Generalized periodic autoregressive models for trend and seasonality varying time series

Francesco Battaglia and Domenico Cucina and Manuel Rizzo

Abstract Many nonstationary time series exhibit changes in the trend and seasonality structure, that may be modeled by splitting the time axis into different regimes. We propose multi-regime models where, inside each regime, the trend is linear and seasonality is explained by a Periodic Autoregressive model. In addition, for achieving parsimony, we allow season grouping, i.e. seasons may consist of one, two, or more consecutive observations. Identification is obtained by means of a Genetic Algorithm that minimizes an identification criterion.

Key words: Genetic algorithms, Structural break, Regime change

1 Introduction

Many seasonal time series exhibit an autocorrelation structure which depends not only on the time between observations but also on the season of the year. Moreover, the time series of observations within a given season is usually second order stationary (Hipel and McLeod, 1994). In order to model appropriately these and similar types of time series, Periodic AutoRegressive models (PAR) can be employed. When fitting a PAR model to periodic time series a separate AR model for each season of the year is estimated. These models are appropriate for describing time series drawn from different areas such as economics, hydrology, climatology and

Francesco Battaglia
Department of Statistical Sciences, University La Sapienza, Rome, Italy e-mail: francesco.battaglia@uniroma1.it

Domenico Cucina
Department of Economics and Statistics, University of Salerno, Italy e-mail: dcucina@unisa.it

Manuel Rizzo
Department of Statistical Sciences, University La Sapienza, Rome, Italy e-mail: manuel.rizzo@uniroma1.it

signal processing (e. g. Franses and Paap, 2004; Hipel and McLeod, 1994; Ursu and Turkman, 2012).

In this study we consider a generalization of PAR models with linear trend in two directions. First, the model may follow different regimes in time, and regime changes may occur at any time. The regime changes may affect the linear trend, the seasonal means and the autoregressive parameters. We also allow a discontinuous trend which can identify changes in level. Second, inside each regime the model structure may be different for each seasonal position (e.g. months) or vary more slowly, changing only according to grouped seasons like quarters or semesters.

The number of regimes and change times (or break points) are assumed to be unknown. The problem of their identification can be treated as a statistical model selection problem according to a specified identification criterion (Aue and Horväth, 2013). This approach has been used for identification of structural breaks e.g. in Davis et al (2008) and Lu et al (2010). In these works Genetic Algorithms (GAs) are proposed to solve the selection problem.

To the best of our knowledge, there are no articles that handle the changing parameters and changing trend problem in PAR models simultaneously. We propose a class of GAs to detect the number of regimes of piecewise PAR models and their locations. Our procedure evaluates several regime patterns where the locations that are possibly change times are simultaneously considered. In this way, GAs deal efficiently with the detection of multiple change times. We also allow subset AR models to be selected. Each piecewise subset PAR configuration is evaluated by an AIC identification criterion.

In our paper, since the seasonal effect on means, variances and correlations may show different speed and pattern, we propose to join appropriately season parameters into groups for each of these three features.

The piecewise linear nature of our model makes forecasting very simple.

2 The Model

Suppose that a time series $\{X_t\}$ of N observations is available. The seasonal period of the series is s and is assumed to be known. Assume that there are $m+1$ different regimes, separated by m change times τ_j so that the first regime contains observations from time 1 to $\tau_1 - 1$, the second regime contains data from time τ_1 to $\tau_2 - 1$, the $(j+1)$-th regime contains data from τ_j to $\tau_{j+1} - 1$, and the last regime data from τ_m to N. To ensure reasonable estimates we assume that the minimum regime length is a fixed constant mrl, thus any regime assignment is defined by the set $\{\tau_j, j = 1, \ldots, m\}$ subject to $mrl < \tau_1 < \tau_2 < \ldots < \tau_m < N - mrl$, $\tau_j \geq \tau_{j-1} + mrl, j = 2, \ldots, m$.

The parameters of the model for regime j will be denoted by a superscript (j).

The seasonal effect on means, variances and correlations may show different speed and pattern, thus it seems advisable, for each of these features, to use a different splitting of the year, determined by a different length of the season inside which that feature remains constant. For example, if the seasonal period is s, we may have exactly s different models, one for each seasonal position; or rather only s/c different structures, when c consecutive observations are supposed to belong to the same season. E. g. if $s = 12$ (monthly data) and $c = 3$, the same model works for each quarter, therefore there are only $s/c = 4$ different seasons. This may be useful when the seasonal variation is slow and more detailed models would be redundant.

We allow a different season grouping for means, correlation and variance. We denote by c_M the number of consecutive observations for which the mean remains constant (c_M divides s), and by $ss = s/c_M$ the number of seasons. In an analogue fashion, we denote by c_{AR} the number of consecutive observations for which the AR parameters remain constant, and $sv = s/c_{AR}$ the related number of seasons. E. g. for $s = 12$, if $c_{AR} = 1$, each month has a different set of AR parameters, then the variances of the 12 months may a priori be different. If on the contrary $c_{AR} > 1$, the variances of c_{AR} contiguous observation all are proportional, through the same coefficient, to the residual variances, thus a variance instability is equivalent to a residual variance instability. Therefore, we allow the possibility that, inside each single season for the AR model (containing c_{AR} consecutive observations) the residual variances may change. Thus we must consider sub-seasons composed by c_V observations, where c_V divides c_{AR}, and allow the residual variance to change every c_V observations, in a total number of seasons (concerning the residual variance) equal to $svar = s/c_V$.

A linear trend and a different mean for each season is assumed. The residuals are treated as zero mean and described by an autoregressive model with maximum order p, and parameters varying with seasons. Let k_t denote the season (for the mean) of the t-th observation ($1 \le k_t \le ss$) and k_t^* the season for AR structure of the t-th observation, denote by $a^{(j)} + b^{(j)}t$ the linear trend in regime j, by $\mu^{(j)}(k)$ the mean of season k in regime j, and by $\phi_k^{(j)}(i)$ the lag-i autoregressive parameter for the model in season k and regime j. Then for $\tau_{j-1} \le t < \tau_j$:

$$X_t = a^{(j)} + b^{(j)}t + \mu^{(j)}(k_t) + W_t \ , \ W_t = \sum_{i=1}^{p} \phi_{k_t^*}^{(j)}(i) W_{t-i} + \varepsilon_t$$

where $\tau_0 = 1$ and $\tau_{m+1} = N + 1$.

The innovations ε are supposed independent and zero-mean, with variances $\sigma^2(j,k)$ possibly depending on the regime and season.

As far as subset selection is concerned, we introduce also $m+1$ binary vectors $\delta^1, \ldots, \delta^{m+1}$, which specify presence or absence of autoregressive parameters in each regime as follows: if $\delta^j[p(k_t^* - 1) + i] = 1$ then $\phi_{k_t^*}^{(j)}(i)$ is constrained to zero. In summary, a model is identified by the following:

External parameters (fixed and equal for all models) N, s, maximum order p, maximum number of regimes, and minimum number of observations per regime *mrl*

Structural parameters (determining the model structure)

m	number of change times
$\tau_1, \tau_2, \dots, \tau_m$	change times or thresholds
$\delta^1, ..., \delta^{m+1}$	denote which ϕ's are zero in each regime and season
c_M, c_{AR}, c_V	season grouping parameters subject to constraints: c_M divides s; c_{AR} divides s; c_V divides c_{AR}

Regression parameters to be estimated by Least Squares (LS) or Maximum Likelihood (ML)

$a_1, a_2, \dots, a_{m+1}$	intercepts
$b_1, b_2, \dots, b_{m+1}$	slopes
$\mu^{(j)}(k)$	seasonal means, $k = 1, \dots, ss$; $j = 1, \dots, m+1$
$\phi_k^{(j)}(i)$	AR parameters, $k = 1, \dots, sv$; $j = 1, \dots, m+1$; $i = 1, \dots, p$ (some of them may be constrained to zero)
$\sigma^2(j,k)$	innovation variances, regime j and season $k = 1, \dots, svar$.

For estimating trend and seasonal means by LS, note that the intercept and the means are linearly dependent, therefore we assume that the seasonal means sum to zero on one cycle: $\mu^{(j)}(1) + \mu^{(j)}(2) + \ldots + \mu^{(j)}(ss) = 0, \forall j$. Therefore the following equations are estimated:

$$X_t = b^{(j)} t + c(j, k_t) \;\;,\;\; \tau_j \le t < \tau_{j+1} \tag{1}$$

and then the parameter vector is $\beta' = \{b^{(1)}, b^{(2)}, \dots, b^{(m+1)}, c(1,1), c(1,2), \dots, c(1,ss), c(2,1), \dots, c(2,ss), \dots, c(m+1,1), \dots, c(m+1,ss)\}$ with dimension $(m+1) \times (ss+1)$ and the estimates are obtained by least squares. From the $\{\hat{c}(j,k)\}$, the intercepts $\hat{a}^{(j)}$ and seasonal means $\hat{\mu}^{(j)}(k)$ are recovered basing on the above assumption. It follows

$$\hat{a}^{(j)} = \frac{1}{ss} \sum_{k=1}^{ss} \hat{c}(j,k) \;\;,\;\; \hat{\mu}^{(j)}(k) = \hat{c}(j,k) - \hat{a}^{(j)}.$$

Moreover it is possible to prescribe trend continuity by imposing that, if the number of regimes is larger than one, the trend values of two consecutive regimes coincide on the first observation of the second regime. A possible level change at $t = \tau_{j+1}$ is estimated if the trend continuity is not imposed.

Conditioning on thresholds, seasonal arrangement and estimated trend and means, the residual series is computed as $\hat{W}_t = X_t - \hat{a}^{(j)} - \hat{b}^{(j)} t - \hat{\mu}^{(j)}(k_t)$.

For each regime and season a separate autoregressive process is considered:

$$\hat{W}_t = \sum_{i=1}^{p} \phi_{k_t^*}^{(j)}(i) \hat{W}_{t-i} + \varepsilon_t.$$

We denote by $I(j,k)$ the set of times belonging to regime j and season k. The corresponding observations $z_{j,k}$ are selected and the LS estimates of the parameters $\{\phi_k^{(j)}(i), i = 1, \dots, p\}$ are obtained. As far as subset selection is concerned, the final estimates are obtained via LS constrained optimization, with constraints given by linear system $H\phi = 0$:

$$\hat{\phi} = \phi_{LS} - (Z'Z)^{-1}H'[H(Z'Z)^{-1}H']^{-1}H\phi_{LS},$$

where $\phi_{LS} = (Z'Z)^{-1}Z'z$ are the unconstrained least squares estimates, Z is the $n_{j,k} \times p$ design matrix including lagged observations and H is the constraints matrix that specifies subset models. The residuals $e = z - Z\hat{\phi}$ give the estimate of the innovations for regime j and season k $\{\varepsilon_t, t \in I(j,k)\}$, which allow to obtain $\hat{\sigma}^2(j,k)$. The structural parameters take discrete values and their combinations amount to a very large number. GAs are naturally suitable for the choice of optimal structural parameters.

3 Genetic algorithms

GAs, initially developed by Holland (1975), imitate the evolution process of biological systems, to optimize a given function. A GA uses a set of candidate solutions, called *population*, instead of one single current solution. In GA terminology, any candidate solution is encoded via a numerical vector called *chromosome*. The GA proceeds iteratively by updating the population in rounds, called generations. In each generation, some of the active chromosomes are selected (parents-chromosomes) to form the chromosomes of the next generation (children-chromosomes). The selection process is based on an evaluation measure called *fitness function*, linked to the objective function, that assigns to each chromosome a positive number. Children are formed by recombining (*crossover*) the genetic material of their two parents-chromosomes and perhaps after a random alteration of some of the genes (single digits of the chromosome), which is called *mutation* (see Holland, 1975; Goldberg, 1989, for a detailed description).

A successful implementation of GAs is certainly crucial to obtain satisfactory results. Before a GA can be applied to a problem some important decisions have to be made. The GA methods require a suitable encoding for the problem and an appropriate definition of objective function. In addition operators of selection, crossover and mutation have to be chosen.

Encoding. An appropriate encoding scheme is a key issue for GAs. It must guarantee an efficient coding producing no illegal chromosome and no redundancy. Details of the adopted method may be found in Battaglia et al (2018).

Fitness function. The most natural objective in building statistical models is to minimize an identification criterion such as AIC, BIC, ICOMP, MDL. They all are based on the estimated residual variance $\hat{\sigma}^2(j,k)$ and the total number of estimated parameters: there are $m+1$ parameters for trend, $(m+1) \times ss$ seasonal means, and in regime j there are $p \times sv - |\delta^j|$ autoregressive parameters (where $|x|^2 = \sum_i x_i^2$). So, the total number of estimated parameters is $P = (m+1)(ss+1) + (m+1)p \times sv - |\delta^1| - |\delta^2| - \ldots - |\delta^{m+1}|$.

If continuity constraints on trend are added, the number of parameters decreases by m.

The most obvious generalization of AIC is the NAIC criterion introduced by Tong (1990, p. 379) for threshold models:

$$NAIC = [\sum_j \sum_k AIC_{j,k}]/N = \left[\sum_j \sum_k n_{j,k} \log \hat{\sigma}^2(j,k) + \pi \times P\right]/N,$$

where $AIC_{j,k}$ is identification criterion for series of regime j and season k, $\sigma^2(j,k)$ is the related residual variance, P is the total number of parameters, π is the penalization term (equal to 2 in the original Akaike's proposal). Other alternatives are possible (see Battaglia et al, 2018).

Since the identification criteria are to be minimized, the fitness function is a monotonically decreasing function of the identification criterion. We adopted a negative exponential transformation.
GA operators. For selection we used the "roulette wheel" rule where the probability of a chromosome being selected as a parent is proportional to its fitness. Each selected couple of parents will produce two "children" by methods of crossover and mutation. We implemented uniform crossover — each child receives each gene from one parent or the other randomly with probability $1/2$.

The entire population of chromosomes is replaced by the offsprings created by the crossover and mutation processes at each generation except for the best chromosome, which survives to the next generation. This *elitist* strategy ensures that the fitness will never decrease through generations (Rudolph, 1994).

Our search strategy is in two steps: in the first one the GA tries to determine the best splitting in regimes for complete models, as the chromosome includes only $m, \tau_1, ..., \tau_m$; in the second step we exhaustively enumerate all possible seasons grouping (specified by c_M, c_{AR}, c_V) and subset models (examining $\delta^1, ..., \delta^{m+1}$). This strategy is hybrid, as it combines an exact method with an approximate method, and it is feasible if order p and seasonal period s are not too large.

4 Applications

We briefly summarize the application of our method to the CET series of monthly mean surface temperatures for a location in the Midlands region (see Proietti and Hillebrand, 2017, and references therein). It is a very long and frequently investigated series (we use 2904 observations for years 1772–2013). Many researchers suggest an upward trend since the beginning of the 20th century, due to global warming, and an evolution of seasonal pattern, identified as a precession of Earth's axis of rotation. Four PAR models were fitted to the time series: a *complete* model, with a different AR(2) model for each month and no constraint on the autoregressive parameters; a *subset* model similar to the previous one, but with some AR parameters constrained to zero in order to maximize fitness; a *grouped* subset model where the season are grouped; and finally a *constant* seasonality model, subset as well, where the autoregressive parameters remain equal in each regime. The series plot appears in Figure 1 (left panel).

Two regimes were identified, with change time at July, 1899 (the trend is drawn as a dotted line in the figure). This confirms with clear evidence the suggested trend change at the beginning of the last century. The grouped season model identified the optimal setting $c_M = c_{AR} = 1$, meaning that any grouping of seasons would not increase the fitness, thus the grouped model coincides with the subset model (with other more parsimonious criteria like *BIC* the best grouping results $c_M = 1, c_{AR} = 4, c_V = 2$). The results appear in Table 1: it may be concluded that many autoregressive parameters may be constrained to zero without a sensible loss of fit. Moreover, the smaller fitness of the constant seasons model indicates an evolution in the seasonal pattern; Figure 1 (right panel) reports the monthly means for the two regimes. More applications may be found in Battaglia et al (2018).

Table 1 Models fitted to the CET series

Model	Complete	Subset	Constant seasons
Residual variance	1.696	1.701	1.752
Number of parameters	74	50	30
Fitness	0.594	0.602	0.595

5 Conclusions

In this paper we have proposed models that are able to explain, on one side, regime changes and structural breaks, and on the other side a seasonal behavior that evolves in time.

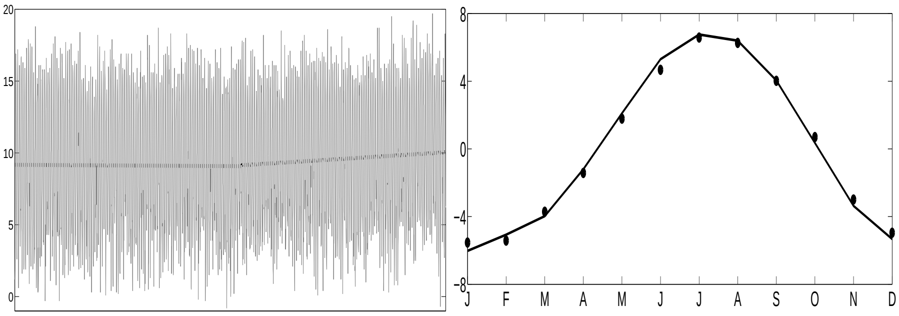

Fig. 1 Left panel: CET series and trend (dotted line). Right panel: monthly means, continuous line: first regime; dotted line: second regime

The complex problem of identifying and estimating such models is solved by GAs. The best model is selected according to a fitness function that is a monotonically decreasing transformation of widely used identification criteria. Experience on real and simulated data suggests that the choice of the fitness function is crucial because a too parsimonious criterion may lead to models that overlook important structure changes.

The results seem to support the usefulness of the proposed methods in detecting relevant changes in the structure of the trend and also possible evolution in the seasonal behavior concerning levels, variance and correlation. The generalized periodic autoregressive models allow a closer analysis of the seasonal behavior, suggesting also the most convenient grouping of seasons in terms of fitness.

References

Aue A, Horväth L (2013) Structural breaks in time series. Journal of Time Series Analysis 34(1):1–16

Battaglia F, Cucina D, Rizzo M (2018) A generalization of periodic autoregressive models for seasonal time series. Tech. Rep. 2, Dept. of Statistical Sciences, University La Sapienza, Rome, Italy, ISBN 2279-798X

Davis R, Lee T, Rodriguez-Yam G (2008) Break detection for a class of nonlinear time series models. Journal of Time Series Analysis 29(5):834–867

Franses PH, Paap R (2004) Periodic Time Series Models. Oxford University Press, New York

Goldberg D (1989) Genetic algorithms in search optimization and machine learning. Addison-Wesley, Reading, MA

Hipel KW, McLeod AI (1994) Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam

Holland J (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and AI. The University of Michigan, Ann Arbor, MI

Lu Q, Lund R, Lee T (2010) An MDL approach to the climate segmentation problem. The Annals of Applied Statistics 4(1):299–319

Proietti T, Hillebrand E (2017) Seasonal changes in central England temperatures. Journal of the Royal Statistical Society A 180:769–791

Rudolph G (1994) Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks 5:96–101

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

Ursu E, Turkman KF (2012) Periodic autoregressive model identification using genetic algorithms. Journal of Time Series Analysis 33:398–405