ELSEVIER

Energy Policy 35 (2007) 5229–5241

ENERGY POLICY

www.elsevier.com/locate/enpol

Integration of genetic algorithm, computer simulation and design of experiments for forecasting electrical energy consumption

A. Azadeh*, S. Tarverdian

Department of Industrial Engineering, Center of Excellence for Intelligent Experimental Mechanics, Department of Engineering Optimization and Research Institute of Energy Management and Planning, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Iran

Received 22 January 2007; accepted 16 April 2007
Available online 28 June 2007

Abstract

This study presents an integrated algorithm for forecasting monthly electrical energy consumption based on genetic algorithm (GA), computer simulation and design of experiments using stochastic procedures. First, time-series model is developed as a benchmark for GA and simulation. Computer simulation is developed to generate random variables for monthly electricity consumption. This is achieved to foresee the effects of probabilistic distribution on monthly electricity consumption. The GA and simulated-based GA models are then developed by the selected time-series model. Therefore, there are four treatments to be considered in analysis of variance (ANOVA) which are actual data, time series, GA and simulated-based GA. Furthermore, ANOVA is used to test the null hypothesis of the above four alternatives being equal. If the null hypothesis is accepted, then the lowest mean absolute percentage error (MAPE) value is used to select the best model, otherwise the Duncan Multiple Range Test (DMRT) method of paired comparison is used to select the optimum model, which could be time series, GA or simulated-based GA. In case of ties the lowest MAPE value is considered as the benchmark. The integrated algorithm has several unique features. First, it is flexible and identifies the best model based on the results of ANOVA and MAPE, whereas previous studies consider the best-fit GA model based on MAPE or relative error results. Second, the proposed algorithm may identify conventional time series as the best model for future electricity consumption forecasting because of its dynamic structure, whereas previous studies assume that GA always provide the best solutions and estimation. To show the applicability and superiority of the proposed algorithm, the monthly electricity consumption in Iran from March 1994 to February 2005 (131 months) is used and applied to the proposed algorithm.

Keywords: Genetic algorithm; Simulation; Electricity consumption

1. Introduction

Increasing worldwide demand for electricity requires development of exact intelligent methods and algorithm for its forecasting. The estimation of electrical energy demand based on economic and non-economic indicators may be achieved by certain statistical, mathematical and simulation models. These forecasting models might be linear or non-linear. Due to the fluctuations of electricity demand indicators, the non-linear forms of the equations could estimate electrical energy demand more effectively. The non-linearity of these indicators and electrical energy demand has lead to search for intelligent solution approach methods such as neural networks, fuzzy regression and genetic algorithms (GAs). GAs are optimizing and stochastic search techniques which possess vast and powerful applications (Azadeh et al., 2006). They consider a solution space and move intelligently towards the best solution while they are able to be trained by the data available and estimate for the kept part of the data called the trial period. The power of GA has recently been noticed due to its powerful search for identification of optimum parameters. The case of this study is concerned with monthly electrical consumption and GA is used as a new tool to show its ability against other searching techniques such as ANNs, which were already used for monthly consumption estimations. Computer simulation could also

*Corresponding author.
E-mail addresses: aazadeh@ut.ac.ir, ali@azadeh.com (A. Azadeh).

0301-4215/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.enpol.2007.04.020

be used as an overlapping approach with GA. Furthermore, previous studies mostly use GA for development of electricity demand function. However, this study uses computer simulation to generate random variables to be used in GA, whereas previous studies only use available raw data for GA. Moreover, the integration of GA and simulation is proposed as an alternative forecasting approach in this study and it is compared with GA and time series.

Computer simulation has excellent capabilities such as proper description of system behavior, scenario analysis and forecasting capabilities. Numerous studies have been conducted in domain of GA or computer simulation; however, this is the first study that integrates GA and simulation for forecasting electricity consumption to be used in the proposed algorithm of this paper. Also, in context of forecasting, simulation is an attractive tool because it allows generation of random variables, which could define the complex behavior of input data for forecasting problems. Simulation could help in modeling of past data to be used for electricity demand process with relatively low cost.

The estimation of Turkey's energy demand based on economic indicators using GA was reported by Ceylan and Ozturk (2004). Ozturk et al. (2005) estimated industrial electricity demand using GA. Haldenbilen and Ceylan (2004) estimated transport energy demand in Turkey by GA. Osman et al. (2005) have combined GA with fuzzy logic controller (FLC) so that the search region is able to adapt toward the promising area and the boundary intervals are monitored by FLC and it is modified each time. Bunning and Sun (2005) presented a stochastic global algorithm for solving constrained optimization problems over a compact search domain. Tang et al. (2005) have used a GA-based Takagi–Sugeno–Kang fuzzy neural network to tune the parameters in Takagi–Sugeno–Kang fuzzy neural network. Stach et al. (2005) have proposed a novel learning method that is able to generate fuzzy cognitive maps models from input historical data and without human intervention, the proposed method is based on GA. Muni et al. (2006) proposed genetic programming methodology simultaneously selects a good subset of features and constructs a classifier using the selected features. Ozturk et al. (2004) have carried out some researches recently to estimate the energy input/output values using GA. Hasheminia and Akhavan Niaki (2006) have introduced a new type of GA to find the best regression model among several alternatives and have assessed its performance by an economical case study. Also, recent studies show the integration of GA and neural networks for short-term estimation and prediction of electrical energy consumption (Azadeh et al., 2007b).

Compton and Wu (2005) projected the electricity consumption in China by Bayesian vector auto regression. However, rapidly developing countries like China, Iran and India face with complex requirements for their demands, which require the use of intelligent tools such as GA. Furthermore, there is no clear-cut between conventional approach and intelligent tools such as GA. However, this study presents a framework to integrate conventional time series with GA and computer simulation through design of experiment (DOE) and relative errors obtained from the three approaches (simulated-based GA, GA and time series) versus actual data. The proposed framework always guarantees best solution whereas previous studies assume either the conventional time series or GA leads to the best electricity forecasting estimations.

The proposed algorithm that is based on GA, time series, simulation, and DOE is discussed in the next section. The input variables used to estimate the best-fit model for electricity monthly consumption estimation for both GA, simulated-based GA and time series approaches are the previous 12 months specified by Auto-Correlation Function (ACF) (discussed in the next sections). To show the applicability and superiority of the proposed algorithm, the monthly electricity consumption in Iran from March 1994 to February 2005 (131 months) is used and applied to the proposed algorithm.

2. The integrated algorithm

The proposed algorithm may be used to estimate energy demand in the future by optimizing parameter values. The input variables are specified by ACF and then the best time-series model is fitted to the data set. The proposed algorithm uses analysis of variance (ANOVA) to select either GA or time series or simulated-based GA for future demand estimation. The term simulated-based GA is defined as integration of computer simulation and GA. Moreover, the raw data is simulated by computer simulation to identify its probability distribution and the mean of probability distribution is then used as input data for time series and GA. This is of course repeated for each month. The advantage of simulated-based is to foresee if the stochastic nature of data has any impact on future demand estimation. Furthermore, if the null hypothesis in ANOVA F-test is rejected, the Duncan Multiple Range Test (DMRT) method is used to identify which model is closer to actual data at α level of significance. It also uses minimum absolute percentage error (MAPE) when the null hypothesis in ANOVA is accepted to select from GA or the other approaches. The significance of the proposed algorithm is two-fold. First, it is flexible and identifies the best model based on the results of ANOVA and MAPE, whereas previous studies consider the best-fit GA model based on MAPE or relative error results. Second, the proposed algorithm may identify time series as the best model for future electricity consumption forecasting because of its dynamic structure, whereas previous studies assume that GA always provides the best solutions and estimation. Fig. 1 depicts the proposed algorithm of this study. The reader should note all steps of the integrated algorithm are based on standard and scientific methodologies which are GA, simulated-based GA, time series,

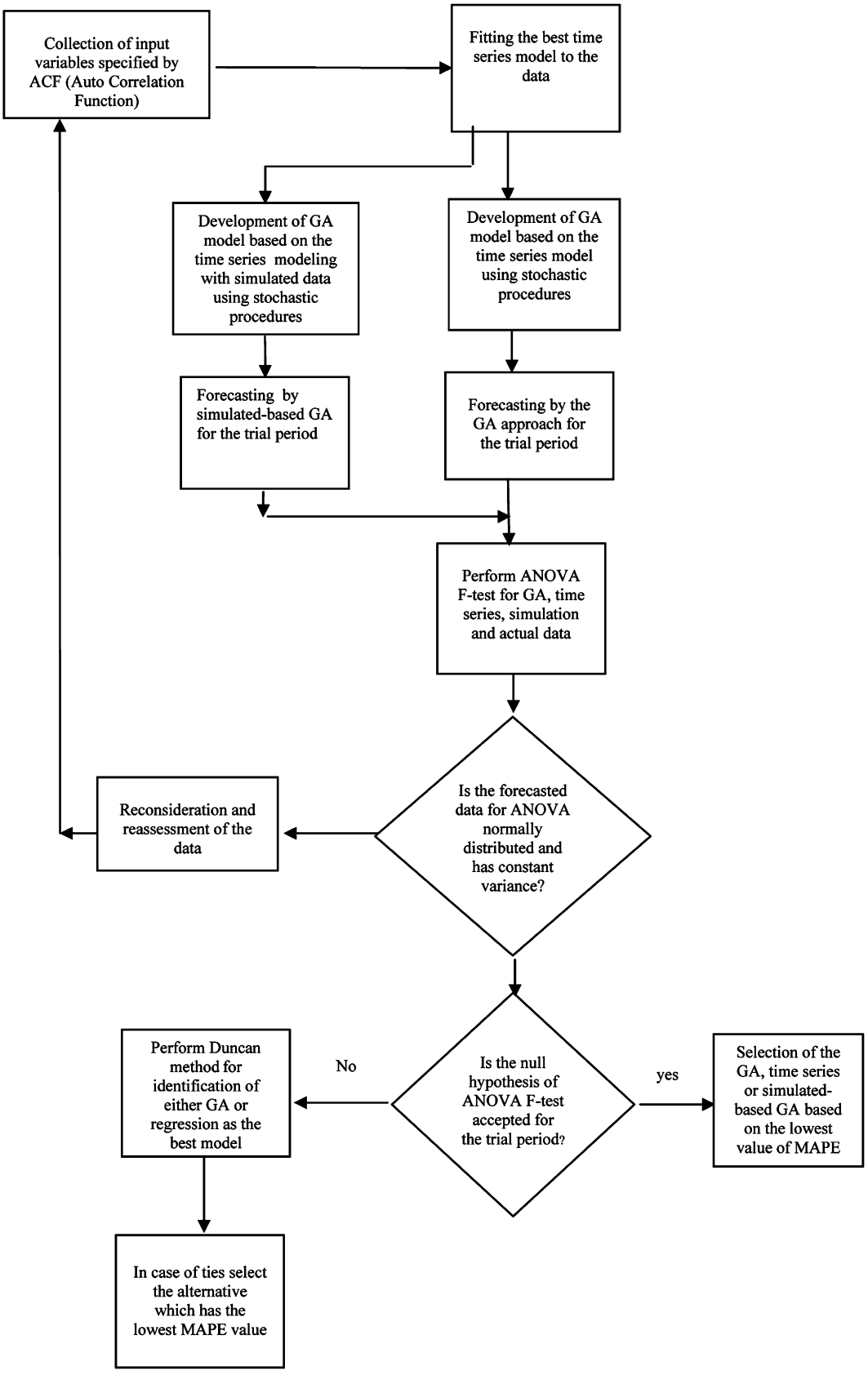

Fig. 1. The integrated GA, simulation, time series and DOE algorithm for electricity energy forecasting.

ANOVA, DMRT and MAPE. Furthermore, the GA and simulated-based GA modeling are based on which time-series model is selected for the data set by ACF. There are two types of experimental period in the algorithm: modeling period and test period. Modeling period concerns the data available for forecasting the three alternatives, namely, time series, GA and simulated-based GA while test period is used to compare the forecasted data by the three approaches via ANOVA, DMRT and MAPE. Next subsections discuss the component of the proposed algorithm.

2.1. Genetic algorithm

GA is similar to the natural evolution process where a population of a specific species adapts to the natural environment under consideration, a population of designs is created and then allowed to evolve in order to adapt to the design environment under consideration. These algorithms were directly described by Goldberg (1989) and have taken attention to solve optimizing problems. The most important advantage of the GA is their ability to use accumulative information about the initial unknown search space in order to move the next searches in to useful spaces (Ceylan and Ozturk, 2004).

The fundamental principal of GAs was first introduced by Holland (1975). In GAs, the better chromosome is the one that is closer to the optimal solution. In applied application of GA, population of chromosomes is created randomly. The number of the populations is different from one problem to another. Some guides about choosing the proper number of population are in different reports (Man et al., 1997).

GAs differ from conventional non-linear optimizing techniques as by preserving a population of the solutions. The key feature of such algorithms is the manipulation of a population whose individuals are characterized by possessing a chromosome. This latter can be coded as a string of characters of given length l. Each string represents a feasible solution to the optimization problem. A chromosome is composed of strings of symbols called bits.

The link between the GA and the problem at hand is provided by the fitness function (F). The F establishes a mapping from the chromosomes to some set of real numbers.

The GA procedure is generative. Each production of GA makes a new population of the existing type. Suppose that the population size is P initially. P individuals are assigned values to their chromosomes, where the assignment can be either random or deterministic. A permutation of such strings can be introduced to construct a population of designs which each design has its own fitness value. The main genetic operators are applied based on fitness function evaluation (Ceylan and Ozturk, 2004).

One of the genetic features is that instead of focusing on one point of the search space of a chromosome, it works on a population of chromosomes. This way at each stage the algorithm has a population of chromosomes that possess the desired properties more than the last one. Each population or generation of the chromosomes has the same size that is referred to as population size. If the number of the chromosomes is too low the possibility of the movement operation by GA will be also low and it only searches small part of the search space. According to the researches, a suitable population size is about 20–30 chromosomes. Of course sometimes a population with 50–100 has lead to best answers (Goldberg, 1989). The GA works with a 'population' of possible answers (e.g. sets of parameter values). Because of this, it does not require initial estimates of the fitting parameters, but requires only the allowable range of each parameter (Vandernoot and Abrahams, 1998).

The goodness or badness of the answer is determined by the value returned by the goal function. The more suitable answers have bigger fitness. GA encompasses three main operators: selection, crossover and mutation and which are described below briefly.

2.1.1. Selection

Selection operator searches according to the fitness of the members based on fitness values. The fitness value of the ith member in the population can participate in this operation on the basis of probabilities. This probability of the ith member in the population is calculated as below, in which n is the population size:

$$P_i = \frac{F_i}{\sum_{k=1}^{n} F_k}. \quad (1)$$

In the selection operation, the members of the population with better fitness can participate several times while the members with worse fitness may be deleted in order to obtain a bigger fitness average. Next, λ off springs are generated. The number of off springs competes with μ chromosomes, which are their parents for survival. After the competition is accomplished, the best μ chromosomes are chosen.

In another approach, which is called binary tournament, first two chromosomes among the initial ones and the off springs generated by the genetic operators are chosen and the best one is sent to the next generation.

2.1.2. Crossover

Crossover operation allows an exchange of the design characteristics between two mating parents. This operation is done by selecting two mating parents in which two random places are selected on each chromosome string and the strings between these two places among the mates are exchanged (Ceylan, 2006). A presentation of the crossover is shown below:

Parent 1 = 1 0 1 0 1 0 1 0 1 1,

Parent 2 = 1 0 0 1 0 0 0 1 1 1,

Child 1 = 1 0 1 0 0 0 0 1 1 1,

Child 2 = 1 0 0 1 1 0 1 0 1 1.

Another pair of crossover used in this paper, which is used randomly with the above type is the way the two random places are chosen in two chromosomes and the genes outside the crossed area are exchanged between two chromosomes diagonally. The crossover operation is applied with a probability of p_c which takes the probabilistic values from 0.2 to 0.8.

2.1.3. Mutation

Mutation operator is another essential operator in GA process and it acts on each chromosome after crossover operator in this way that a random number is produced for each bite of a chromosome, if this number is smaller than P_m mutation will occur in that bite and otherwise it is not happened. If mutation is not applied, after crossover the offspring will enter the new generation. According to the researches, P_m shows best while varying between 1% and 5%. Mutation operation prevents losing unexpected valuable genetic information in the population during selection and crossover operation. This operator acts at a random place of a chromosome with a low probability of P_m (Ozturk et al., 2004; Canyurt et al., 2004).

2.1.4. The fitness function

To introduce fitness function, it has to be said that the variables should be put in the model and then the difference of the estimated values from the actual data for each chromosome should be calculated and in each generation the individual with minimum difference must be returned. Individual parameters are selected randomly and after putting in the model the fitness is calculated. To obtain this purpose, the fitness function is to cover this goal and it is shown below which is called MAPE error, where D_{actual} and $D_{estimated}$ show the actual and estimated energy demand, respectively, and m is the number of observations. As the fitness function is minimization, individual with less amount of fitness are chosen for each generation:

$$\min f = \frac{1/m \sum_j^m (D_{actual} - D_{estimated})}{D_{actual}}. \quad (2)$$

2.1.5. Tuning GA

For the chosen GA operators; crossover and mutation, and also the number of generations and the size of the algorithm population tuning was done in which all these parameters were tested for different rates altogether and finally the best parameter values were obtained yielding minimum relative error.

2.2. Time-series modeling

Time-series models are quite well known to predict a variable behavior in the future by knowing its behavior in the past. One of the most famous time-series model is Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model belongs to a family of flexible linear time-series models that can be used to model many different types of seasonal as well as non-seasonal time series. In the most popular multiplicative form, the ARIMA model can be expressed as

$$\Phi_p(L) y_t = \theta_q(L) \varepsilon_t$$

with

$$\Phi_p(L) = 1 - \Phi_1 L - \ldots - \Phi_p L^p,$$
$$\Phi_q(L) = 1 - \theta_1 L - \ldots - \theta_q L^q, \quad (3)$$

where s is the seasonal length L is the back shift operator defined by $L^k y_t = y_{t-k}$ and ε_r is a sequence of white noises with zero mean and constant variance. Eq. (4) is often referred to as the *ARIMA* (p, q) model. Box et al. (1994) proposed a set of effective model building strategies to identification, estimation, diagnostic checking, and forecasting for ARIMA models. In the identification stage, the sample ACF (ACF shows the correlation of the two variables, if this amount is zero then there is no correlation between two variables) is plotted. A slowly decaying ACF suggests non-stationary behavior. In such circumstances, Box and Jenkins recommend differentiating the data. A common practice is to use a logarithmic transformation if the variance does not appear to be constant. After preprocessing, if needed, ACF and Partial Auto-Correlation Function (PACF) of preprocessed data are examined to determine all plausible ARIMA models.

One of the well-known models of the ARIMA family is AR (p) models, in which the value of y_t is dependent on the previous values of the consumption and a stochastic factor of ε_t. The rank of this model is dependent to the last amount of p. This model of p rank is described by

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_3 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t. \quad (4)$$

Some non-linear time-series patterns were also developed mainly by Granger and Pristly (Chatfield, 2003). One of these non-linear models is referred to as bilinear of which the first rank model of the bilinear model is shown by

$$X_t = aX_{t-1} + bZ_t + cZ_{t-1}X_{t-1}. \quad (5)$$

In which, Z_t is the stochastic procedure and a, b and c are the model parameters. It should be noted that only the last part of the above equation is non-linear. Another type of non-linear models is the Threshold Auto Regressive (TAR) model in which the parameters are dependent on the past values of the procedure. One example of such models is described by

$$X_t = \begin{cases} \alpha_1 X_{t-1} + Z_{t-1}^{(1)} & \text{if } X_{t-1} < d, \\ \alpha_2 X_{t-1} + Z_{t-1}^{(2)} & \text{if } X_{t-1} \geqslant d. \end{cases} \quad (6)$$

2.3. Simulated data

One of the unique features of this study is development of simulated data for monthly consumption based on

stochastic behavior of raw data for a 30-day period (1 month). Furthermore, raw data for each month is examined and the best probability distribution for that set of data with lowest *P*-value is identified. Therefore, the probability distributions of all months with respect to energy consumption are identified. The probability distributions are then generated by 1000 runs of simulation and their parameters are estimated. This is done to make sure the steady state is obtained. The estimated parameters (mostly mean) are therefore used as input data for the GA estimation. Therefore, the GA estimation considers both deterministic and stochastic input data. Considering the stochastic nature of the input data in GA modeling may be quite important when compared with actual data and time series.

2.4. Design of experiment

The estimated results of GA, simulated-based GA, time series and actual data are compared by ANOVA *F*-test. The experiment should be designed such that variability arising from extraneous sources can be systematically controlled. Time is the common source of variability in the experiment that can be systematically controlled through blocking (Montgomery, 2001). Therefore, a blocked design of ANOVA may be applied. The hypothesis is

$$\mathbf{H0}: \mu_1 = \mu_2 = \mu_3 = \mu_4,$$
$$\mathbf{H1}: \mu_i \neq \mu j \quad i,j = 1,2,3,4 \text{ and } i \neq j, \tag{7}$$

where μ_1, μ_2, μ_3 and μ_4 are the average estimation values obtained from actual data, GA, simulated-base GA and time series, respectively, tested at α significance level. If the null hypothesis is accepted then the preferred model is the one which has lower MAPE error. Otherwise, if the null hypothesis is rejected, DMRT is used to compare treatment means and to select the preferred model. Montgomery (2001) prescribes the use of either LSD or Duncan's Multiple Range experiments. There are $3(b-1)$ degrees of freedom for error at α level of significance where b is number of blocks or observation for the four treatments (actual, GA, simulated-base GA and time series) to perform DMRT method the following equation is used where MSE is the mean square error obtained from ANOVA:

$$S_{\bar{y}_{i.}} = \sqrt{\frac{\mathrm{MS}(error)}{b}}, \tag{8}$$

where b is the number of blocks and R_p as calculated as follows:

$$R_p = r_\alpha(p,f) S_{\bar{y}_{i.}}. \tag{9}$$

In which, $r_\alpha(p,f)$ is obtained from Duncan tables, α is the level of significance and f is the degree of freedom.

3. The case study

The proposed algorithm is applied to estimate and forecast monthly electricity consumption with the data from March 1994 to February 2005. Fig. 2 shows the electricity consumption during the aforementioned months.

3.1. Data preprocessing

As seen from Fig. 2, the data are not stationary and needs to become stationary. By definition, an ARIMA model is covariance stationary if it has a finite and time-invariant mean and covariance. This is one of the basic assumptions and also using preprocessed data is more useful in most heuristic methods (Box et al. (1994); Chatfield, 2003; Zhang and Oi, 2005), so the stationary assumption should be studied for the models. If the models are not covariance stationary, the most suitable preprocessed method should be considered and applied to the data set. In forecasting models, an appropriate preprocessing method should have two main features. It should make the process stationary and have the capability of transforming the preprocessed data in to its original scale (referred to as post processing). As was mentioned, the first step in the Box–Jenkins method is to transform the data to stationary from non-stationary format. The difference method was also proposed by Box–Jenkins. In this method, the following transformation should be applied and the data are stationary to the average:

$$y_t = x_t - x_{t-1}. \tag{10}$$

But in another method, which is more useful, the data are also variance stationary and therefore this method is applied as

$$y_t = \log(x_t) - \log(x_{t-1}). \tag{11}$$

Therefore, we also use this common method on the data to make them stationary. Fig. 3 shows the data being preprocessed by the Eq. (11).

In order to feed the data to the GA, first we have to know that consumption in month t depends on which

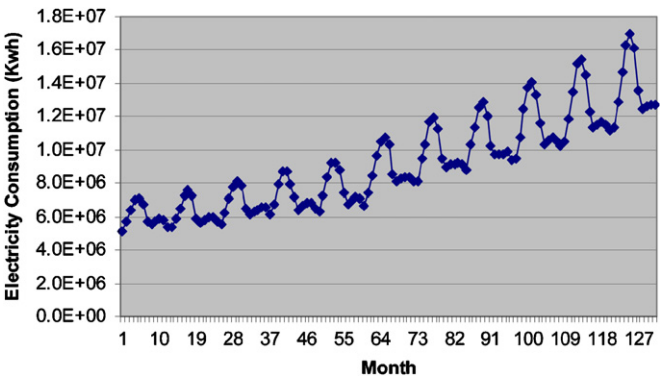

Fig. 2. Monthly electricity consumption in from March 1994 to February 2005 (Kwh).

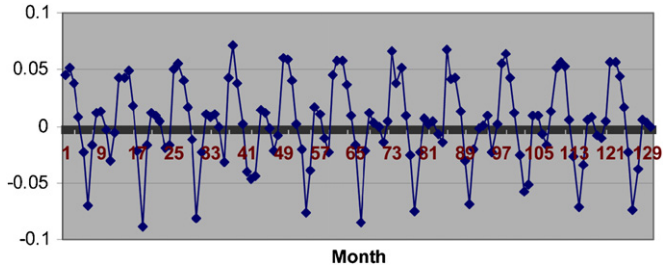

Fig. 3. The preprocessed data by the logarithmic difference method.

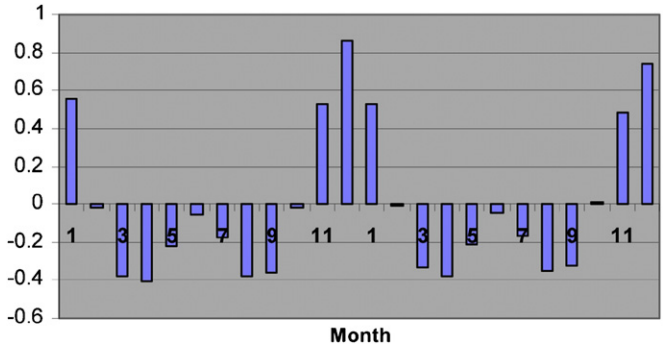

Fig. 4. ACF for the preprocessed data.

months. For this purpose, ACF is used. Fig. 4 shows this relation for y_t which is consumption in month t. It shows the auto-correlation for 2 years and this procedure is repeated periodically for the rest of the years.

According to Fig. 4, y_t is the function of consumption in the 3rd, 4th, 8th, 9th, 11th and 12th. Therefore, we have the series, which is ordered on this basis and according to this structure, we have 118 (after being preprocessed and ordered the data are 118 series) rows of data to be fed to the GA, simulated-based GA and time series. Due to linearity of the data the linear time series seems to be ideal. Therefore, the following linear time-series model is developed according to the above findings:

$$y_t = \alpha_1(y_{t-3}) + \alpha_2(y_{t-4}) + \alpha_3(y_{t-8}) + \alpha_4(y_{t-9}) + \alpha_5(y_{t-11}) + \alpha_6(y_{t-12}). \quad (12)$$

3.2. Genetic algorithm

For the GA development the 118 series of data are fed to Eq. (12). The reason for using linear equation for the estimation of monthly electricity consumption is that, we can compare it exactly with conventional time series and previous study for the estimation of electricity consumption in Iran has shown linear form of time series are more appropriate than non-linear ones (Sadeghi, 1999, 2003; Rahmati, 2004). Some studies (Darbellay and Slama, 2000) showed that the short-term prediction comply with linear behavior and using non-linear models like ANN is more appropriate for mid-term predictions. Another research (Hwarng and Ang, 2001) indicates that for complex and uncertain, non-linear methods are used. Moreover, the non-linearity of the data for the case study is being dominated by the data preprocessing approach and hence linear approach may be used. We have kept 12 series of data for test procedure and we deal with 106 series of data instead of 118 in our algorithm since the data was preprocessed and ordered based on the Eq. (12). According to this equation the chromosomes encompass six genes as we have six coefficients, all these coefficients are chosen randomly. The GA algorithm has the following parameters:

- population size (n): 100
- iterations (number of the generation): 200
- mutation rate: 0.04
- crossover rate: 94%.

The reason why this rate of crossover has been chosen is because of the lowest MAPE error as shown by Table 1. It shows the different crossover rates and their relative MAPE errors.

After applying GA with the above parameters, we get

$$y_t = 0.04(y_{t-3}) - 0.09(y_{t-4}) - 0.01(y_{t-8}) + 0.04(y_{t-9}) + 0.04(y_{t-11}) + 0.82(y_{t-12}). \quad (13)$$

For calculating the MAPE error of the above estimation for monthly electricity consumption, there is a need to post

Table 1
The fitness values for different crossover rates while other parameters are constant

Crossover rates (%)	50	60	70	80	84	90	94
MAPE values	0.051	0.05	0.045	0.032	0.024	0.019	0.014

Table 2
Estimated fitness values by GA and the real data for the testing period (Kwh)

The fitness values using GA	Actual data
11,275,831	11,202,808
11,586,663	11,328,070
12,543,789	12,920,315
14,455,562	14,715,280
16,224,999	16,291,332
16,430,043	16,944,113
15,887,329	16,086,172
13,908,969	13,581,415
12,625,007	12,435,256
12,491,345	12,612,147
12,745,199	12,690,363
12,671,315	12,676,607

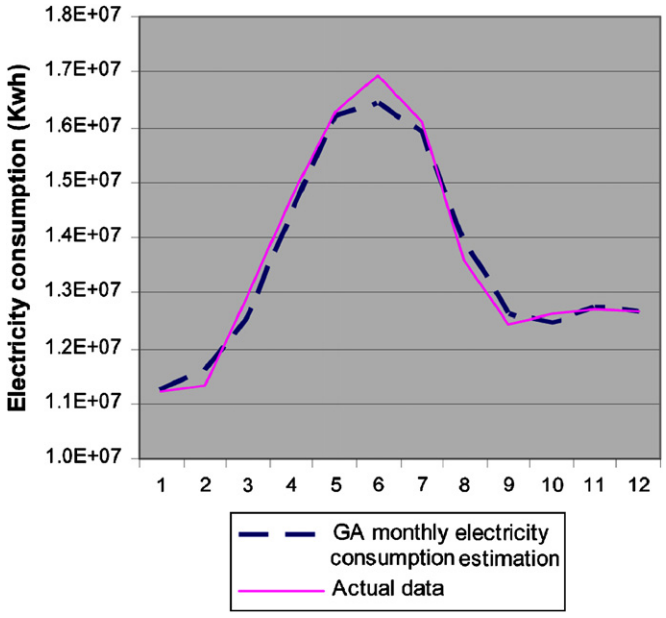

Fig. 5. GA versus actual data for the test data.

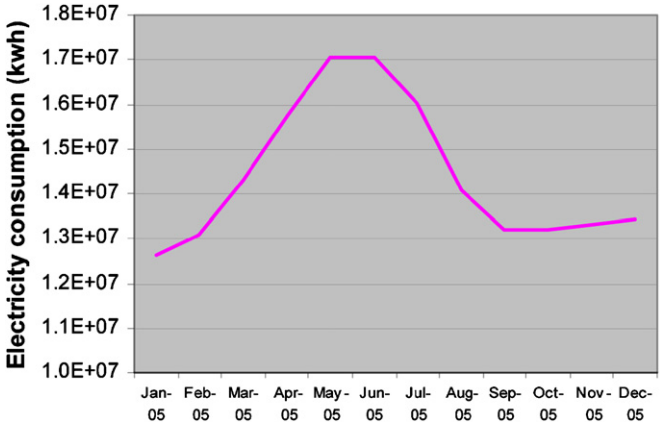

Fig. 6. Monthly electrical energy consumption prediction from February 2005 to January 2006.

process the fitness values as they have been once preprocessed. After doing so, the MAPE error returned by this estimation is 0.014. Table 2 shows the estimated values by GA and real data. Fig. 5 shows the fitness of GA versus actual data, which shows the good fitness for the trial period. To see the prediction ability of the GA model, it has been used to forecast the electricity consumption for the case of this study. Furthermore, GA can be used for short-term prediction and for mid or long term is not as exact as the short-term prediction. The GA prediction has been performed for 12 months. Moreover, because the data was available up to February 2005, the GA prediction is conducted for the next 12 months, which would be February 2005 to February 2006 (Fig. 6).

3.3. Time-series forecasting

In this section, we will estimate electricity consumption using time series approach by the same procedure which was mentioned in Section 3.2. Again, we have kept 12 series of data for testing procedure and we deal with 106 series of data instead of 118. The consumption in month t is again the function of the 3rd, 4th, 8th, 9th, 11th and 12th month. According to Eq. (12), we use Eviews software (Gujarati, 2003) to estimate the coefficients of this equation. The data are preprocessed like in Section 3.2. The result of this software is as follows:

$$y_t = 0.03(y_{t-3}) - 0.09(y_{t-4}) - 0.02(y_{t-8}) - 0.07(y_{t-9}) + 0.12(y_{t-11}) + 0.81(y_{t-12}). \quad (14)$$

The fitness values returned by Eq. (14), for the kept series of data must again post processed so that we can have the MAPE error for this estimation. Fig. 7 depicts the fitness values versus actual data and Table 3 shows the data of Fig. 7. The MAPE error for this estimation is 0.042.

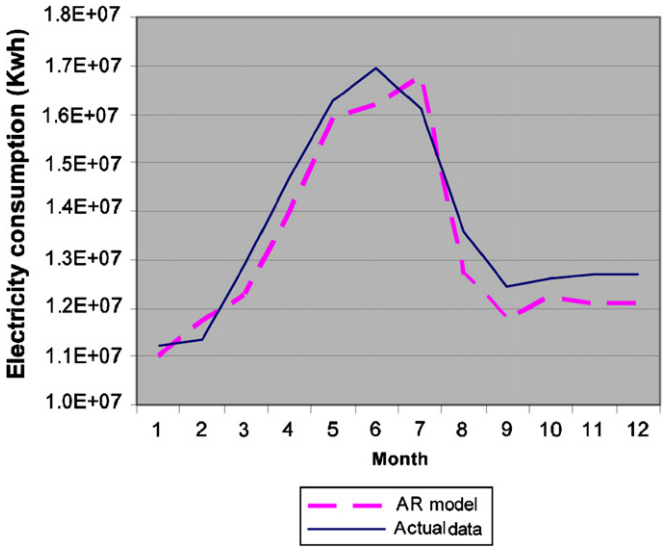

Fig. 7. Fitness values returned by time-series approach (AR modeling) versus actual for the testing period.

Table 3
The comparison of the estimated data using time series versus actual data

Fitness values using time series	Actual data
11,027,430	11,202,808
11,727,050	11,328,070
12,259,999	12,920,315
14,011,596	14,715,280
15,874,495	16,291,332
16,210,037	16,944113
16,761,144	16,086,172
12,674,796	13,581,415
11,755,537	12,435,256
12,245,900	12,612,147
12,110,746	12,690,363
12,104,043	12,676,607

3.4. Simulation

For this purpose, the distribution function of each month is calculated and then by using the related function for each month, the average value of that month is obtained. By this way, instead of using the deterministic value for each month we have its average value from the probable distribution the real value belongs to. When the distribution of each month is found, the amount of the square error and p-value of that distribution is also returned. An example of the selected distribution functions for each month is shown in Table 4. The selected distribution functions are selected from a series of distribution functions according to their p-values and square errors. Table 5 shows an example of such selection for the period of March 2001 to June 2001.

Using these results, the average values for each month is simulated by Visual Slam (Pristker et al., 1999). The selected distribution functions for each month is then generated 1000 times to obtain steady state. An example of the Visual Slam network can be seen in Fig. 8. The outputs of the simulation are the average and standard deviation of daily consumption values. Then, the upper and lower limits are constructed by $\mu \pm 3\sigma$. Next, the daily results are multiplied by the number of days in per month and consequently monthly consumption values are obtained.

Now with this new simulated data, electricity consumption must be again estimated. The data need to be preprocessed once again using Eq. (11). We have to know that consumption in month t (y_t) is dependent on previous months. For this reason, specifying the ACF of the data is necessary. Fig. 9 depicts that consumption in month t is reliable on the consumption of 3rd, 4th, 9th, 11th and 12th months before. It shows the auto-correlation for 2 years and this procedure is repeated periodically for the rest of the years.

Eq. (15) shows this relation. After knowing the form of the equation and therefore the number of genes in the chromosome (5 here), the data are fed to GA using the same GA parameters in Section 3.2. Eq. (16) further shows the output of GA for the coefficients of the equation:

$$y_t = \alpha_1(y_{t-3}) + \alpha_2(y_{t-4}) + \alpha_3(y_{t-9}) + \alpha_4(y_{t-11}) + \alpha_5(y_{t-12}), \tag{15}$$

$$y_t = -0.01(y_{t-3}) - 0.01(y_{t-4}) - 0.03(y_{t-9}) + 0.02(y_{t-11}) + 0.1(y_{t-12}). \tag{16}$$

In order to obtain the error for this approach in the estimation of electricity consumption estimation, the data are returned to their original status so that they can be compared to real data. Table 6 shows the values estimated by GA using Eq. (16) and real values. Fig. 10 also shows the fitness values of the estimated electricity consumption values by GA on simulated data versus actual ones. The MAPE error of this approach is 0.018.

Fig. 11 shows the three approaches against actual data. As shown in this figure, monthly electricity consumption by GA and simulated-based GA have very good fitness with actual data, whereas time series approach estimation has more distance from real data than the other two approaches. Furthermore, it is obvious that GA and simulated-based GA provide better estimation than time series. However, a formal approach, namely ANOVA is conducted next to make a formal comparison between the

Table 4
The distribution function for 2001

Distribution	Date
Normal (285,000, 20,900)	3/2001
TRIA (297,000, 332,000, 536,000)	4/2001
Normal (365,000, 18,300)	5/2001
Normal (403,000, 14,200)	6/2001
Normal (415,000, 12,300)	7/2001
Normal (387,000, 19,000)	8/2001
Normal (341,000, 17,200)	9/2001
Normal (325,000, 11,400)	10/2001
Normal (323,000, 14,700)	11/2001
Normal (323,000, 10,000)	12/2001
Normal (330,000, 12,000)	1/2002
Normal (285,000, 20,900)	2/2002

Table 5
Distribution functions for March 2001 to June 2001

Month	Distribution function	Square error	P-value
3/2001	Uniform	0.077	0.045
	Erlang	0.153	0.038
	Exponential	0.153	0.038
	Normal	0.100	>0.15
	Triangular	0.099	<0.01
	Gamma	0.160	<0.01
	Lognormal	0.249	<0.01
4/2001	Uniform	0.138	>0.15
	Erlang	0.225	<0.01
	Exponential	0.225	<0.01
	Normal	0.082	>0.14
	Triangular	0.074	>0.15
	Gamma	0.245	<0.01
	Lognormal	0.348	<0.01
5/2001	Uniform	0.046	<0.01
	Erlang	0.144	<0.01
	Exponential	0.144	<0.01
	Normal	0.025	>0.14
	Triangular	0.002	>0.15
	Gamma	0.150	<0.01
	Lognormal	0.257	<0.01
6/2001	Uniform	0.123	<0.01
	Erlang	0.225	<0.01
	Exponential	0.225	<0.01
	Normal	0.093	>0.15
	Triangular	0.061	0.124
	Gamma	0.233	<0.01
	Lognormal	0.308	<0.01

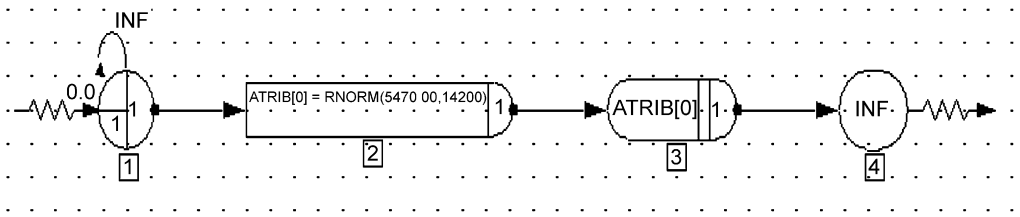

Fig. 8. The network to simulate electricity consumption average values.

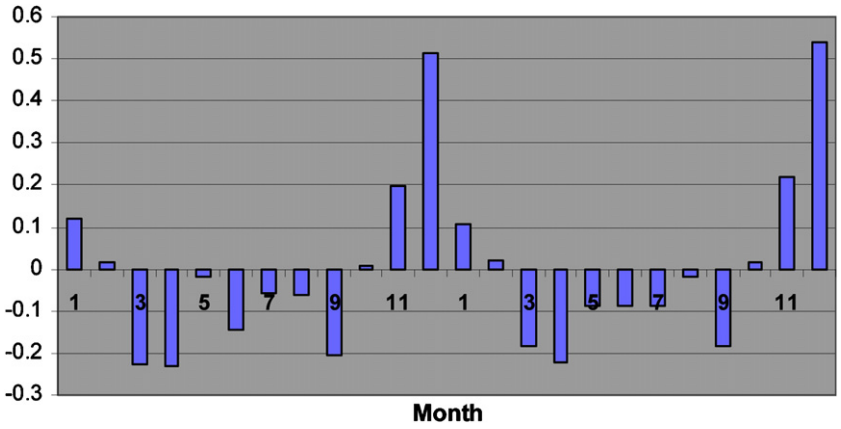

Fig. 9. The ACF of the simulated data for two test period.

Table 6
The comparison of the estimated values using simulated-based GA versus actual data

Fitness values of simulated-based GA	Actual data
11,595,161	11,187,128
11,128,301	11,192,604
12,492,455	12,709,220
13,896,839	14,283,109
15,834,338	16,271,932
16,032,426	16,178,585
15,673,535	16,065,205
13,331,456	13,567,828
12,320,346	12,197,568
12,801,220	12,574,024
12,619,154	12,439,073
12,539,857	12,681,651

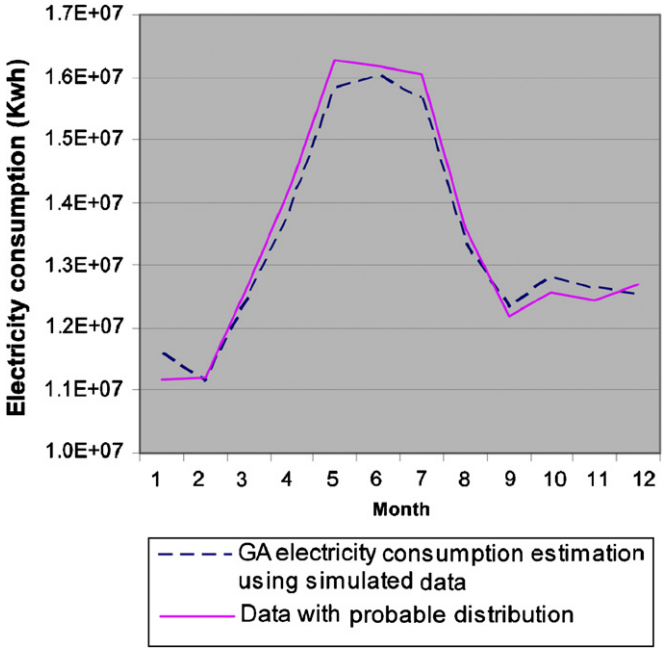

Fig. 10. The simulated-based GA versus actual data for the test period.

GA, simulated-based GA, time series and actual data. This is of course according to layout of the proposed algorithm presented in previous section.

4. Analysis of variance

The estimated results of GA, simulated-based GA, time-series method and actual data are compared by ANOVA *F*-test. The experiment was designed such that variability arising from time can be systematically controlled through blocking. Therefore, a blocked design of ANOVA is applied according to Eq. (4) and the results are shown in Table 7.

According to ANOVA results in Table 7, with $\alpha = 0.05$ the null hypothesis is rejected and therefore, further analysis needs to be performed to foresee which treatment pairs caused the rejection of null hypothesis. Furthermore, DMRT is used to compare treatment means. There are 33

degrees of freedom for error and at $\alpha = 0.05$ and Eq. (5) is used and R_p is calculated as

$S_{\bar{y}i.} = 0.0842,$

$r_{0.05}(2, 33) = 2.88,$

$r_{0.05}(3, 33) = 3.03,$

$r_{0.05}(4, 33) = 3.09,$

$R_2 = r_{0.05}(2, 33)S_{\bar{y}i.} = 0.242,$

$R_3 = r_{0.05}(3, 33)S_{\bar{y}i.} = 0.255,$

$R_4 = r_{0.05}(4, 33)S_{\bar{y}_{i.}} = 0.260.$

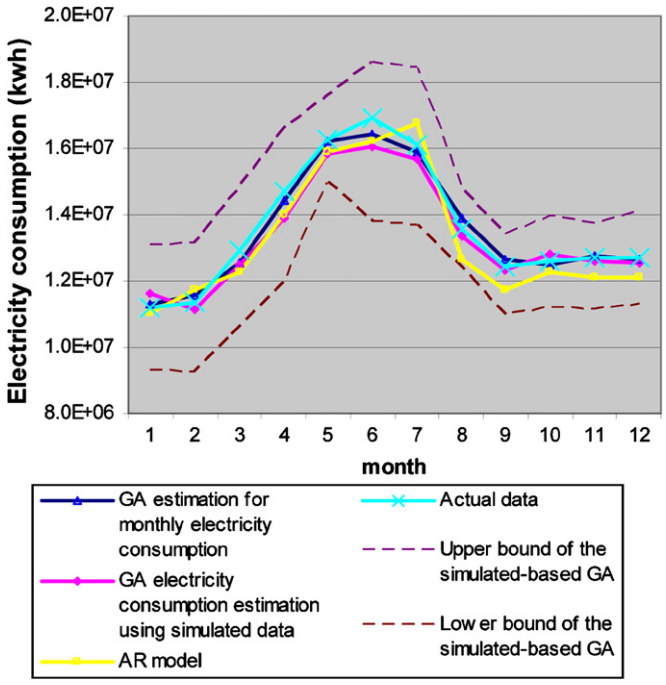

Fig. 11. The comparison of all the approaches versus actual data for electricity consumption estimation.

Also the treatment averages are as follows:

$\bar{y}_{1.} = 13.5774 \quad \bar{y}_{2.} = 13.5705 \quad \bar{y}_{3.} = 13.35 \quad \bar{y}_{4.} = 13.23.$

In the above, $\bar{y}_{1.}$ is the average of the actual data, $\bar{y}_{2.}$ is the average values estimated by GA, $\bar{y}_{3.}$ is the average of simulated-based GA value and $\bar{y}_{4.}$ is the average values estimated by time-series model.

Comparing treatment 1 with $2 = 13.577 - 13.570$
$= 0.007 < 0.242 \rightarrow \mu_1 = \mu_2,$
Comparing treatment 1 with $3 = 13.577 - 13.350$
$= 0.227 < 0.255 \rightarrow \mu_1 = \mu_3,$
Comparing treatment 1 with $4 = 13.577 - 13.230$
$= 0.347 > 0.260 \rightarrow \mu_1 \neq \mu_4.$

According to the above, it is found that the cause of the inequality of the averages is the fourth treatment which is time-series model. Furthermore, GA and simulated-based GA are statistically equal with the actual data. However, GA provides lower MAPE error than that of the simulated-based. Therefore, GA is selected as the best approach for forecasting electricity consumption according to the proposed algorithm.

Other prediction techniques such as Artificial Neural Networks (ANNs) are also used for forecasting purpose (Azadeh et al., 2007a) in which the same data set were used and the result shows that MLP network with back propagation learning method and two hidden layer is the preferred network for Iran monthly electricity consumption modeling. The MAPE value of such network is reported 0.156 while the result of the present study will show that GA dominates ANN for this data set (Table 8).

Table 7
ANOVA table for comparison of genetic algorithm, simulated-based GA, time series and actual data

Source of variation	Sum square	Degrees of freedom	Mean square	F	F ($\alpha = 0.05$)	P- value
Between groups (treatment)	1.042	3	0.34	4.08	2.89	0.014
Blocks (months)	143.65	11	13.05	153.40		
Within groups	2.80	33	0.085			
Total	147.5	47				

Table 8
Comparing four ANN MLP models with GA and simulated-based GA

MLP Model number	1	2	3	4	5	6
Learning method	BP, momentum, weight decay	BP, momentum	BP, momentum, weight decay	BP, momentum	GA	Simulated-based GA
Number of neurons in first hidden layer	20	16	16	16		
Number of neurons in second hidden layer	6	6	6	0		
MAPE	0.0165	0.016	0.0156	0.022	0.014	0.018

5. Conclusion

This paper presented an integrated algorithm based on GA, simulated-based GA, time series and DOE (ANOVA and DMLT) to forecast electricity energy consumption. To show the applicability and superiority of the proposed framework actual data for energy consumption in Iran from 1994 to 2005 (131 months) was used. The proposed algorithm may be used to estimate energy demand in the future by optimizing parameter values. The GA applied in this study has been tuned for all its parameters and the best coefficients with minimum error are identified, while all parameter values are tested concurrently. The proposed algorithm first identifies the best time-series model with lowest possible relative error. It then uses the type of model selected by time series for GA and simulated-based GA modeling and estimation. It uses ANOVA *F*-test to select either GA, simulated-based GA or conventional time series for future demand estimation based on the test data. Furthermore, if the null hypothesis in ANOVA *F*-test is rejected, the DMRT is used to identify which model is closer to actual data at α level of significance. It also uses MAPE when the null hypothesis in ANOVA is accepted to select from GA, simulated-based GA or time-series model. One major assumption for using the algorithm is that the data should be normally distributed with constant variance. The significance of the proposed algorithm is three fold. First, it is flexible and identifies the best model based on the results of ANOVA and MAPE, whereas previous studies consider the best-fit GA model based on MAPE or relative error results. Second, the proposed algorithm may identify conventional time series as the best model for future electricity consumption forecasting because of its dynamic structure, whereas previous studies assume that GA always provide the best solutions and estimation. Third, it generates and uses probabilistic distribution for monthly consumption as a new alternative (simulated-based GA). This is done to make sure the stochastic nature of data is also considered and modeled as an alternative. To show the applicability and superiority of the proposed algorithm the data for monthly electricity consumption March 1994 to February 2005 (131 months) was used and applied to the proposed algorithm. The results showed that GA is superior to conventional time series and simulated-based GA according to ANOVA *F*-test, MAPE and DMRT experiment. However, it is argued that the utilization of GA, simulated-based GA and conventional time series assures the best-fit for an actual data set. This in turn minimizes the bias of utilizing only one approach like previous studies and enhances the overall performance of future forecasting and estimation.

Acknowledgment

This research is partially supported by a grant from the Research Institute of Energy Management and Planning, Faculty of Engineering at the University of Tehran and Ministry of Energy, Iran.

References

Azadeh, A., Ghaderi, S.F., Tarverdian, S., 2006. Electrical energy consumption estimation by genetic algorithm. In: Proceedings of IEEE Conference on Industrial Electronics, 9–13 July, Montreal, Canada.

Azadeh, A., Ghaderi, S.F., Sohrabkhani, S., 2007a. Forecasting electrical consumption by integration of Neural Network, time series and ANOVA. Applied Mathematics & Computations 186, 1753–1761.

Azadeh, A., Ghaderi, S.F., Tarverdian, S., Saberi, M., 2007b. Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. Applied Mathematics & Computations 186, 1731–1741.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. Prentice-Hall, Englewood Cliffs, NJ.

Bunning, D., Sun, M., 2005. Genetic algorithm for constrained global optimization in continuous variables. Applied Mathematics & Computation 171, 604–636.

Canyurt, O.E., Ozturk, H., Hepbasli, A., 2004. Energy demand estimation based on two-different genetic algorithm approaches. Energy Sources 26 (14), 1313–1320.

Ceylan, H., 2006. Types of Crossover. Personal communications.

Ceylan, H., Ozturk, H., 2004. Estimating energy demand of Turkey based on economic indicators using genetic algorithm approach. Energy Conversion and Management 45, 2525–2537.

Chatfield, C., 2003. The Analysis of Time Series: An Introduction.

Compton, P., Wu, Y., 2005. Energy consumption in China: past trends and future directions. Energy Economics 27 (1), 195–208.

Darbellay, A., Slama, M., 2000. Forecasting the short-term demand for electricity, do neural networks stand a better chance? International Journal of Forecasting 16, 71–83.

Goldberg, D.E., 1989. Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley, Harlow, England.

Gujarati, D.N., 2003. Basic Economics, fourth ed. McGraw-Hill higher education, New York.

Haldenbilen, S., Ceylan, H., 2004. Genetic algorithm approach to estimate transport energy demand in Turkey. Energy Policy 33, 89–98.

Hasheminia, H., Akhavan Niaki, S.T., 2006. A genetic algorithm approach to fit the best regression/econometric model among the candidates. Applied Mathematics & Computation 183, 337–349.

Holland, J.H., 1975. Adoption in Neural and Artificial Systems. The University of Michigan Press, Ann Arbor, MI, USA.

Hwarng, H.B., Ang, H.T., 2001. A single neural network for ARMA(p,q) time series. Omega 29, 319–323.

Man, K.F., Tang, K.S., Kwong, S., Halang, W.A., 1997. Genetic Algorithms for Control and Signal Processing. Springer, London, pp. 1–5.

Montgomery, D.C., 2001. Design & Analyze of Experiments. Wiley, New York.

Muni, D., Pal, N., Das, J., 2006. Genetic programming for simultaneous feature selection and classifier design. IEEE Transactions on Systems, Man and Cybernetics 36 (1).

Osman, M.S., Abo-sinna, M.A., Mousa, A.A., 2005. A combined genetic algorithm-fuzzy logic controller (GA-FLC) in non-linear programming. Applied Mathematics & Computation, 821–840.

Ozturk, H., Canyurt, H., Hepbasli, A., Utlu, Z., 2004. Three different genetic algorithm approaches to the estimation of residential energy input/output values. Building and Environment 39, 807–816.

Ozturk, H., Ceylan, H., Canyurt, O.E., Hepbasli, A., 2005. Electricity estimation using genetic algorithm approach: a case study of Turkey. Energy 30, 1003–1012.

Pristker, A.B., O'Reilly, J.J., Laval, D.K., 1999. Simulation with Visual Slam and AweSim. Wiley, New York.

Rahmati, R., 2004. A comparative study of neural network and Box–Jenkins ARIMA modeling in monthly electricity demand in Iran. MS Thesis, Faculty of Economics, University of Tehran, Iran.

Sadeghi, M., 1999. Demand stability for energy in Iran. Ph.D. Dissertation, Faculty of Economics, University of Tehran, Iran.

Sadeghi, N., 2003. Forecasting and modeling electricity demand by an econometric model. MS Thesis, Faculty of Economics, University of Tehran, Iran.

Stach, W., Kurgan, L., Pedricz, W., Reformat, M., 2005. Genetic learning of fuzzy cognitive maps. Fuzzy Sets and Systems 153, 371–401.

Tang, A., Quek, C., Ng, G., 2005. GA-TSKfnn: Parameters tuning of fuzzy neural network using genetic algorithms. Expert Systems with Applications 29, 769–781.

VanderNoot, T.J., Abrahams, I., 1998. The use of genetic algorithms in the non-linear regression of immittance data. Journal of Electro Analytical Chemistry 448, 17–23.

Zhang, G.P., Oi, M., 2005. Neural networks forecasting for seasonal & trend time series. Journal of Operational Research 160, 501–514.