Research Article

DNA Optimization Threshold Autoregressive Prediction Model and Its Application in Ice Condition Time Series

Xiao-Hua Yang¹ and Yu-Qi Li²

¹ State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, China

² School of Geography and Remote Sensing Science, Beijing Normal University, Beijing 100875, China

Correspondence should be addressed to Xiao-Hua Yang, xiaohuayang@bnu.edu.cn

Received 24 August 2011; Accepted 18 September 2011

Academic Editor: Carlo Cattani

Copyright © 2012 X.-H. Yang and Y.-Q. Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There are many parameters which are very difficult to calibrate in the threshold autoregressive prediction model for nonlinear time series. The threshold value, autoregressive coefficients, and the delay time are key parameters in the threshold autoregressive prediction model. To improve prediction precision and reduce the uncertainties in the determination of the above parameters, a new DNA (deoxyribonucleic acid) optimization threshold autoregressive prediction model (DNAOTARPM) is proposed by combining threshold autoregressive method and DNA optimization method. The above optimal parameters are selected by minimizing objective function. Real ice condition time series at Bohai are taken to validate the new method. The prediction results indicate that the new method can choose the above optimal parameters in prediction model (IGATARPM) and standard genetic algorithm threshold autoregressive prediction model (SGATARPM), DNAO-TARPM has higher precision and faster convergence speed for predicting nonlinear ice condition time series.

1. Introduction

Many natural phenomena, such as ice condition, runoff, are usually nonlinear, complex, and dynamic processes. Prediction of ice conditions is of primary importance for weather forecasting, agriculture, geosciences, and marine transportation safety. The simulation of the nonlinear time series was very difficult with the traditional deterministic mathematic models, which cause new challenges to calibrate the parameters [1, 2]. There are many methods for predicting nonlinear time series [3–10]. Threshold autoregressive (TAR) models are typically applied to time series data as an extension of autoregressive models for higher degree of flexibility in model parameters through a regime switching behavior. TAR models were introduced by Tong and Li in 1977 and more fully developed in the seminal paper [11]. The threshold autoregressive model is a special case of Tong's general threshold autoregressive models. The latter allows the threshold variable to be very flexible, such as an exogenous time series in the open-loop threshold autoregressive system [11–13]. For a comprehensive review of developments over the 30 years since the birth of the model, see Tong [14]. However, the uncertainties in determining the parameters of the threshold autoregressive coefficients, and the delay time exist in the developed threshold autoregressive model. So as to improve the prediction accuracy, the key problem is how to determine the parameters in the prediction model.

The global optimization in determining all the parameters is intractable mathematically. Once an objective function has many local extreme points, the traditional optimization methods may not obtain the global optimal solution. A genetic algorithm (GA) based on the genetic evolution of a species was proposed by Holland [15]. GA is a global optimization algorithm. However, the computational amount is very large and premature convergence phenomena exist in GA [16–20]. Recently, Adleman [21] showed that DNA can be used to solve a computationally hard problem. Many scientists used DNA computation to solve real problems [22–24].

In this study, DNA optimization threshold autoregressive prediction model (DNAO-TARPM) is presented to determine the parameters and to improve the calculation precision for predicting ice condition time series. In order to validate the new method, some real ice condition time series are used.

2. DNA Optimization Threshold Autoregressive Prediction Model (DNAOTARPM)

The TAR model is a tool for predicting future values in time series assuming that the behavior of the time series changes once the time series shifts to a different regime. The switch from one regime to another depends on the p past values of the x series. The model consists of k autoregressive (AR) parts for each different regime. The model is usually referred to as the TAR (k, p) model where k is the number of regimes and p is the order of the autoregressive part. Since those can differ between regimes, the p portion is sometimes dropped and models are denoted simply as TAR (k). A k-regime TAR (d; p_1 , p_2 , ..., p_k) model for time series x(i) (i = 1, 2, ..., n) has the form

$$x(i) = b(j,0) + \sum_{l=1}^{p_j} b(j,l) x(i-l) + e(j,i), \quad x(i-d) \in (r(j-1), r(j)], \quad (j = 1, 2, \dots, k),$$
(2.1)

where $r(0) = -\infty$, $r(k) = +\infty$, r(j) (j = 1, 2, ..., k-1) are k-1 nontrivial threshold parameters dividing the domain into k different regimes; d is the delay time parameters, b(j, l) is the regressive coefficients in the jth regime, e(j, i) stands for white-noise error term with constant variance, and p_j is the autoregressive order in the jth regime of the model. The threshold parameters satisfy the constraint:

$$-\infty = r(0) < r(1) < r(2) < \dots < r(k-1) < r(k) = +\infty.$$
(2.2)

Mathematical Problems in Engineering

Here *d*, *k*, r(1), r(2),..., r(k - 1), p_1 , p_2 ,..., p_k , and b(j, l) are parameters in TAR model. It is very difficult to determine these parameters with the traditional methods.

In this paper, we use DNA optimization method to determine the parameters and improve model accuracy. The new model, DNA optimization threshold autoregressive prediction method (DNAOTARPM), is described as follows.

Step 1 (Determine the delay time d and the number of regressive coefficients). The delay time d is determined by the autocorrelation function method [21]. The autocorrelation function R(j) for delay time j is calculated as

$$R(j) = \frac{\sum_{i=j+1}^{n} (x(i) - m_x) (x(i-j) - m_x)}{\sum_{i=1}^{n} (x(i) - m_x)^2},$$

$$m_x = \sum_{i=1}^{n} \frac{x(i)}{n}.$$
(2.3)

The delay time *d* is selected when autocorrelation function R(j) [25] satisfies the following condition:

$$R(j) \notin [R1(j), R2(j)],$$
 (2.4)

$$R1(j) = \frac{-1 - u_{\alpha/2} \cdot (n-j-1)^{0.5}}{n-j}, \qquad R2(j) = \frac{-1 + u_{\alpha/2} \cdot (n-j-1)^{0.5}}{n-j}, \qquad (2.5)$$

where $u_{\alpha/2}$ is the upper $100 \cdot (\alpha/2)$ percentage point of the normal distribution for $1 - \alpha$ confidence level. The number of regressive coefficients $p_l \leq \max(j)$, l = 1, 2, ..., k. Some of the *j* values are regarded as the delay time.

Step 2 (Determine the number and ranges of threshold parameters). Considering a set $\{(x(i), x(i-d)) \mid i = 1, 2, 3, ...; d = 1, 2, ...\}$ from the time series (*i*) (*i* = 1, 2, ..., *n*), we divide x(i-d) into *s* regimes (s > k). Suppose there are N_j number of x(i-d) in the *j*th part, and the corresponding x(i) is regarded as x(i, j). In the *j*th part, the conditional expectation of x(i) given the event X = x(i-d) is

$$E\left(\frac{x(i)}{x(i-d)}\right)_{j} = \sum_{i=1}^{N_{j}} \frac{x(i,j)}{N_{j}}, \quad (j = 1, 2, \dots, s; \ d = 1, 2, \dots).$$
(2.6)

Let x(i - d) be horizontal axis, and let E(x(i)/x(i - d)) be vertical axis; we can get the scatter plots. When the scatter plots are piecewise linear map, we can estimate the number and ranges of threshold parameters. The piecewise number of piecewise linear map is the number of threshold parameters, and the ranges of the piecewise points are the ranges of threshold parameters.

Step 3 (Construct the objective function). The parameter estimation for DNAOTARPM can be obtained by the following objective function, namely, the mean of least residual absolute value sum:

$$\operatorname{Min} f = f(r(1), r(2), \dots, r(k-1); b(j,l)) = \frac{\sum_{i} |(x(i) - x'(i))|}{n}.$$
(2.7)

Step 4 (Solve objective function by DNA optimization method). Solving the parameters of $r(1), r(2), \ldots, r(k-1)$; b(j, l), $(j = 1, 2, \ldots, k; l = 1, 2, \ldots, p_j)$, in the optimization objective function (2.7) is one nonlinear optimization problem. It is rather difficult to deal with it using a traditional optimization method. The above optimal model can be solved by the following DNA optimization method [24]. The *k*-regime prediction formula will be seen in the following application part in detail.

If we solve objective function (2.7) with improved genetic algorithm, we call the method improved genetic algorithm [18] threshold autoregressive prediction method (IGATARPM), and if we solve objective function (2.7) with standard genetic optimization method [15], we call the method standard genetic algorithm threshold autoregressive prediction method (SGATARPM).

3. DNA Optimization Method (DNAOM)

Consider the following optimization problem:

Min
$$f(c_1, c_2, ..., c_p)$$

st. $a_j \le c_j \le b_j$ for $j = 1, 2, ..., p$, (3.1)

where $c = \{c_j, j = 1, 2, ..., p\}$, c_j is a parameter to be optimized, f is an objective function, and $f \ge 0$, $[a_i, b_j]$ is the range of c_j .

The procedure of DNAOM is shown as follows [25].

Step 1 (DNA encoding). Suppose DNA-encoding length is *m* in every parameter, the *j*th parameter range is the interval $[a_j, b_j]$, and then each interval is divided into $2^m - 1$ subintervals:

$$c_j = a_j + \theta_j \cdot (2^m - 1) \cdot h_j, \tag{3.2}$$

where the length of subinterval of the *j*th parameter $h_j = (b_j - a_j)/(2^m - 1)$ is constant. The searching location $I_j = \theta_j \cdot (2^m - 1)$ is an integer, and $0 \le I_j < 2^m$, θ_j is a random variable, and $0 \le \theta_j \le 1$, for j = 1, 2, ..., p.

The DNA code array of the *j*th parameter is denoted by the grid points of $\{d(j,k)|k = 1, 2, ..., m\}$ for every individual:

$$I_j = \sum_{k=1}^m d(j,k) \cdot 2^{k-1}.$$
(3.3)

Mathematical Problems in Engineering

DNAOM's process operates on a population of individuals (also called DNA code array, strings, or chromosomes). Each individual represents a potential solution to the problem. For corresponding $1 \times 1 \rightarrow T$, $1 \times 0 \rightarrow A$, $0 \times 1 \rightarrow G$, $0 \times 0 \rightarrow C$. The first position value "1" or "0" expresses the position of DNA code and the second position value "1" or "0" expresses the true value of binary code and the value of DNA code.

Step 2 (creating the initial population). To cover the whole solution space and to avoid individuals entering into the same region, large uniformity random population is selected in this algorithm. Once the initial father population has been generated, the decoding and fitness evaluation should be done.

Step 3 (evaluating fitness value of each individual). The smaller the value f(i) is, the higher the fitness of its corresponding *i*th chromosome is (i = 1, 2, ..., N). So the fitness function of *i*th chromosome is defined as follows:

$$F(i) = \frac{1}{\left[f(i)\right]^2 + 1.0}.$$
(3.4)

Step 4 (selection). Select chromosome pairs randomly depending on their fitness value from the initial population. Two groups of *N*-chromosomes $d_1(j,k,i)$ and $d_2(j,k,i)$, are gotten (j = 1, 2, ..., p; k = 1, 2, ..., m; i = 1, 2, ..., N).

Step 5 (two-point crossover and two-point mutation). Perform crossover and mutation on chromosomes the same as GA.

Step 6 (DNA evolution). Repeat Steps 3–6 until the evolution times q = Q (Q is the total evolution times) or the termination condition is satisfied.

Step 7 (accelerating cycle). The parameter ranges of n_e -excellent individuals obtained by Q-times of the DNA-encoded optimal evolution alternating are regarded as the new ranges of the parameters, and then the whole process is back to the DNA-encoding. The DNAOM computation is over until the algorithm running times reaches the designed T times or there exists an optimal chromosome C_{fit} whose fitness satisfies a given criterion. In the former case the C_{fit} is the fittest chromosome or the most excellent chromosome in the population. That is, the chromosome C_{fit} represents the solution [25].

The parameters of the DNAOM are selected as follows. The length m = 10, population size N = 100, the number of excellent individuals $n_e = 10$, the times of evolution alternating Q = 3, the crossover probability $p_c = 1.0$, and the mutation probability $p_m = 0.5$.

4. Application in Ice Condition Time Series

The real ice condition time series in this study are chosen as the annual ice condition at Bohai in China for the period of 1966 to 1994 (29 years) [25]. For the ice condition time series, the first modeling data set is the data during the period of 1966 to 1993 (28 years). The prediction lead time is the year of 1970–1994 (25 years).



Figure 1: The autocorrelation function figure for the observed time series.

4.1. The Autocorrelation Function R(j) for Delay Time j

The changes of the autocorrelation functions for the time series are presented at the confidence level 70% in Figure 1.

From Figure 1, we can see that only the values of R(1), R(3), R(4) satisfy condition (2.4). So the delay time *d* is 1, 3, or 4 in DNAOTARPM.

4.2. The Number and Ranges of Threshold Parameters

The number and ranges of threshold parameters of the above ice condition time series are determined by the conditional expectation of x(i) given the event X = x(i - d). The scatter plot of the conditional expectation is shown in Figure 2.

From Figure 2, we can see that there are two piecewise linear maps, and the piecewise point is around the mean value of the time series. So we suppose y(i) = x(i) – mean value, and the *k*-regime TAR(*d*; $p_1, p_2, ..., p_k$) model has the following form for d = 1, 3, 4:

$$y'(i) = \begin{cases} b(1,1)y(i-1) + b(1,3)y(i-3) + b(1,4)y(i-4), y(i-d) \le r(1), \\ b(2,1)y(i-1) + b(2,3)y(i-3) + b(2,4)y(i-4), y(i-d) > r(1), \\ x'(i) = y'(i) + \text{mean value.} \end{cases}$$
(4.1)

The parameters of r(1), b(j, l) (j = 1, 2; l = 1, 3, 4) are required in this model. In this work, the three parameters are estimated with respect to one criterion, namely, the mean of least residual absolute value sum shown in (2.6).



Figure 2: The scatter plot of the conditional expectation: (a) E(x(i)/x(i-1)): x(i-1); (b) E(x(i)/x(i-3)): x(i-3); (c) E(x(i)/x(i-4)): x(i-4).

4.3. Result Comparison between DNAOTARPM, IGATARPM, and SGATARPM

The time series were predicted by DNAOTARPM, IGATARPM, and SGATARPM, respectively.

Mean least residual absolute value sum f is 0.5737 for DNAOTARPM. The evaluation number of the objective function is 900. The computational results of the above model are given in Table 1.

For IGATARPM, the evaluation number of the objective function is 2700, and the prediction error f is 0.6016.

For SGATARPM, the evaluation number of the objective function is 2700, and the prediction error f is 0.6380.

From Table 1, we can see that prediction results for DNAOTARPM are better than those with the other methods. The prediction results of the practical example are shown in Figure 3 with different methods.

From Table 1 and Figure 3, we can see that the results achieved with our DNAO-TARPM method are satisfactory in global optimum and prediction precision.

| Methods | Parameters | | | | | | | Objective function minimum f |
|------------|-------------|------------|------------|------------|------------|------------|------------|------------------------------------|
| | r(1) | b(1,1) | b(1,3) | b(1, 4) | b(2,1) | b(2,3) | b(2, 4) | |
| | [-1.0, 1.0] | [-0.5,0.5] | [-0.5,0.5] | [-0.5,0.5] | [-0.5,0.5] | [-0.5,0.5] | [-0.5,0.5] | |
| DNAO-TARPM | -0.02 | 0.34 | 0.24 | -0.36 | 0.13 | 0.39 | -0.46 | 0.5737 |
| IGA-TARPM | -0.75 | 0.12 | 0.40 | -0.20 | 0.20 | 0.28 | -0.44 | 0.6016 |
| SGA-TARPM | 0.90 | 0.26 | 0.33 | -0.50 | -0.06 | -0.35 | 0.36 | 0.6380 |

Table 1: The comparison of the prediction results for DNAOTARPM, IGATARPM, and SGATARPM at



Figure 3: Comparison of prediction results with DNAOTARPM, IGATARPM, and SGATARPM at Bohai.

Compared with IGATARPM and SGATARPM, DNAOTARPM has a faster convergence speed and higher precision. And it is useful for parameter optimization of the nonlinear ice condition time series model.

5. Conclusions

In order to improve prediction precision and reduce the uncertainties in determination of the parameters for forecasting nonlinear ice condition time series, a new DNA optimization threshold autoregressive prediction model (DNAOTARPM) is proposed in this paper. The ice condition time series at Bohai in China are studied by using DNAOTARPM. The main conclusions are given as follows.

(1) DNAOTARPM is established by using DNA optimization method and threshold autoregressive model. The delay time *d* is selected with autocorrelation function, and the results indicate x(i - 1), x(i - 3), x(i - 4) have significant influence on the ice condition time series ($\alpha = 0.30$) at Bohai.

Bohai.

Mathematical Problems in Engineering

- (2) DNA optimization method is proposed for optimizing all parameters in DNAO-TARPM model. The optimal parameters, that is, the threshold value, autoregressive coefficients, and the delay time, are obtained for predicting the ice condition time series at Bohai by using DNA optimization method.
- (3) The prediction errors are 0.5737, 0.6016, and 0.6380 with DNAOTARPM, IGATARPM, and SGATARPM at Bohai, respectively. DNAOTARPM can reduce the calculation errors. It provides a new way to forecast nonlinear ice condition time series.
- (4) The evaluation number of the objective function is 900, 2700, and 2700 with DNAOTARPM, IGATARPM, and SGATARPM at Bohai, respectively. Compared with IGATARPM and SGATARPM, DNAOTARPM model has a faster convergence speed. The new model (DNAOTARPM) can be used in predicting other nonlinear systems in the future and its theory will be further studied.

Acknowledgments

The present work is supported by the Program for the National Basic Research Program of China (no. 2010CB951104), the Project of National Natural Science Foundation of China (nos. 50939001 and 51079004), the Specialized Research Fund for the Doctoral Tutor Program of Higher Education (no. 20100003110024), and the Program for Changjiang Scholars and Innovative Research Team in University (no. IRT0809).

References

- [1] X. H. Yang, D. X. She, Z. F. Yang, Q. H. Tang, and J. Q. Li, "Chaotic Bayesian method based on multiple criteria decision making (MCDM) for forecasting nonlinear hydrological time series," *International Journal of Nonlinear Sciences and Numerical Simulation*, vol. 10, no. 11-12, pp. 1595–1610, 2009.
- [2] M. Li, "Modeling autocorrelation functions of long-range dependent teletraffic series based on optimal approximation in Hilbert space-A further study," *Applied Mathematical Modelling*, vol. 31, no. 3, pp. 625–631, 2007.
- [3] C. T. Cheng, C. P. Ou, and K. W. Chau, "Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall-runoff model calibration," *Journal of Hydrology*, vol. 268, no. 1–4, pp. 72–86, 2002.
- [4] W. Jia, B. Ling, K. W. Chau, and L. Heutte, "Palmprint identification using restricted fusion," Applied Mathematics and Computation, vol. 205, no. 2, pp. 927–934, 2008.
- [5] K. W. Chau, "Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River," *Journal of Hydrology*, vol. 329, no. 3-4, pp. 363–367, 2006.
- [6] N. Muttil and K. W. Chau, "Neural network and genetic programming for modelling coastal algal blooms," *International Journal of Environment and Pollution*, vol. 28, no. 3-4, pp. 223–238, 2006.
- [7] C. T. Cheng, K. W. Chau, and Y. Z. Pei, "A hybrid adaptive time-delay neural network model for multi-step-ahead prediction of sunspot activity," *International Journal of Environment and Pollution*, vol. 28, no. 3-4, pp. 364–381, 2006.
- [8] E. I. Debolskaya, "Numerical modeling the turbulent structure of flows under ice," Water Resources, vol. 27, no. 2, pp. 144–151, 2000.
- [9] E. I. Debolskaya, V. K. Debolskii, and M. V. Derbenev, "Numerical modeling of pollutant distribution at catastrophic floding in conditions of ice difficulties," *Water Resources*, vol. 34, no. 6, pp. 673–681, 2007.
- [10] A. M. Wasantha Lai and H. T. Shen, "A mathematics model for river ice processes," *Journal of Hydraulic Engineering*, vol. 117, no. 7, pp. 851–867, 1991.
- [11] H. Tong and K. S. Li, "Threshold autoregression, limit cycles and cyclical data," *Journal of the Royal Statistical Society, Series B*, vol. 42, pp. 245–292, 1980.

- [12] H. Tong, Threshold Models in Nonlinear Time Series Analysis, vol. 21 of Lecture Notes in Statistics, Springer, New York, NY, USA, 1983.
- [13] H. Tong, "Birth of the threshold time series model," Statistica Sinica, vol. 17, no. 1, pp. 8–14, 2007.
- [14] H. Tong, "Threshold models in time series analysis—30 years on (with discussions by P. Whittle, M. Rosenblatt, B. E. Hansen, P. Brockwell, N. I. Samia and F. Battaglia)," *Statistics and Its Interface*, vol. 4, pp. 107–136, 2011.
- [15] J. H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, Mich, USA, 1975.
- [16] K. Chau, "Calibration of flow and water quality modeling using genetic algorithm," in AI: Advances in Artificial Intelligence, vol. 2557 of Lecture Notes in Computer Science, p. 720, Springer, Berlin, Germany, 2002.
- [17] K. W. Chau, "A two-stage dynamic model on allocation of construction facilities with genetic algorithm," Automation in Construction, vol. 13, no. 4, pp. 481–490, 2004.
- [18] J. Andre, P. Siarry, and T. Dognon, "An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization," *Advances in Engineering Software*, vol. 32, no. 1, pp. 49–60, 2001.
- [19] X. H. Yang, Z. F. Yang, and Z. Y. Shen, "GHHAGA for environmental systems optimization," Journal of Environmental Informatics, vol. 5, no. 1, pp. 36–41, 2006.
- [20] X. H. Yang, Z. F. Yang, G. H. Lu, and J. Li, "A gray-encoded, hybrid-accelerated, genetic algorithm for global optimizations in dynamical systems," *Communications in Nonlinear Science and Numerical Simulation*, vol. 10, no. 4, pp. 355–363, 2005.
- [21] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," Science, vol. 266, no. 5187, pp. 1021–1024, 1994.
- [22] C. W. Yeh, C. P. Chu, and K. R. Wu, "Molecular solutions to the binary integer programming problem based on DNA computation," *BioSystems*, vol. 83, no. 1, pp. 56–66, 2006.
- [23] R. J. Lipton, "DNA solution of hard computational problems," Science, vol. 268, pp. 542–545, 1995.
- [24] X. H. Yang and J. Q. Li, "DNA accelerating evolutionary algorithm and its application in the parameter optimization of Muskingum routing model," in *Proceedings of the The 6th International Conference on Natural Computation and the 7th International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 8, pp. 3914–3917, Yantai, China, 2010.
- [25] X. H. Yang and Z. Y. Shen, Intelligent algorithms and their applications in resource and environment system modeling, Beijing Normal University Press, Beijing, China, 2005.



Advances in **Operations Research**

The Scientific

World Journal





Mathematical Problems in Engineering

Hindawi

Submit your manuscripts at http://www.hindawi.com



Algebra



Journal of Probability and Statistics



International Journal of Differential Equations





International Journal of Combinatorics

Complex Analysis









Journal of Function Spaces



Abstract and Applied Analysis





Discrete Dynamics in Nature and Society