Identification of multiregime periodic autoregressive models by genetic algorithms

Domenico Cucina, Manuel Rizzo, and Eugen Ursu

Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy
dcucina@unisa.it
Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00100 Rome, Italy
manuel.rizzo@uniroma1.it
GREThA UMR-CNRS 5113, Université de Bordeaux, Avenue Léon Duguit, 33608 Pessac cedex, France
eugen.ursu@u-bordeaux.fr

Abstract. This paper develops a procedure for identifying multiregime Periodic AutoRegressive (PAR) models. In each regime a possibly different PAR model is built, for which changes can be due to the seasonal means, the autocorrelation structure or the variances. Number and locations of changepoints which subdivide the time span are detected by means of Genetic Algorithms (GAs), that optimize an identification criterion. The method is evaluated by means of simulation studies, and is then employed to analyze shrimp fishery data.

Keywords: Seasonality, Structural changes, Genetic algorithm

1 Introduction

This paper is concerned with seasonal time series which may display many discontinuities, that can be specified by changepoints in time (or structural changes). As defined in [16] a changepoint is "a time where the structural pattern of a time series first shifts". In many cases, the changepoints are located at known times and it is easy to take into account their effects. When changepoints are located at unknown times and their features are ignored, the time series estimation can be misleading [20]. In fact, an undetected changepoint can lead to: misinterpretation of the model, biased estimates and less accurate forecasting [9]. Taking all these into account, changepoint detection becomes a demanding job especially if its identification is required soon after occurrence. In the past four decades several techniques have employed for changepoint detection [4, 6, 27]. For a recent review of changepoint analysis in time series see [1].

Periodic time series models have been introduced because standard seasonal autoregressive integrated moving average (SARIMA; [3]) cannot be filtered to achieve second-order stationarity, and this is because the correlation structure of these time series depends on the season [26]. [22] also showed that seasonal differencing maintains the seasonal correlation structure, whereas the periodic term is completely removed by seasonal standardization or by spectral analysis. General overviews of periodic models and their applications are presented in [10, 7].

We shall focus on time series recorded monthly which display periodic dynamics and possible structural changes, which may imply the existence of several regimes in time. Changepoint detection procedures for periodic series have been studied in [17], which focused on mean shifts. In our proposal we allow also the whole model structure to switch at each changepoint time, as far as periodic data usually display both seasonal effects and various kind of discontinuities.

We propose a procedure based on Genetic Algorithms to detect changepoints and estimate resulting PAR models. These kind of methods are well suited for complex global optimization, as they have been widely applied to hardly tractable identification and estimation problems [25, 6]. The procedure is based on an identification criterion, such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) or MDL (Minimum Description Length). The resulting method will also allow to perform subset selection, as we allow intermediate parameters to be constrained to zero. This modification leads to gain in parsimony, and could also contribute to improve forecasting ability of resulting models.

The article is organized as follows: Section 2 describes proposed methodology for model building; in order to illustrate the efficiency of the proposed procedure some simulations are presented in Section 3; an application to French Guiana shrimp fishery is included in Section 4; comments close the paper in Section 5.

2 Methodology

2.1 Model description

We consider a periodic time series of period s, observed for N years and possibly subdivided into M regimes in time. The multiregime PAR model is specified as follows:

$$X_{(n-1)s+k} = a^j + b^j[(n-1)s + k] + Y_{(n-1)s+k}, \tag{1}$$

where $j = 1, ..., M$ is the index of regimes, $k = 1, ..., s$ is the index of periods, a^j and b^j are trend parameters, which may vary with the regime, and μ_k^j are the seasonal means. $X_{(n-1)s+k}$ is referred to the observation in season k of year n ($n = 1, ..., N$), while $Y_{(n-1)s+k}$ follows a PAR given by:

$$Y_{(n-1)s+k} = \sum_{i=1}^{p} \phi_i^j(k) Y_{(n-1)s+k-i} + \epsilon_{(n-1)s+k}, \tag{2}$$

where $i = 1, \ldots, p$ denotes the lag, $\phi_i^j(k)$, $i = 1, ..., p$ indicate the autoregressive parameters of regime j and season k. As far as our identification procedure will allow to identify subset models, the autoregressive parameters can be constrained to zero, in order to get more parsimonious models. The error process is a periodic white noise with $E(\epsilon_{(n-1)s+k}) = 0$ and $\text{var}(\epsilon_{(n-1)s+k}) = \sigma_j^2(k)$, so that also the residual variances are allowed to change in each regime and season. We shall assume that each regime is periodic stationary with period s [18].

The regimes are specified by $M - 1$ changepoint years $\tau_1, ..., \tau_{M-1}$, defined in such a way that τ_{j-1} and $\tau_j - 1$ denote, respectively, the first and the last year of regime j ($j = 1, \ldots, M$). We also assume that $\tau_0 = 1$ and $\tau_M = N + 1$. In order to ensure reasonable estimates, we require that each regime contains at least a minimum number ω of years, therefore $\tau_j \geq \tau_{j-1} + \omega$ for any regime j. For the sake of simplicity we assume that the total number of observations T is a multiple of s ($T = N \times s$).

2.2 Model building

The identification of our multiregime PAR model consists in the choice of changepoints $M-1$, the changepoint times $\tau_1, \ldots, \tau_{M-1}$ and the specification of subset models (we shall assume the same maximum autoregressive order p for all models). The discrete search space is prohibitively large, so we shall base the procedure on GAs [11].

They are a nature-inspired optimization method, often employed when it is required to find an optimal solution from a prohibitively large discrete set. In GAs metaphor, the search strategy is based on the evolution of a population of individuals, coded in binary vectors named chromosomes to suitably represent the problem solutions, towards populations which are better able to adapt to the environment. The goodness of individuals in such populations is called fitness, and it is related to the objective function of the problem at hand. At each iteration (named generation in the GA terminology) the evolution takes place by means of three main operators: the selection, which chooses the individuals that will generate

the offspring; the crossover, that allows pairs of individuals to combine, producing possibly better solutions; mutation, which simulates the rare random changes happening in nature, and facilitates the exploration of the search space. Lastly, the elitist strategy ensures that the best solution is always retained in each generation of the algorithm (for an account on GA operators and strategies see [8]). The flow of generations generally stops when a prefixed criterion is met, for example the reaching of a fixed number of generations.

We shall optimize a fitness function based on an identification criterion, such as AIC, BIC, Hannan-Quinn, that combine a measure of goodness of fit and a penalization on the number of parameters. In particular, we will consider a criterion inspired by the Normalized Akaike's Information Criterion (NAIC), introduced in [24] for threshold models:

$$g = [\sum_{j=1}^{M}\sum_{k=1}^{s} n_{j,k}\log(\hat{\sigma}_j^2(k)) + IC\sum_{j=1}^{M}\sum_{k=1}^{s} P_{j,k}]/T, \tag{3}$$

where $\hat{\sigma}_j^2(k)$ is the model residual variance of series in regime j and season k, $n_{j,k}$ and $P_{j,k}$ are, respectively, sample size and number of parameters of regime j and season k, IC is the penalization term. The choice of IC specifies the magnitude of penalization on number of parameters: for example a value equal to 2 resembles the structure of an AIC, while $IC = ln(N)$ leads to the analogous to BIC criterion. The final fitness f will be a scaled exponential transformation of g: $f = \exp(-g/\beta)$, where β is a problem dependent constant. This is a quite common procedure in GAs [8, 13] as it allows to control the shape of fitness function without changing the solutions ranking.

The fitness evaluation step is carried out conditioning on the model structure of a generic solution, and the model parameters are estimated consequently. These latter are the trend intercepts and slopes a^j and b^j, the seasonal means μ_k^j, the autoregressive parameters $\phi_i^j(k)$ and the residual variances $\sigma_j^2(k)$, for all $j = 1, \ldots, M$; $k = 1, \ldots, s$; $i = 1, \ldots, p$. These parameters are estimated by Ordinary Least Squares, except for the autoregressive ones, which must account also for the subset selection constraints.

With respect to the model structure, we shall adopt the following strategy in our GA: the generic chromosome will binary encode only the regime structure $[M-1, \tau_1, ..., \tau_{M-1}]$. Conditioning on such structure, all the possible 2^p subset autoregressive models will be enumerated and evaluated in the fitness evaluation step, and only the best will be retained.

This is an exact strategy with respect to the subset selection, and it is computationally feasible only if the maximum autoregressive order p is small.

The chromosome encoding works as follows: the first two or three bits (depending on the maximum number of regimes allowed) give the number of changepoints $M-1$; subsequent bit intervals, whose length is custom fixed, produce changepoint times $\tau_1, ..., \tau_{M-1}$. This part of encoding must ensure the constraints on minimum number of years per regime ω. We shall adopt a procedure introduced in [2] for multiregime nonlinear models identification, which allows each chromosome to be legal (the constraints on minimum regime length ω always hold), so there is no computational time wasted on evaluating infeasible solutions.

3 Simulation studies

To illustrate the efficiency of the proposed procedure, we use a set of one thousand simulated series and apply our method to each series. The series will contain a century ($N = 100$) of monthly data ($s = 12$). We consider the following options for the objective function, in order to study the sensitivity of penalization: values of IC equal to 2 and $ln(N)$, which resemble a generalization of AIC and BIC criteria, and also $IC = 3$, successfully adopted in [2] for the identification of nonstationary nonlinear models by GAs. Concerning the choices on GA implementation, we employed a population of 50 solutions and used operators of tournament selection, bit-flip mutation (rate 0.1) and parameterized uniform crossover (rate 0.7). As far as GA as stochastic methods, namely each GA run may lead to a different results, we shall report the correct number of changepoints identification rate, and also the mean and standard deviation of changepoint locations.

Model 1: 1 changepoint with no trend Our first simulation model consists of 1 changepoint located at the end of the 60-th year ($\tau_1 = 61$). For the second regime, we consider that the variances are four times smaller than for the first regime. We consider the same autoregressive structure in both regimes and an order equal to one for the PAR model.

We apply our method to the realization in Figure 1. Table 1 reports empirical frequency distribution of the estimated changepoints. We see that BIC ($IC = \ln N$) has the best percentage rate (100%) of correct identification. The $IC = 3$ has a correct rate of identification of 73%; 19.5% of runs estimate 2 changepoints and 6.2% of runs estimate 3 changepoints.

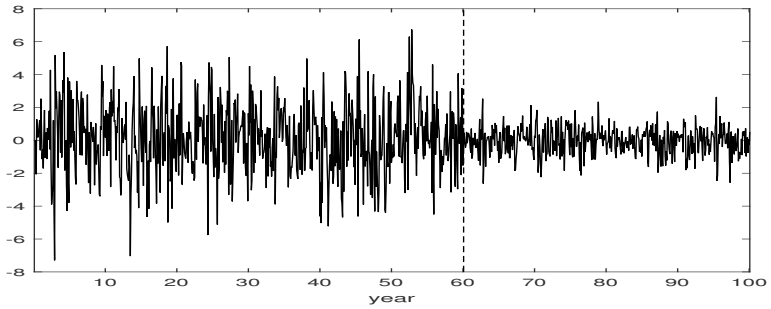

Fig. 1: A realization from the process in *Model 1*. The vertical dash line indicates the true changepoint location at the end of the 60-th year.

The AIC criterion seems to overestimate the number of changepoints as 84.4% of runs estimates 4 changepoints or more.

Number of	IC=ln N			IC=3			IC=2		
changepoints	%	mean	se	%	mean	se	%	mean	se
0									
1	100	60.97	0.68	73.0	60.96	0.59	0.1	24	0
2				19.5	50.40 70.34	15.29 9.05	4.3	48.76 69.86	16.43 9.62
3				6.2	50.90 67.11 81.73	14.50 9.88 8.18	11.3	48.27 65.18 79.92	17.45 12.12 9.73
≥ 4				1.3			84.4		

Table 1: Summary of the estimated changepoints for the *Model 1*. The true number of changepoints is 1.

As for where the changepoints are estimated, the mean and the standard-error of the estimated locations are also reported in Table 1. The proposed procedure performs very well in locating the changepoints for our method combined with BIC.

Model 2: 2 changepoints with different trends

In the second simulation experiment we consider a model with two changes in trend parameters at times $\tau_1 = 31$ and $\tau_2 = 61$. We use a PAR model of order 1 for each regime with same parameters from one regime

to other. For illustrative purposes, Figure 2 shows a typical realization of the model defined above.

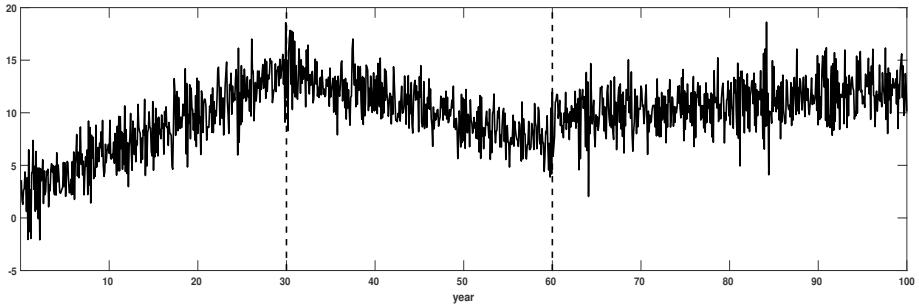

Fig. 2: A realization from the process of *Model 2*. The vertical dash line indicates the true changepoint locations at the end of 30-th and 60-th year.

For our method combined with BIC criterion, the selected number of changepoints is generally equal to 2, while only for 80 of the 1000 analyzed series 3 changepoints were selected and in 15 cases a the number of changepoints greater or equal to 4 was chosen. The $IC = 2$ and $IC = 3$ criteria seem to overestimate the number of changepoints. Lastly, in order to evaluate the changepoints location, we report the mean and the standard-error for estimated locations in Table 2.

Number of	IC=ln N			IC=3			IC=2		
changepoints	%	mean	se	%	mean	se	%	mean	se
0									
1	0.01	32.00	0.00						
2	90.4	30.94 60.97	1.48 0.45	38.1	30.86 60.99	1.43 0.14	1.4	31.00 61.00	1.17 0
3	8.0	27.51 51.05 71.40	6.33 12.58 10.42	22.7	28.07 50.19 70.01	6.23 11.95 9.96	4.1	25.58 44.51 65.68	7.64 12.07 8.45
≥ 4	1.5			26.6			94.5		

Table 2: Summary of the estimated changepoints for the *Model 2*. The true number of changepoints is 2.

4 Data analysis

We illustrate the main findings by analyzing the shrimp French Guiana fishery, a case study that has been accounted in [23]. Two shrimp species are mainly exploited in this fishery, the brown and the pink shrimps (respectively, Farfantepenaeus subtilis and Farfantepenaeus brasiliensis). The F. subtilis represents more than 85% of shrimp landings. We denoted by C the total catch of this shrimp in tons for the whole French Guiana fleet. This catch C is the product of the catchability coefficient q, the fishing effort measured by the number of days at sea E and the abundance of the fish population B. Based on the Schaeffer relation $C = qEB$, the catch-per-unit-effort (CPUE) is equal to the ratio C/E. CPUE is the catch extracted from one unit of fishing effort. We use the data collected by IFREMER (French institute of research for the exploitation of the sea) on C and E between January 1989 to December 2014 to get the CPUE. The data are represented in Figure 3.

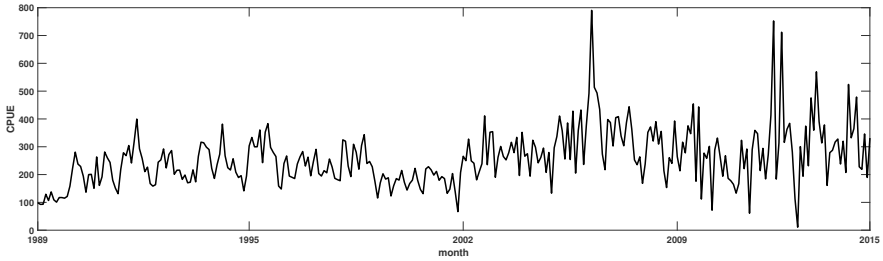

Fig. 3: Plot of monthly average of catch per unit effort between 1989 and 2014.

We will build various kind of PAR models using the BIC criterion, and also evaluate the forecasting accuracy, a standard one-step-ahead procedure, of resulting models by means of the measures Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) [12]. The first model considered in our analysis is a subset PAR without changepoints (denoted by *Model 1* in Table 3). We then estimated a PAR model with at least one changepoint: we impose an upper bound for the order of the PAR models on each regime equal to one. To avoid having too few observations in any regime we set a minimum span of $\omega = 10$ (*Model 2* in Table 3).

	Years of changepoint	$RMSE$	MAE	$MAPE$	$Fitness$
Model 1	/	108.19	85.58	29.05	0.413
Model 2	2002	89.29	67.98	22.53	0.426
Model 3	1996,2002	89.29	67.98	22.53	0.429

Table 3: Results on evaluation criteria of the forecast errors for CPUE

Using our model we found one changepoint corresponding to 2002 (Figure 4).

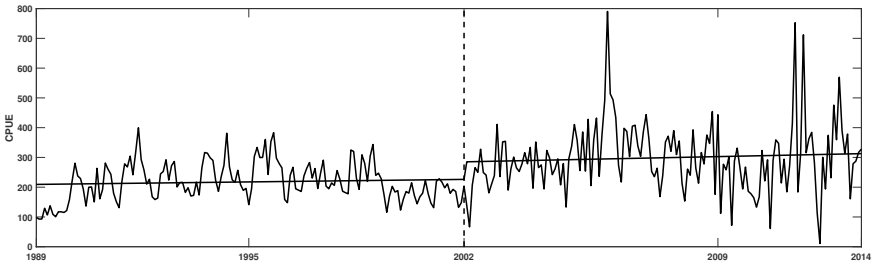

Fig. 4: Changepoint detected on years 2002 for CPUE

To ascertain the type of changes in the time series data due to changepoints, we calculate the 12 seasonal means and the 12 seasonal variation for all years up until the first changepoint. From the changepoint onwards we calculate the seasonal means or the seasonal standard deviations for each period until a detection of a new changepoint. To note that the model used by [17], designed to detect mean shifts, fails to identificate the changepoint. This could be explained by important changes in variance and not so important changes in mean (Table 4).

The changepoint corresponding to 2002 could be linked to the comments made in [15]: it reports a strong correlation between the Southern Oscillation Index (SOI) and the fish recruitment between $2002-2009$ ($R^2 = 0.81$), and a lack of correlation between $1990-2001$ ($R^2 = 0.001$). The El Niño and La Niña variables are captured through the SOI. The El Niño yields a disruption of temperature in the tropical Pacific Ocean that has important weather and climate consequences around the globe and are associated with physical and biological changes in our oceans that affect fish abundance and distribution. El Niño usually currents last for several months, resulting in the reduction of nutrients and a corresponding dissi-

Month	percentage change in mean	percentage change in variance
Jan.	16.69	107.53
Feb.	41.14	-63.80
Mar.	32.43	365.05
Apr.	69.86	1028.67
May	54.39	270.73
June	71.16	96.70
July	69.15	251.81
Aug.	16.63	78.41
Sep.	2.55	180.97
Oct.	26.87	-9.12
Nov.	22.78	140.97
Dec.	54.87	171.18

Table 4: Percentage change in mean and variance before and after 2002 of CPUE

pation of fish stocks. The La Niña is opposite for this other phase of the SOI, when sea surface temperatures in the central and eastern tropical Pacific are unusually low and when the trade winds are very intense. The sea surface temperature (SST) is a good indicator of global warming due to greenhouse gases. A change in temperature could have an impact on the movement of shrimp populations, on their rate of growth and/or mortality ([21]). The recruit abundance as well as the stock biomass and the fishing mortality were monthly performed by virtual population analysis (VPA) calculations.

If we set the minimum span $\omega = 6$ and we use our model combined with BIC we found two changepoints corresponding to 1996 and 2002 (Figure 5). The evolution of CPUE shows an increase from 1990 to 1996, followed by a decrease until 2002. The changepoint corresponding to 1996 could be linked to the evolution of biomass which has been decreasing steadily over time since 1996 followed by an improvement between 2003 and 2005, but without affecting the overall trend of decline ([14]). Fishing effort was concentrated in the shallow waters until 1995, for which the biomass was highest. Moreover, the French Guiana marine fishing area might be affected by changes in the SST since the latter significantly increased between 1970 and 2004, with an accentuation of this phenomenon since 1995 ([5]). These results, however, should be interpreted with caution as the number of observations is very important for good estimation results.

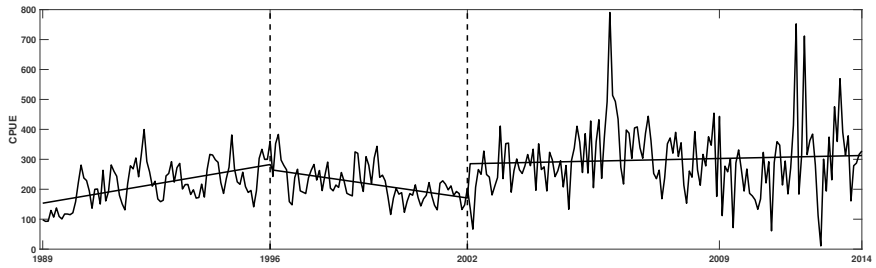

Fig. 5: Changepoint detected on years 1996 and 2002 for CPUE

5 Conclusions

The goal of our research was to develop a computational procedure for building multiregime models in time series with a periodic behaviour. Our procedure has been found effective both in simulation studies and in analysis of time series related to catch per unit effort of shrimps in French Guyana. The reasons for such changepoints are possibly due to both human activities and climatic oscillations. It is hoped that the results presented in this article will be useful in hydrology and finance, where interest lies in detecting changes in the volatility of time series due to changes in instrumentation and institutional changes, respectively.

Acknowledgments. The authors thank professor Francesco Battaglia for his valuable and constructive remarks. Part of this work has been carried out with the financial support of the French National Research Agency in the frame of the Investments for the future Programme, within the Cluster of Excellence COTE (ANR-10-LABX-45).

References

1. Aue A, Horváth L (2013) Structural breaks in time series. Journal of Time Series Analysis 34:1–16
2. Battaglia F, Protopapas MK (2012) Multi-regime models for nonlinear nonstationary time series. Computational Statistics 27:319–341
3. Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden-Day, San Francisco, CA
4. Davis RA, Lee TCM, Rodriguez-Yam GA (2006) Structural break estimation for nonstationary time series models. Journal of the American Statistical Association 473:223–239
5. Diop B, Sanz N, Duplan Y.J.J., Guene E.H.M., Blanchard F, Pereau J-C,Doyen L (2018) MEY fishery management facing climate warming. Ecological Economics:

6. Doerr B, Fischer P, Hilbert A, Witt C (2017) Detecting structural breaks in time series via genetic algorithms. Soft Computing 21(16):4707–4720
7. Franses PH, Paap R (2004) Periodic time series models. Oxford University Press
8. Goldberg DE (1989) Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley
9. Hansen BE (2001) The new econometrics of structural change: dating breaks in U.S. labor productivity. Journal of Economic Perspectives 15:117–128
10. Hipel KW, McLeod AI (1994) Time series modelling of water resources and environmental systems. Elsevier, Amsterdam
11. Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press
12. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22:679–688
13. Kreinovich V, Quintana C, Fuentes O (1993) Genetic algorithms: what fitness scaling is optimal? Cybernetics and Systems 24(1):9–26
14. Lampert, L (2011) Etude de la crise de la pêche de la crevette en Guyane, VOL.1. IFREMER:1–79
15. Lampert, L (2013) Etude de la crise de la pêche de la crevette en Guyane, VOL.2. IFREMER:1–56
16. Li S, Lund R (2012) Multiple changepoint detection via genetic algorithms. Journal of Climate 25:674–686
17. Lu Q, Lund R, Lee TCM (2010) An MDL approach to the climate segmentation problem. The Annals of Applied Statistics 4:299–319
18. Lund RB, Basawa IV (1999) Modeling and inference for periodically corelated time series. In: Gosh S (ed) Asymptotics, Nonparametrics and Time Series, Marcel Dekker, New York, Statistics : textbooks and monographs, vol 158, pp 37–62
19. Lund RB, Basawa IV (2000) Recursive prediction and likelihood evaluation for periodic ARMA models. Journal of Time Series Analysis 21:75–93
20. Lund RB, Wang XL, Lu Q, Reeves J, Gallagher C, Feng Y (2007) Changepoint detection in periodic and autocorrelated time series. Journal of Climate 20:5178–5190
21. Magraoui A, Baulier L, Blanchard F (2014) Effet du changement climatique sur le stock guyanais de crevettes pénéïdes. IFREMER:1–31
22. Moeeni H, Bonakdari H, Fatemi SE (2017) Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. Journal of Hydrology 547:348–364
23. Sanz N, Diop B, Blanchard F, Lampert L (2016) On the influence of environmental factors on harvest: the French Guiana shrimp fishery paradox. Environmental Economics and Policy Studies pp Online First Articles, DOI 10.1007/s10,018–016–0153–6
24. Tong H (1990) Non-linear time series: a dynamical system approach. Oxford University Press
25. Ursu E, Pereau JC (2015) Application of periodic autoregressive process to the modeling of the Garonne river flows. Stochastic Environmental Research and Risk Assessment 30(7):1785–1795
26. Vecchia AV (1985) Periodic autoregressive-moving average (PARMA) modeling with applications to water resources. Water Resources Bulletin 21:721–730
27. Yau CY, Tang CM, Lee TCM (2015) Estimation of multiple-regime threshold autoregressive models with structural breaks. Journal of the American Statistical Association 511:1175–1186