International Journal of Pure and Applied Science, Volume 1, Number 2, 2008

TIME SERIES PREDICTION BASED ON GENETIC ALGORITHM WITH APPLICATION IN FINANCE

R.A. Adeyemi[1] and G. M. Oyeyemi[2]
[1]Federal University of Technology, Minna , Niger State, Nigeria
[2]Department of Statistics, University of Ilorin, Kwara State, Nigeria

ABSTRACT

Real world problems are described by non-linear and chaotic processes, which makes them hard to model and predict. The aim of this paper is to determine the structure and weights of a time series model using genetic algorithm (GA). The paper first describes the traditional procedure of estimating time series models, which are commonly used in financial forecasting. These traditional estimation methods may not be adequate enough to capture stochastic nature of the financial due to its complexity. This article gives a brief background of Genetic algorithm method and its estimation procedure. This approach was later applied to model the Naira exchange rates against other currencies and it yielded a mean square error of 0.0058, 0.00799, 0.03744, 1.212 and 0.1108 for U.S dollars, British Pound, Japanese Yen, CFA franc and Swiss franc respectively.

Keywords: Genetic algorithm, Mean square error, Variation criterion, Exchange rate

INTRODUCTION

The increasing awareness in the financial industries (both private and regulators) of the consequences of extreme risks (the possibility of losing large amount of money) in tradable portfolios has called for effective risk management systems to be put in place for financial institutions, such as banks and investment firms. A predictive model can be used to quantify such risk ahead of time and estimate its market worth as essential management alternatives used for internal or external requirements parallel with other models. Quite natural, trend analysis may be defined as the analysis of changes in a given item/variable or sets of items/variables of information regarding a system over a period of time. Such items could be environmental, sociological and economical data, e.g. exchange rates, investment, hydrological data which deals with time series data. These typically consist of time resolved observations reflecting a system's response to some (usually unknown or not quantified) external driving force(s). A time series can be defined as any kind of (timely) ordered data series:

$X = (x_1, x_2, x_3, \ldots, x_t)$ with $t>0$ (1)

An analysis of such time series is commonly carried out with the purpose of i) its description, ii) its modelling and iii) a prognosis of the future evolution of the time series [Schlittgen & Streitberg 1984]. A central idea of i) is a regression (linear or non-linear) analysis in a least squares sense. A typical application involves the time series decomposition into a trend (long-term development of series), a cyclical (referring to regular, for instance, seasonal fluctuations of known period) and an irregular residual component. These forms of regression models, however, require a number of prerequisites (e.g. independency of the individual components) which may be called unrealistic when dealing with natural time series [e.g. Schlittgen & Streitberg 1984]. Therefore, time series modelling (ii) usually comprises the formulation and fitting of stochastic models. The prognosis (iii), i.e. the forecasting of future values of the time series from its historic data record, then essentially requires the validity of the fitted (stochastic) model. As time series are observed for a variety of different variables many different scientific disciplines have been concerned with their study, e.g. economics, meteorology, sociology. The methods applied to time series are therefore manifold and this paper can only cover a selection of them comprising least-squares regression and Fourier analysis as standard methods, as well as autoregressive integrated moving average models (ARIMA), and genetic algorithm on the more advanced side. Furthermore, a main focus of this paper (as mentioned in the title) are genetic algorithms approach to time series model and it describes the working principle and implement the method on naira exchange rate to other currencies.

METHODS

Time series analysis methods can either be applied to the time domain (e.g. regression, ARIMA models) or to the frequency domain (e.g. Fourier analysis and extensions). Most of the methods are introduced by assuming a single, i.e. univariate time series. However, the generalization of the methods to several, i.e. multivariate time series usually holds. In the following we begin with two standard analysis methods, which offer important concepts. However, the presentation of these methods is intentionally brief, as our main focus lies on the more advanced methods of the later sections.

LEAST SQUARES REGRESSION

Least squares regression requires the fitting of a model curve to the time series data. In this deterministic approach, one has to decide beforehand, which function class (e.g. linear, polynomial, logarithmic) fα(t) to apply (where α Λ and Λ represents a set of possible values for the function class' free parameters). In the univariate or one-dimensional case (simple curve-fitting), a decision for the function class can be made by visual inspection of a time-value plot. The function parameters α are found by minimizing some form of squared residuals, the typical difference or error measure being the mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x(t_i) - f_\alpha(t_i))^2 \qquad (2)$$

where $x(t_i)$ is the observed time series value at time t_i and $f_\alpha(t_i)$ is the computed function value for time t_i for a fixed α. According to e.g. Gottman [1981], Streitberg & Schlittgen [1984] a least square fit is also a canonical method to de-trend time series for further analysis such as ARMA models,. Future predictions could also be made using the formula of the fitted curve.

FOURIER ANALYSIS

A standard method for analyzing time series with respect to periodicity and seasonality is a Fourier analysis. Here, the (discretely sampled) time series is transformed from the time domain into the frequency or spectral domain, where perfectly sinusoidal oscillations correspond to a single frequency peak. Using the Euler representation of complex numbers $e^{i\varphi} = \cos(\varphi) - i\sin(\varphi)$ a convenient form of a Fourier transform of frequency n is (after Gottman [1981]):

$$f_n = \sum_{i=1}^{N-1} e^{-2\pi i n \frac{i}{N}} x_i \qquad (3)$$

(3)

From this Fourier-transform the periodogram can be computed by

$$I(f) = \frac{1}{2\pi T}\left|\sum_{i=1}^{N-1} e^{-2\pi i n \frac{i}{N}} x_i\right|^2$$

From the periodogram one can easily detect the key frequencies of seasonalities and remove them from the time series by building the inverse Fourier transform without the corresponding frequency peaks. However, there are clear limits of this approach when applied to natural time series, e.g. Jenkin & Watts (1968). Hsieh [2004]) suggested that the trend should be removed prior to the analysis and Fourier analysis scatters the energy of non-sinusoidal signals into its higher harmonics, such that a signal reconstruction becomes complicated, i.e. the number of frequency peaks needed to capture the characteristic features of the time series is not significantly lower than the number of data describing it in the time domain. Nevertheless, a Fourier analysis may still serve as tool pointing the direction of further analyses. According to some authors (Gottman., 1981; Schlittgen & Streitberg., 1984; and Kantz, & Schreiber., 1997), some characteristic features of the periodogram may be indicative of the underlying system dynamics (for detail see M. Strickert (2003):

GENETIC ALGORITHM (GA)

Genetic algorithms are stochastic algorithms whose search methods are based on the mechanics of natural selection and natural genetics. John Holland (1975) originally developed them. The aim of Holland's work was to develop a theory of adaptive systems that retain the mechanisms of natural systems. The features of natural systems of self-repair, self-guidance and reproduction intrigued early researchers in this field to be applied in problem solving. Problem solving can be thought of as a search through a space of potential solutions. The desired output of such a search is the best solution. Thus, this task can be viewed as an

optimization process. Traditional optimization methods such as *hill climbing* have been used in many applications but they require the existence of the derivative of an objective function and continuity over its domain. Random search optimization schemes have also been used, but they lack efficiency. These conventional optimization schemes are not robust for a broad field of problems. Genetic algorithms try to overcome the problem of robustness by being a directed search process using random choice as a tool.

BASIC CONCEPT

In a genetic algorithm, the first step is to define and code the problem to be solved. A typical single-variable optimization problem can be outlined as

$$Maximize\ g(x) = x^2 \qquad (4)$$

$$Variable\ bound: x_{min} \le x \le x_{max}$$

The problem is defined with the use of an objective function that indicates the *fitness* of any potential solution, and for the above problem is x^2. The decision variables are coded as a finite length string called *chromosome*,

$$A = a_h a_{h-1} a_{h-2} a_{h-3} \cdots a_1 \qquad (5)$$

, where h is the string length. The *alphabet* of a coding defines the possible values of the bit or *gene*:i, i.e. in a binary coding the alphabet is $\{0, 1\}$. For example, if four-bit binary strings are used to code the variable x, the string $(0\ 0\ 0\ 0)$ is decoded to the value x_{min}, the string $(1\ 1\ 1\ 1)$ is decoded to the value x_{max}, and any other string is decoded to a value in the range (x_{min}, x_{max}), uniquely.

In natural terminology, the values of the alphabet are called *alleles* and the position of the gene, indicated by i, is called *locus*. The choice of the string length l and the alphabet determine the accuracy of the solution and the computation time required to solve the problem, Deb (1996). The principle of minimal alphabets defines that the smallest alphabet that permits a natural expression of the problem should be selected.

Genetic Algorithms begin with a population of chromosomes created randomly. Following, the initial population is evaluated. Three main operators -*reproduction, crossover* and *mutation*- are used to evolve the initial population towards better solutions. The population is evaluated, and if the termination criteria are not met, the three main operators are applied again. One cycle of these operators and the evaluation procedure is known as a *generation* in GA terminology. More details about genetic schemes can be found in (Back., 1996; Chen and Smith., 1999; Coli et.al., 1996; Goldberg., 1989; Mitchell., 1999; and Michalewicz., 1996).

The GA procedure is encoded as follows:

Start
Choose a coding to represent variables;
Initialize the population;
Evaluate the population;
Repeat
Reproduction;
Crossover;
Mutation;
Evaluate the population;
Until termination criteria;
End.

ESTIMATION PROCEDURE BASED ON GMGH

The method used to solve the time series prediction problem is based on the Group Method of Data Handling (GMDH) method proposed by Farlow (1984)) and was modified by Ivakhnenko(1994). The approach produces mathematical models of complex systems by handling data samples of observation. It is based on the sorting –out procedure, which chooses a set of models-candidates, in accordance with given external criteria on a separate part of data samples. Thus, GMDH algorithms solve the argument

$$\tilde{g} = arg \min_{g \in G} CR(g)$$

where G is the set of candidate models and $CR(g)$ is an external criterion of model quality.

Most GMDH algorithms are polynomial reference functions which form the set of candidate models. All possible combinations of pairs of the independent variables are formed to determine the parameters of the function used to evaluate the output of each node. The function used to evaluate the output of a node employed Kolmogorov-Gabor theorem can show that any function, $y = f(\vec{x})$ can be represented

$$y = a_0 - \sum_i a_i x_i - \sum_i \sum_j a_{ij} x_i x_j - \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k - \cdots \qquad (6)$$

, where x_i is the independent variable in the input variable vector $\bar{x}$ and $\bar{a}$ is the coefficient vector. GMDH algorithms are then used to determine the coefficients and terms of the reference functions used to partially described system. GMDH algorithms are multi-layered, and at each layer the partial description is obtained and it is conveyed to the next layers to gradually obtain the final model of the complex system. Combinatorial GMDH algorithm (COMBI) is the simplest of GMDH algorithms which can be employed to estimate the coefficient of the models. First n observations of regressions – type data are taken. These observations are divided into sets: training set and the validating set. The training set consists of m observations, while the validation in made of n-m observations. As in original GMDH concept, COMBI algorithm is multi-layered; at each layer, it obtains a candidate model of the system and once the models of each layer are obtained, the best one is chosen to be the output model. The first layer is obtained by using the information contained in every column of the training sample of observations. The candidate models for the first layer have the form

$$y = a_0 - a_1 x_i, \; i = 1,2 \ldots\ldots p \qquad (7)$$

To obtain the values of the coefficients a_0 and a_1 for each of the n models, a system of Gauss normal equations is solved. In the case of the first layer, the system of Gauss normal equation for the i^{th} model will be

$$\begin{bmatrix} m & \sum_{i=1}^{m} x_{ij} \\ \sum_{i=1}^{m} x_{ij} & \sum_{i=1}^{m} x_{ij}^2 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_{ij} y_i \end{bmatrix} \qquad (8)$$

,where m is the number of observations in training set.

After all possible models from this layer have been formed, the one with the minimum *regularity criterion AR*(*s*) [68] is chosen. The regularity criterion is defined by the formula

$$AR(s) = \frac{1}{m} \sum_{i=n-m+1}^{n} (y_i - \hat{y}_i)^2 \qquad (9)$$

where m is the number of observations in the validation set, *n* is the total number of observations, $\hat{y}$ is the estimated output value and *s* is the model whose fitness is evaluated.

A small number of variables that give the best results in the first layer, are allowed to form second layer candidate models of the form

$$y = a_0 - a_1 x_i - a_2 x_j, \; i,j = 1,2 \ldots\ldots p \qquad (10)$$

Models of the second layer are evaluated for compliance with the criterion, and again the variables that give best results will proceed to form third layer candidate models. This procedure is carried out as long as the criterion decreases in value, and candidate models at the *p*th layer will have the form

$$y = a_0 - a_1 x_i - a_2 x_j - \ldots - a_p x_l, i,j,l = 1,2, \ldots\ldots p$$

After the best models of each layer have been selected, the output model is selected by the *discriminating criterion* termed as y^*. A possible discriminating criterion is the *variation criterion RR*(*s*) proposed by Belogurov V. P. (1990) defined as

$$RR(s) = \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2} \qquad (11)$$

, where $\bar{y}$ is the mean output value and *s* is the model whose fitness is evaluated. The model with the minimum value of the variation criterion *RR*(*s*) is selected as the output model. Other discriminating criteria can be used that make a

compromise between the accuracy and complexity of a model.

MEASURING PREDICTION ACCURACY

Mean Square Error (MSE) & Root Mean Square Error (RMSE)

The mean square error measures the prediction accuracy. It averages the sizes of prediction errors avoiding the cancelling of positive and negative terms. The MSE, instead of using the absolute value of the prediction

errors, uses their square value. The advantage of using the square value of the prediction errors is that it gives more weight to large prediction errors. The formula for the mean square error is

$$MSE = \frac{\sum (f(x) - \hat{f}(x))^2}{N} \quad (12)$$

, where N is the number of records in the data set
The MSE is measured in the squares of the units of the original series, which makes it harder to be interpreted. For this reason, the root mean square error can sometimes be evaluated, that is given simply by the equation

$$RMSE = \sqrt{MSE} \quad (13)$$

and is measured in the same units as the original time series.

4.2 Mean Absolute Percent Error (MAPE)

The mean absolute percent error (MAPE) is a unit-free evaluation measurement. This allows the comparison of the accuracy of the same or different models on different time series. It is evaluated by expressing each prediction error as a percentage according to actual value of the time series according to the formula [20], Farnum N. R. and Stanton L. W. ((1989)

$$MAPE = \frac{\sum \left| \frac{e_i}{f(x)} \right|}{N} 100\% \quad (14)$$

, where N is the number of records in the data set.

APPLICATION TO NAIRA EXCHANGE RATES

The data set used to test the method was obtained from CBN statistical bulletin of 2006, is the exchange rates for the U.S dollar, British pound, , the Japanese Yen, the CFA franc and the Swiss franc against the Nigerian Naira. The data used are the yearly Autonomous Foreign Exchange Market (AFEM) observed exchange rates for all the above mentioned currencies from the 31st of December 1995 to the 31st of December 2006. The data were again normalised to take values from zero to one, before they were used as inputs to the polynomial neural networks. The predictions of the exchange rates were based on the last three vales in the series. Thus, the output pattern is

$x(t) = f(x(t-1), x(t-2), x(t-3))$

The COMBI algorithm converged to networks with only on layer, to model the exchange rates of all the currencies against Nigerian Naira. The models for the exchange rates of each currency are given below.

U.S dollar against Nigerian Naira

$y = 0.171 + 0.979x_2 - 0.576x_3 - 1.03x_1^2 - 0.86x_2^2$

MSE= 0.0058 , R-Sq = 99.6%

British pound against Nigerian Naira

$y = -0.183 - 0.989x_2 + 0.576x_3 - 0.342x_2x_3 + 0.04786x_1^2$

MSE= 0.00799, R-Sq = 99.3%

Japanese Yen against Nigerian Naira

$y = 0.1493 + 0.894x_2 - 0.0288x_3 - 0.4112x_1^2 - 0.06147x_2^2$

MSE = 0.03744, R-Sq = 96.6%

CFA franc against Nigerian Naira

$y = -0.871 - 7.19x_2 - 1.426x_3 + 3.373x_1^2 - 5.04x_2^2$

MSE=1.212, R-Sq = 69.7%

Swiss franc against Nigerian Naira

$y = -0.174 - 7.19x_2 - 1.426x_3 + 3.373x_1^2 - 5.04x_2^2$

MSE=0.1108, R-Sq = 84.8%

The results obtained with COMBI method are shown above for U.S dollars, British Pound, Japanese Yen, CFA franc and Swiss franc respectively. It should be noted that the models obtained by COMBI algorithm

used past variables x_2 and x_3, except for the British pound exchange rates. The mean square error and its associated coefficient of determination (R^2) were presented under the model. The model for U.S dollar exchange rate against Naira is the best which explained about 99.6% of the variation and CFA franc exchange rate is the worst, as explains 69.7 % of the variation.

CONCLUSION

The result obtained is very encouraging and showed the applicability and suitability of the genetic algorithm for Time series prediction.

REFERENCE

Back T.(1996), Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford University Press.

Belogurov V. P. (1990), A criterion of model suitability for forecasting quantitative processes, Soviet Journal of Automation and Information Sciences, Vol. 23, No. 3, pp. 21-25.

CBN(2006) Central Bank of Nigeria, Statistical Bulletin vol.17. December 2006 pp 233.

Chen S. and Smith S. F.(1999), Putting the "genetics" back into genetic algorithms, Foundations of Genetic Algorithms 5, Morgan Kaufmann.

Coli M., Gennuso G. and Palarazzi P.(1996), A new crossover operator for genetic algorithms, Proceedings of 1996 IEEE International Conference on Evolutionary Computation (ICEC '96), pp. 201-206.

Deb, K. (1996), Genetic algorithms for function optimization, Genetic Algorithms and Soft Computing, pp. 3-29.

Farlow S. J. (1984), The GMDH algorithm, Self-Organizing Methods in Modeling, pp. 1-24.

Farnum N. R. and Stanton L. W. ((1989)., Quantitative forecasting methods, PWS-KENT.

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

Gottman, J.M. (1981), Time-series analysis A comprehensive introduction for social scientists, Cambridge University Press, Cambridge, USA.

Holland J. H. (1975), Adaptation in natural and artificial systems, University of Michigan Press.

Hsieh, W.W. (2004), Nonlinear multivariate and time series analysis by neural network methods, Reviews of Geophysics, 42, RG1003, doi:10.1029/2002RG000112.

Ivakhnenko A. G (1994), An inductive sorting method for the forecasting of multidimensional random processes and events with the help of analogues forecast complexing, Pattern Recognition and Image Analysis, Vol. 4, No. 2, pp. 177-188.

Jenkins, G.M., and D.G. Watts (1968), Spectral analysis and its applications, Holden-Day, San Francisco, USA.

Kantz, H., and T. Schreiber (1997), Nonlinear time series analysis, Cambridge University Press, Cambridge, UK.

Michalewicz Z.(1996), Genetic algorithms + data structures = evolution programs, 3rd edition, Springer-Verlag.

Mitchell M.(1999), An introduction to genetic algorithms, MIT Press.

Schlittgen, R. and Streitberg, B.H.J. (1984), Zeitreihenanalyse, R. Oldenburg Verlag, Munich, Germany.
Strickert, M. (2003), Time Series and Data Analysis, Lecture script, Applied System Science, University of Osnabrück, Germany, Internet-resource :http://luna2.informatik.uni-osnabrueck.de/marc/lectures/zra_ss03/scriptparts/ (04.10.2005)