Int. J. Agricult. Stat. Sci. Vol. 13, No. 2, pp. 615-621, 2017 *ISSN : 0973-1903*

ORIGINAL ARTICLE

GENETIC ALGORITHM APPROACH FOR ESTIMATION OF PARAMETERS OF VECTOR AUTOREGRESSIVE MODELS UNDER HETEROSCEDASTICITY

B. S. Yashavanth*, K. N. Singh[1] and A. K. Paul[1]

ICAR-National Academy of Agricultural Research Management, Hyderabad - 500 030, India.
[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110 012, India.
E-mail : yashavanthbs@gmail.com

Abstract : Forecasting is one of the core focuses of statisticians working in agricultural research. Obtaining timely as well as accurate forecasts under all possible circumstances is the need of the hour. Most of the forecasting techniques make one or the other assumptions limiting their applications. Vector Autoregression is one such widely used multivariate forecasting technique where homoscedasticity of errors is assumed for estimation of parameters by ordinary least square (OLS) method. This study proposes genetic algorithm (GA), a heuristic search algorithm, which does not make any such assumptions for estimating the parameters under such situation. The developed methodology is empirically validated using simulated bivariate vector autoregressive model of order 1 under heteroscedasticity. The relative error of parameter estimates and Mean Absolute Percentage Error have shown that GA performs better than OLS estimation under heteroscedasticity. The proposed methodology is also tested under homoscedasticity using bivariate data of fish landings. The results indicated that both GA and OLS are equally efficient in estimating the parameters.

Key words : Vector autoregression, Heteroscedasticity, Least squares, Genetic algorithm.

1. Introduction

Forecasting techniques in agriculture include, inter alia, forecasting of production, yield, area, prices of crops and forewarning of incidence of crop pests and diseases. Agricultural production and price are highly varying as they are largely influenced by several eventualities. Natural calamities like droughts, floods and attacks by pests and diseases make these unpredictable leading to a considerable risk and uncertainty in the process of modeling and forecasting. Forecasts of agricultural production and prices are intended to be useful to the farmers, governments and agribusiness industries. Policy makers need internal forecasts to execute policies that provide technical and market support for the agricultural sector.

In time series forecasting, the past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. During the past few decades, a lot of effort has been directed towards developing and improving time series forecasting models. Literature has been flooded with the application of univariate time series models, mostly Autoregressive Integrated Moving Average (ARIMA) models, for this purpose. But lately, multivariate time series techniques like Vector autoregressive (VAR) models are used extensively. In VAR models, all the series are modeled at a go capturing the relations between different series, which helps in arriving at better forecasts than those given by univariate time series models. Sathianandan (2007) used VAR type of models to model and discover the relationships between landings of eight commercially important marine fish species/groups using quarterwise landings in Kerala during 1960-2005. Kilian (2011) forecasted the price of oil using Vector Autoregression. Trujiello-Barrera *et al.* (2013) forecasted hog prices in United States using VAR models. Gutierrez (2014) employed VAR methodology to analyze the world wheat market.

The VAR modeling method is simple since all variables considered are endogenous and the Ordinary

Author for correspondence** ***Received *April 12, 2017* ***Revised*** *July 29, 2017* ***Accepted*** *August 30, 2017*

Least Square technique (OLS) can be applied for estimation of parameters making it advantageous over other multivariate modeling techniques like simultaneous equation models. Despite these advantages, VAR models require certain assumptions to be satisfied for their successful application. One such assumption is the homoscedasticity of error terms. If one ignores heteroscedasticity and uses OLS technique for parameter estimation, the properties of unbiasedness and consistency of parameter estimates are not violated, but they are no more efficient. Also, the estimates of the variances of the parameters are no longer unbiased. There are possibilities to find an alternative unbiased linear estimate that has a lower variance than OLS estimate. Moreover, if we persist in using the usual testing procedures despite heteroscedasticity, the conclusions drawn or inferences that are made may be very misleading [Gujarati *et al.* (2009)].

If the assumption of homoscedasticity is not met, one can go for Generalized Least Squares (GLS) technique for estimation of the parameters instead of OLS technique. But, the GLS estimator is usually not available and whenever available, it yields local optimum values. So there is a need to study methods, which give global optimum values. This can be achieved by using heuristic search algorithms which do not make any assumptions. One such algorithm is the genetic algorithm (GA), developed by Holland (1975), which finds application in many situations. Use of GAs for optimization problems is conspicuous in literature. GAs have been successfully used for estimating parameters in regression under heteroscedasticity [Iquebal *et al.* (2008)], ARMA [Hung (2008)] and their better performance over the GLS method has been established [Parviz *et al.* (2010), Abo-Hammour *et al.* (2012)]. Based on these available results, this study extends the application of GAs for estimation of parameters in VAR models under heteroscedasticity. The performance of the proposed technique is evaluated by comparing with the existing OLS technique.

2. Materials and Methods

Vector Autoregression

A univariate autoregression of order *p* involves one variable where it is regressed on *p* lags of itself. In contrast, a vector autoregression of order *p* involves N different equations, one each for N variables. In each equation, we regress a variable on *p* lags of it and *p* lags of every other variable. Thus, the right hand side variables are the same in every equation – *p* lags of every variable. The key point is that, in contrast to the univariate case, vector autoregressions allow for cross-variable dynamics. Each variable is related not only to its own past, but also to the past of all the other variables in the system.

Suppose there are *k* time series components $\{Y_{1t}\},\{Y_{2t}\},\ldots,\{Y_{kt}\}$ for $t = 0,1,2,3,\ldots,n$ at equally spaced time intervals. We can represent these components by a vector $Y_t = (Y_{1t}, Y_{2t}, \ldots, Y_{kt})^T$, which is called a vector of time series. A vector time series with *k* components can be modeled by a vector autoregressive model of order *p* denoted by VAR (*p*) and its expression is

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t \qquad (1)$$

where, μ is the mean vector of the series, β_i ($i = 1, 2, \ldots, p$) are $k \times k$ parameter matrices, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{kt})^T$ are independently and identically distributed random innovation vectors having zero mean. For example, a bivariate VAR (1) is written as below in matrix notations:

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \qquad (2)$$

Genetic Algorithm

GAs are stochastic search algorithms based on the Darwinian concepts of natural selection and evolution to solve optimization problems. The procedure of GAs starts by randomly choosing an assortment of chromosomes (a numerical value or values that represent a candidate solution to the problem), which serves as the first generation (initial population). Then each chromosome in the population is evaluated by the fitness function to test how well it solves the problem of interest. Once all the individuals in the population have been evaluated, their fitnesses are used as the basis for selection. Selection is implemented by eliminating low-fitness individuals from the population and inheritance is implemented by making multiple copies of high-fitness individuals. These inherited chromosomes are treated later by further genetic operations, such as mutation (flipping individual bits) and crossover (exchanging substrings of two individuals to obtain two offspring), which are applied probabilistically to the selected individuals to produce a new population of individuals. The evolution process is terminated based on some convergence criteria like the maximum number of generations is defined or when a sufficiently large number of generations have passed without any improvement in the best fitness value.

For the problem under study, the Genetic Algorithm approach to estimate the parameters of a VAR model goes through following stages.

(i) Initiation : On the basis of number of unknown parameters that are to be estimated *i.e.*, k^2p+k parameters for a k variable VAR of order p, a population of individuals is created. Here, each parameter is an individual in the population. The population contains different sets of solutions called chromosomes.

(ii) Evaluation : The evaluation of chromosomes (solutions) is performed based on the fitness function. To evaluate the degree of goodness of a chromosome, the following fitness function, F, is used.

$$F = \frac{1}{1 + \sum_{l=1}^{k} RSS_l} \tag{3}$$

where, $RSS = \sum_{i=1}^{t} \left(Y_i - \hat{Y}_i\right)$ and k is the number of variables in the VAR model. The above fitness function is maximized since the parameters of the VAR model are estimated by minimizing the total of Residual Sum of Squares for each variable.

(iii) Selection : Chromosomes are chosen from the current population and entered into the mating pool to create new children. These new children constitute the next generation. The chromosomes are selected based on the fitness value, larger the fitness, higher the probability that the chromosome will contribute one or more children for the next generation. Only the best chromosomes are selected to continue. Selecting the fraction of the population, referred to as elitism, that survives for the next step of mating is usually kept 5% of the population. In this study, fitness proportionate selection (roulette wheel selection) is used for selecting the candidate solutions.

(iv) Crossover : Each pair of chromosomes is crossed over at some randomly chosen point to produce two new segments. Usually, children inherit some genes from each parent; however, they have their own structures compared with their parents. The crossover operation is not usually applied to all the chromosomes that are selected for mating. However, the choice is made randomly with a probability of crossover being between 0.5 and 1. The 'local arithmetic crossover' where in some arithmetic operation such as addition or multiplication is performed to make a new offspring is used in this study.

(v) Mutation : This operation is used to provide a small amount of random search to guard against any premature convergence. The mutation operation is applied to each child individually once the crossover operation is performed. In nature, the probability of mutation is very low and hence, it is kept below 0.2. In this study, the 'uniform random mutation' where the value of the chosen gene is replaced with a uniform random value is made use of.

(vi) Termination : The GA is terminated when some convergence criterion is met, such as the maximum number of generations is reached, a desired fitness value is reached or when a sufficiently large number of generations have passed without any improvement in the best fitness value. Once the algorithm is terminated, the Information Criteria (Akaike Information Criteria, Bayesian Information Criteria, etc.) can be used to choose the order of the VAR model.

There is no definite set rule to select the parameters of the genetic algorithm *viz.*, population size, elitism crossover and mutation probabilities. Different combinations of these parameters are to be tried and the combination, which gives the best fitness in least number of generations is chosen. For illustrations in this study, the following set of combination of parameters of the GA is chosen for the estimation: Population size = {100, 200, 300}

Crossover probability = {0.7, 0.8, 0.9}

Mutation probability = {0.1, 0.2, 0.3}

Illustration

The proposed GA technique is tested both under homoscedastic as well as heteroscedastic conditions. To test the algorithm under heteroscedastic condition, a two variable VAR of order 1 is generated, which is given as

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} + \begin{bmatrix} 1.2 & -0.5 \\ 0.6 & 0.3 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{4}$$

where, r_i (i = 1, 2) are the residual series. Also, to test the algorithm under homoscedastic conditions, the data on estimated annual landings of oil sardines and mackerel fish species in India is considered. The data consist of 64 observations for each variable (1950-2013). The first 58 observations are used for model fitting and last 6 observations are used for model evaluation. The data is available at www.cmfri.org.in. Mean Absolute Percentage Error (MAPE) is used to evaluate the performance of the models. SAS 9.3 is used for simulating the data and estimation of parameters using OLS whereas R 3.2.1 is used for GA technique of parameter estimation.

3. Results and Discussion

Simulated Data : The first step in time series analysis is to plot the data to get an idea about the behavior of the series. Fig. 1 shows the time series plot of the simulated variables given in Equation (4).

A basic assumption in time series modeling is that the series under study is stationary in nature. A perusal of Fig. 1 reveals that there is no trend over time hinting at the stationary nature of both the series. To confirm this, a non-parametric stationarity test, Phillips-Perron unit root test is performed. The results are given in the Table 1, which confirm the stationarity of the series at 5% level of significance.

Subsequently, a VAR (1) was fitted for the simulated data using OLS technique and the residuals are tested for presence of autocorrelation and heteroscedasticity using Durbin-Watson (DW) test and Lagrange's Multiplier (LM) test, respectively. The DW statistic ranges between 0 and 4, close to 2 indicating absence of autocorrelation. The result of the DW test indicates that the residual series are not autocorrelated. But the Lagrange's Multiplier (LM) test for heteroscedastic residuals rejects the null hypothesis of homoscedasticity. The results are given in Table 2.

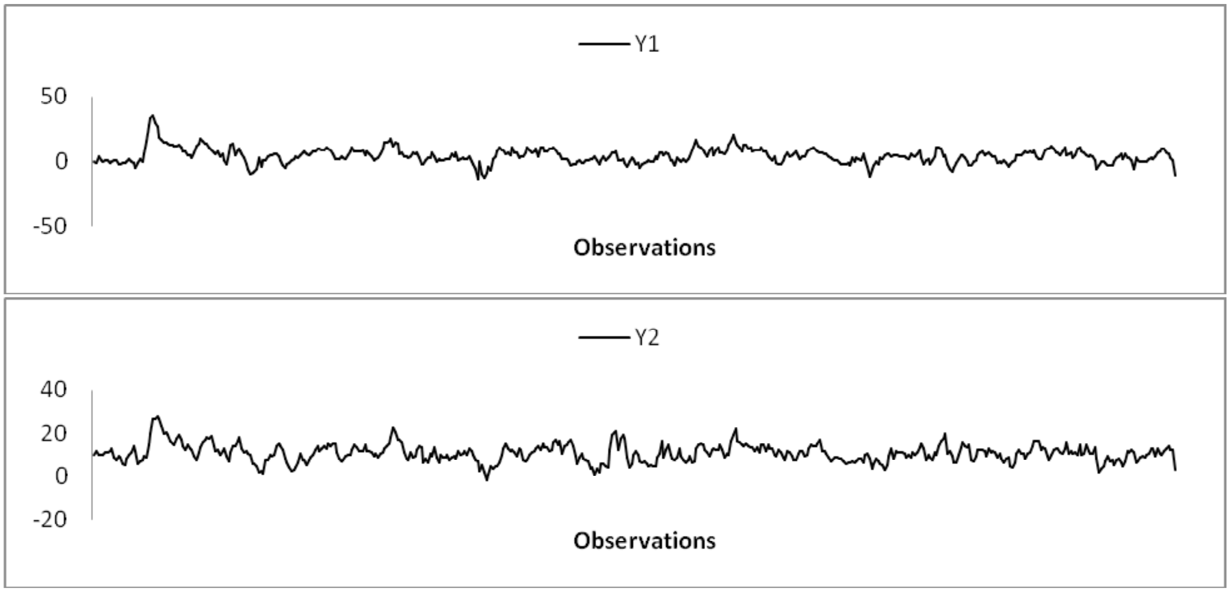

Fig. 1 : Time series plot of the simulated variables.

Table 1 : Results of PP test for stationarity of simulated variables.

Variable	Phillips-Perron Unit Root Tests				
	Type	Rho	Pr < Rho	Tau	Pr < Tau
Y1	Zero Mean	-48.1897	<.0001	-5.0028	<.0001
	Single Mean	-79.6831	0.0017	-6.4797	<.0001
	Trend	-82.1127	0.0007	-6.5864	<.0001
Y2	Zero Mean	-9.7983	0.0295	-2.2152	0.0259
	Single Mean	-99.7582	0.0017	-7.3792	<.0001
	Trend	-101.688	0.0007	-7.4456	<.0001

Table 2 : Results of the residual analysis.

Variable	Autocorrelation (Durbin Watson statistic)	LM test for Heteroscedasticity	
		F Value	Pr > F
Y1	2.04	181.03	<.0001
Y2	1.86	100.82	<.0001

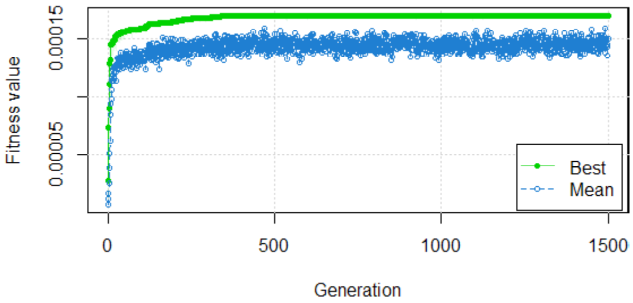

Fig. 2 : Plot showing fitness values obtained for each generation.

Table 3 : Parameters of the genetic algorithm for the simulated data.

Parameter	Value
Population size	100
Number of generations	1500
Elitism	5
Crossover probability	0.7
Mutation probability	0.2
Fitness function value	0.000169

Table 4 : Comparison of parameters estimated from GA and OLS techniques.

Parameter	True value	Value from		Absolute Difference		Relative Error (%)	
		GA	OLS	GA	OLS	GA	OLS
a_{11}	2	2.124	2.211	0.124	0.211	6.200	10.550
a_{21}	4	3.650	3.422	0.350	0.578	8.750	14.450
b_{11}	0.9	0.929	0.946	0.029	0.046	3.222	5.111
b_{12}	-0.15	-0.173	-0.183	0.023	0.033	15.333	22.000
b_{21}	0.3	0.296	0.292	0.004	0.008	1.333	2.667
b_{22}	0.5	0.542	0.565	0.042	0.065	8.400	13.000

Table 5 : Comparison of forecast performance.

Variable	MAPE (%)	
	GA	OLS
Y1	2.586	3.446
Y2	0.198	2.538

Table 6 : Results of PP test for stationarity of fish data.

Variable	Phillips-Perron Unit Root Tests				
	Type	Rho	Pr < Rho	Tau	Pr < Tau
Oil sardines	Zero Mean	0.185	0.722	0.224	0.748
	Single Mean	-14.68	0.032	-2.938	0.047
	Trend	-22.810	0.022	-3.746	0.026
Mackerel	Zero Mean	-0.224	0.628	-0.247	0.592
	Single Mean	-18.754	0.009	-3.241	0.022
	Trend	-28.051	0.005	-4.292	0.006

In such a situation, though the parameter estimates are unbiased, their standard errors are biased and inefficient, making them non-reliable. For the same data, a VAR (1) was fitted by adopting the proposed Genetic Algorithm approach of parameter estimation by maximizing the fitness function,

$$F = \frac{1}{1+\sum_{\ell=1}^{2} RSS_{\ell}} \qquad (5)$$

where, $RSS = \sum_{i=1}^{t}\left(Y_i - \hat{Y}_i\right)$. Fitness values obtained for each generation are plotted in Fig. 2. The fitness value kept getting better and better for each generation till 1300th generation after which, there is a microscopic improvement. Hence, the iteration was stopped after 1500th generation.

The combination of parameters that gave best fitness value is given in Table 3. The best fitness value

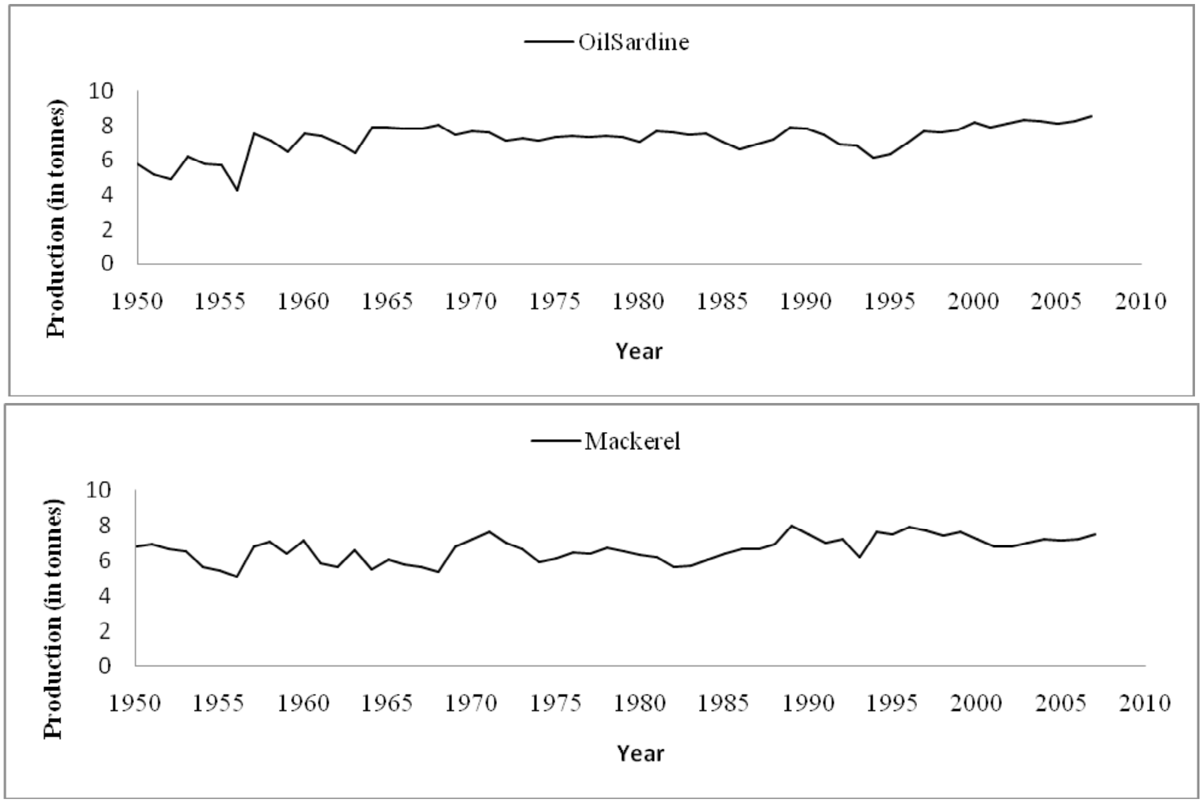

Fig. 3 : Time series plot of the annual fish landings.

obtained is 0.000169, which did not improve even after next 100 generations.

A comparison of the model parameters obtained by adopting OLS and Genetic Algorithm approach are given in Table 4 along with the true values of the simulated VAR (1) model. The absolute difference between true and estimated parameters is found out and is used to find the percentage of relative error.

By looking at relative error in the Table 4, it is evident that the estimates obtained by GA are closer to the true parameters and have consistently performed better than OLS estimates. The forecasting performance is also better for parameters estimated by GA which is apparent from the Mean Absolute Percentage Error (MAPE) values given in Table 5.

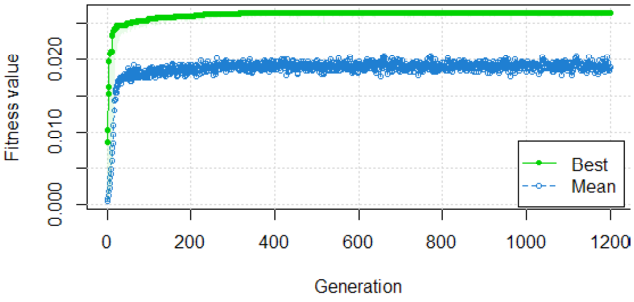

Fig. 4 : Plot showing fitness values obtained for each generation.

Table 7 : Results of the residual analysis.

Variable	Autocorrelation (Durbin Watson statistic)	LM test for Heteroscedasticity	
		F Value	Pr > F
Oil Sardines	2.03	117.99	<.0001
Mackerel	1.70	580.90	<.0001

Real data : The real data of annual production (in tonnes) of oil sardines and mackerel fish species is used to compare the performance of the proposed method with OLS method under homoscedasticity. Fig. 3 shows the time series plot of the data corresponding to annual production of oil sardines and mackerel fish species.

An inspection of Fig. 1 reveals that there is no trend over time indicating the stationarity of both series which is also confirmed by the results of PP unit root test, at 5% significance level, given in Table 6.

VAR (1) is chosen as the best model by using the information criteria and the residuals are tested for presence of autocorrelation and heteroscedasticity. The results of DW test and LM test (Table 7) indicate the absence of autocorrelation and heteroscedasticity, respectively.

The combination of parameters that gave best fitness value is given in Table 8 and the fitness values for different generations are also plotted (Fig. 4).

The parameters obtained by OLS and GA are given in Table 9, which are identical indicating that under absence of heteroscedasticity, both estimation techniques are equally efficient.

The measure of forecast accuracy MAPE is given in Table 10. It is evident from the table that the GA approach has performed on par with the OLS technique when homoscedasticity assumption is met by the data.

Table 8 : Parameters of the genetic algorithm for the fish data.

Parameter	Value
Population size	300
Number of generations	1200
Elitism	15
Crossover probability	0.9
Mutation probability	0.2
Fitness function value	0.0264

Table 9 : Comparison of parameters estimated from GA and OLS techniques.

Parameter	OLS	GA
a_{11}	1.323	1.326
a_{21}	1.519	1.519
b_{11}	0.694	0.693
b_{12}	0.052	0.051
b_{21}	-0.0059	-0.0057
b_{22}	0.659	0.659

Table 10 : Comparison of forecast performance.

Series	Training		Testing	
	OLS	GA	OLS	GA
Oil sardines	0.100	0.100	0.126	0.124
Mackerel	0.104	0.104	0.133	0.132

4. Conclusion

Time series techniques are being used since decades for efficiently modeling and forecasting agricultural data series. VAR models have found vast applications due to their better performance over univariate time series models. But their application is limited to datasets satisfying the assumptions made by VAR. One such assumption is the homoscedasticity of the error series under which usual OLS estimation of parameter becomes unreliable. This study makes an attempt to address this problem by using genetic algorithm, which is a heuristic search algorithm, as an alternative to OLS estimation. In GA approach, a fitness function is decided and its best value is obtained by using the selection, crossover and mutation operations. The performance of the proposed GA estimation technique is tested under both homoscedastic as well as heteroscedastic error series using statistical measures of performance for simulated and real data. The GA approach is found to be an efficient alternative for estimating parameters under heteroscedasticity.

References

Abo-Hammour, Z. S., O. M. K. Alsmadi, A. M. Al-Smadi, M. I Zaqout and M. S. Saraireh (2012). ARMA model order and parameter estimation using genetic algorithms. *Mathematical and Computer Modelling of Dynamical Systems: Methods, Tools and Applications in Engineering and Related Sciences,* **18(2)**, 201-221.

Gujarati, D. N., D. C. Porter and S. Gunasekar (2009). *Basic Econometrics*. Tata McGraw-Hill, New Delhi.

Gutierrez, L., F. Piras and P. P. Roggero (2014). A global vector autoregression model for the analysis of wheat export prices. *American Journal of Agricultural Economics Advance Access*, 1-18.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press. (Second edition: MIT Press, 1992.)

Hung, J. C. (2008). A genetic algorithm approach to the spectral estimation of time series with noise and missed observations. *Information Sciences*, **178**, 632-4643.

Iquebal, M. A., Prajneshu and H. Ghosh (2012). Genetic algorithm optimization technique for linear regression models with heteroscedastic errors. *Ind. J. Agricult. Sci.*, **82(5)**, 422-425.

Kilian, L. (2011). Real-time forecasts of the real price of oil. *J. Business and Econ. Stat.*, **30(2)**, 326-336.

Parviz, L., M. Kholgi and A. Hoorfar (2010). A comparison of the efficiency of parameter estimation methods in the context of streamflow forecasting. *Journal of Agricultural Science and Technology*, **12**, 47-60.

Sathianandan, T. V. (2007). Vector time series modeling of marine fish landings in Kerala. *Journal of the Marine Biological Association of India*, **49(2)**, 197-205.

Trujillo-Barrera, A., P. Garcia and M. Mallory (2013). Price density forecasts in the US hog market : Composite Procedures. *Proceedings of the NCCC-134 conference on Applied Commodity Price Analysis, Forecasting and Market Risk Management*. St. Louis, MO.