Indian Journal of Agricultural Sciences **83** (12): 1406–8, December 2013/Short Communication

Fitting of SETAR Three-regime nonlinear time series model to Indian lac production data through genetic algorithm

M A IQUEBAL[1], HIMADRI GHOSH[2] and PRAJNESHU[3]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 31 July 2012; Revised accepted: 15 October 2013

Key words: Forecasting, Genetic algorithm, Normalized Akaike information criterion, Nonlinear time-series model, SETAR three-regime model

In the field of agriculture, quite often the time-series data depicts cyclical fluctuations. Some examples of such behaviour are: Indian annual lac production/export data, Indian summer monsoon rainfall data and population-sizes of several fish species having prey-predator type of interactions. Prajneshu *et al.* (2002) used "Structural time-series approach" for modeling of all-India lac production data. The fitted model has shown strong periodicity of five years with estimated cycle variance, cycle frequency and cycle amplitude estimated from Cycle plus noise model, Trend plus cycle model, and Cyclic trend model. However, one limitation of this work is that the underlying model is "linear". During last three decades or so, the area of "Nonlinear time-series modelling" has been rapidly developing. To this end, we study an important subclass of such models, viz. Self-Exciting Threshold Auto Regressive (SETAR) family of models proposed by Tong (1980) and discussed in Fan and Yao (2003). A heartening aspect of this model is that it is capable of describing cyclical data having sudden rise and fall.

Nampoothiri and Balakrishna (2000) fitted the SETAR two-regime model to monthly coconut oil prices at Cochin market by following "Recursive estimation" method. However, a drawback of this work is that the estimation procedure is ad hoc in nature as is not based on any sound statistical optimization principle. Accordingly, the estimates obtained are not, in general, globally optimum. Ghosh *et al.* (2006) tried to fit the same model by "Search algorithm" (Tong 1990) to Indian lac export data exhibiting prominent cycles. However, main limitation of this algorithm is that the number of possible models to be searched is extremely large. Specifically, if the number of regime autoregressive models is M, largest order in each regime is L, number of threshold values is S, and number of values of delay parameter is T, then number of models to be computed is $\binom{S}{M-1}L^M T$, which is very large. The concept of apriori information of S threshold values is also not tenable because these are to be estimated from a continuum of threshold values in R^S. Recently, Iquebal *et al.* (2010) fitted the SETAR two-regime model to Indian lac export data through the powerful stochastic optimization technique of Genetic algorithm (GA) and demonstrated the superiority of this approach over the Search algorithm procedure for modelling as well as forecasting purposes.

The objective of this paper is to thoroughly study GA-methodology for estimation of parameters of SETAR three-regime model and apply the same to real data.

A SETAR three-regime model, written as SETAR (3; k_1, k_2, k_3) model, can be expressed as

$$X_t = \begin{cases} a_0^{(1)} + \sum_{i=1}^{k_1} a_i^{(1)} X_{t-i} + \varepsilon_i^{(1)}. & \text{if } X_{t-d} \le r_1 \\ a_0^{(2)} + \sum_{i=1}^{k_2} a_i^{(2)} X_{t-i} + \varepsilon_i^{(2)}. & \text{if } r_1 < X_{t-d} \le r_2 \\ a_0^{(3)} + \sum_{i=1}^{k_3} a_i^{(3)} X_{t-i} + \varepsilon_i^{(3)}. & \text{if } X_{t-d} > r_2 \end{cases} \quad (1)$$

where k_1, k_2, k_3 are orders of three AR models; $\{a_i^{(1)}\}$, $\{a_i^{(2)}\}$, $\{a_i^{(3)}\}$, are autoregressive coefficients; $\varepsilon_t^{(1)}$, $\varepsilon_t^{(2)}$, $\varepsilon_t^{(3)}$ are the white noise terms; d is the delay parameter (where the controlled threshold occurs) and r, j = 1, 2 represent threshold values.

It may be noted that, for fixed threshold values and threshold variable, eq. (1) is linear and so the parameters of the SETAR model can be estimated by using Conditional least squares method. Antonio *et al.* (2009) have recently released their tsDyn package, Ver. 0.7-1 in R, which is a very

[1] Scientist, Centre for Agricultural Bioinformatics, [2] Senior Scientist and [3]Principal Scientist and Head, Division of Statistical Genetics

versatile package and can be employed for fitting several types of nonlinear time-series models including the SETAR model. However, one limitation of the estimation procedure used is that global optima cannot be ensured. To this end, the powerful stochastic search optimization procedure of Genetic algorithm (GA), motivated by the principles of Genetics and natural selection, may be used. GA combines Charles Darwin's principle of "natural selection" and "survival of the fittest" with computer-constructed evolution mechanism to select better species from the original population. The information is further exchanged among them, which produces more number of meaningful schemata (specific sequence of '0', '1' and '*' indicating region in search space). This leads to superior offspring in the sense of improved average fitness values in the successive generations of GA. Besides, in order to avoid missing some good species and becoming a local optimization, several mutations must be processed. This problem may be overcome by using GA in which robust search algorithms that require minimal problem information are artificially constructed. The working principle of GA is very different from that of the classical optimization techniques, in which more than one solution is allowed to direct the search space in each iteration. The real-coded GA is used in present study. A heartening feature of the latter is that, unlike other competing optimization methods, it is capable of exploring the complete search space (Deb 2002).

The GA procedure is based on encoding the parameter values into an appropriate range with finite-length digital strings (i.e. chromosomes, usually binary strings of length α). A widely used formula for decoding is

$$c + L + \{A/(2^B-1)\}\times(U-L) \qquad (2)$$

where c is the encoded value of a chromosome, U and L are upper and lower bounds of the parameter to be estimated, and A and B are respectively the decoded value and number of digits of the chromosome.

The three operators, viz. selection, crossover, and mutation make GA an important tool for optimization. When a string (parameter solution) is created by GA, it is evaluated in terms of its fitness, which is the Normalized Akaike information criterion (NAIC) of SETAR (m; k_1, k_2,…, k_m) defined as

$$\mathrm{NAIC} = \left\{\sum_{j=1}^{3} n_j \log(S_j / n_j) + 2\sum_{j=1}^{3}(p_j+1)\right\} \Big/ (\text{Effective sample size}) \qquad (3)$$

where n_j is the number of observations that belong to the regime j and S_j is the Residual sum of squares for the fitted SETAR j^{th} regime model. The details of GA methodology are available in Iquebal *et al*. (2010).

The data on all-India annual lac production for the period 1930-31 to 2002-03, obtained from the annual reports of Shellac Export Promotion Council, Kolkata, is considered for fitting the SETAR three-regime model. The lac production for 1930-31 and 2002-03 are 13.26 and 17.50 respectively. Real-coded GA with SBX (crossover operator) where (η_c = 2) is applied for estimation of parameters. The detail of SBX operator is given in Deb and Agrawal (1995). Computer programs for fitting SETAR three-regime model are developed in C. However, to save space, only codes for objective function are appended as Annexure-1. The GA parameters, viz. population size, crossover probability, and mutation probability for minimization of NAIC are respectively 100, 0.9, 0.01 with number of generations as 100. The proposed algorithm enables us to select SETAR (3; 1, 1, 1) model, i.e. d = 1, k_1 = 1, k_2 = 1 and k_3 = 1. The optimal threshold values come out as 25 and 40 million tonnes respectively. The best fitted three-regime SETAR model on the basis of minimum NAIC value, viz. 6.23, is

$$X_t = \begin{cases} 12.00 + 0.51\, X_{t-1}, & \text{if } X_{t-1} \le 25 \\ 29.94 + 0.39 X_{t-1}, & \text{if } 25 < X_{t-1} \le 40 \\ 25.01 + 0.60 X_{t-1}, & \text{if } X_{t-1} > 40 \end{cases} \qquad (4)$$

with $\mathrm{Var}(\varepsilon_t^{(1)}) = 2.15$, $\mathrm{Var}(\varepsilon_t^{(2)}) = 4.31$, and $\mathrm{Var}(\varepsilon_t^{(3)}) = 8.61$. The standard errors of parameter estimates ($a_0^{(1)}$, $a_1^{(1)}$, $a_0^{(2)}$, $a_1^{(2)}$, $a_0^{(3)}$, $a_1^{(3)}$,) are respectively computed as (1.24, 0.03, 2.41, 0.11, 3.17, 0.08). To get a visual idea, the fitted SETAR (3; 1, 1, 1) model along with data points is exhibited in Fig 1.

A mechanistic interpretation of fitted SETAR model is as follows. The above fitted model given by eq. (4) can be written as

$$X_t - X_{t-1} = \begin{cases} 12.00 + 0.49\, X_{t-1}, & \text{if } X_{t-1} \le 25 \\ 29.94 + 0.61 X_{t-1}, & \text{if } 25 < X_{t-1} \le 40 \\ 25.01 + 0.40 X_{t-1}, & \text{if } X_{t-1} > 40 \end{cases} \qquad (5)$$

In the lower regime, i.e. $X_{t-1} \le 25$, $X_t - X_{t-1}$ tends to be positive but small, implying slow increase in lac production. In the middle regime, i.e. $25 < X_{t-1} \le 40$, $X_t - X_{t-1}$ tends to be positive but large, implying comparatively faster increase in lac production. However, in the higher regime, i.e. $X_{t-1} >$

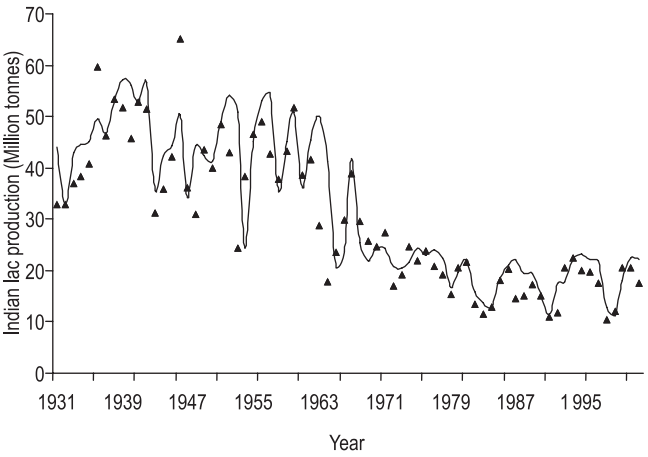

Fig 1 Fitted SETAR (3; 1, 1, 1) model along with observed data of Indian lac production

Table 1 Forecasting for hold-out and out-of-sample data of all-India lac production data (metric tonnes) by SETAR three-regime model

Year	Actual	Forecast
2003-04	20.05	20.77
2004-05	21.30	22.05
2005-06	18.00	22.67
2006-07	23.23	21.02
2007-08	20.64	23.64
2008-09	17.18	22.34
2009-10	16.50	20.60
2010-11	09.04	10.26
2011-12	17.90	16.53
2012-13		22.50
2013-14		23.27
2014-15		23.66
2015-16		23.85
2020-21		24.04

40, $X_t - X_{t-1}$ tends to be negative, implying decrease in the lac production. This type of behavior leads to periodicity, which is in agreement with observed lac production data.

Forecasting for hold-out data

The forecast value for Indian lac production is presented in Table 1 for the year 2003-04 to 2011-12 using the fitted model. The forecast performance for hold-out data on the basis of Root mean square error and Mean absolute error for the fitted SETAR model are computed as 2.58 and 3.05 respectively. The forecast value obtained for the year 2012-13 to 2015-16 and 2020-21 is also shown in Table 1.

SUMMARY

In this paper, utility of Genetic algorithm for fitting of SETAR three-regime model is highlighted. The proposed procedure is successfully applied for modelling and forecasting of Indian lac production data. It is hoped that, applied statisticians would also start employing Genetic algorithm for fitting other nonlinear time-series models.

REFERENCES

Antonio F D N, Aznarte J L and Stigler M. 2009. tsDyn: Time Series Analysis Based On Dynamical Systems Theory Package.

Deb K and Agrawal R B. 1995. Simulated binary crossover for continuous search space. *Complex Systems* **9**: 115–48.

Deb K. 2002. *Multi-Objective Optimization using Evolutionary Algorithms.* John Wiley, Singapore.

Fan J and Yao Q. 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Ghosh H, Sunilkumar G and Prajneshu. 2006. Self exciting threshold autoregressive models for describing cyclical data. *Calcutta Statistical Association Bulletin* **58**: 115–32.

Iquebal M A, Ghosh H and Prajneshu. 2010. Application of genetic algorithm for fitting self-exciting threshold autoregressive nonlinear time-series model. *Journal of the Indian Society of Agricultural Statistics* **64**: 391–8.

Nampoothiri C K and Balakrishna N. 2000. Threshold autoregressive model for a time series data. *Journal of the Indian Society of Agricultural Statistics* **53**: 151–60.

Prajneshu Ravichandran S and Wadhwa S. 2002. Structural time-series models for describing cyclical fluctuations. *Journal of Indian Society of Agricultural Statistics* **55**: 70–8.

Tong H. 1980. *Threshold Models in Non-linear Time Series Analysis*. Springer, New York.

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