithms*, Hamburg: Verlag Dr. Kovač.

Teräsvirta, T. (1994), "Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association*, 89, 208–218.

Teräsvirta, T. (1998), "Modeling Economic Relationships with Smooth Transition Regressions," in *Handbook of Applied Economic Statistics*, eds. Ullah, A. and Giles, D. E. A., New York: Marcel Dekker, pp. 507–552.

Tong, H. (1990), *Non Linear Time Series: A Dynamical System Approach*, Oxford: Oxford University Press.

Wu, B. and Chang, C.-L. (2002), "Using Genetic Algorithms to Parameters (d,r) Estimation for Threshold Autoregressive Models," *Computational Statistics & Data Analysis*, 38, 315–330.