The Third International Conference "Problems of Cybernetics and Informatics"
September 6-8, 2010, Baku, Azerbaijan. Section #3 "Modeling and Identification"
www.pci2010.science.az/3/29.pdf

INTERACTIVE SYSTEM FOR CONSTRUCTING AND ANALYZING LINEAR AND NONLINEAR STATISTICAL MODELS

Kamil Aida-zade[1], Marina Khoroshko[2]

[1]Azerbaijan State Oil Academy, Baku, Azerbaijan
[2]Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*kamil_aydazade@rambler.ru*, [2]*khoroshko@yahoo.com*

The problems of constructing mathematical models that adequately describe the process under investigation have become an integral part of developing any information system in many areas of science, industry and business. Far not always the researchers can know the kind of mathematical model that best describes the investigated process (object). In particular, this problem may arise in the process of modeling new, poorly studied phenomena. For example, when modeling economic processes of the transition to a market economy, researchers must determine which model accurately approximates one or another economic indicator.

Choosing suitable kind of a mathematical model is important not only to study and control the processes, but also to reveal general tendency of their development. The problem of choosing a class and a kind of mathematical model of investigated process concerns to the first stage of modeling, i.e. to structural identification. At the moment there are no mathematically rigorous methods and algorithms to obtain the solution of this problem. The essential role at this stage is played by the researcher experiences, and by degree of scrutiny of the object. It allows researchers practically at once to proceed to the second stage of modeling, that is, to parametrical identification.

Because of a small degree of scrutiny of the object researchers need to test different kinds of mathematical models for adequacy in describing the process. Therefore working out an interactive system, which allows to construct statistical models on the basis of available statistics in an interactive mode, and also to carry out the analysis of adequacy of the constructed models using statistical criteria, is rather actually for modeling. Method of statistical modeling is widely applied on condition that the drift of mathematical models is small.

The developed interactive system for constructing and analysing linear and non-linear statistical models is presented in the paper. The interactive system consists of the following units:

- Inputting, updating and analyzing statistical data reflecting observing over processes;
- Adding new kinds of models to database of models;
- Parametrical identifying models for which the hypothesis is put forward that one or several functions from the selected classes of functions can adequately describe the tendency of development of analyzed process;
- Choosing the model which most adequately describes the regularity of process;
- Calculating the predicted values for the investigated phenomenon on the basis of the chosen model;
- Displaying the results computed including the use of graphic tools.

The user of interactive system has an opportunity to calculate parameters for all kinds of mathematical models (which are available in base of models) on the basis of the statistical information on process or phenomenon, and to choose from them the best one. Also the users can specify new kind of a model, if there is no suitable model in database of models. During one interactive system session, the user has a possibility to select different kinds of models repeatedly and to recalculate model parameters in user-friendly mode. The system provides convenient interface for inputting, viewing and updating statistical data. For presentation of

results there is a possibility to display them in a graphic mode. Also there exists a possibility to add new kinds of models to database of models.

To identify the parameters $p \in R^l$ of mathematical models of investigated process $y = f(x, p), x \in R^n$ two criteria can be used: root-mean-square criterion, and Chebyshev's min-max criterion. The criteria correspond to two different approaches to parametrical identification. We use method of least squares for first criterion. The objective function is the following:

$$p = \arg\min_{p} \left\{ \frac{1}{m} \sum_{j=1}^{m} \left[f(x_j, p) - y_j^{ob} \right]^2 \right\}, \quad (1)$$

where $m = [3/4 * k]$, k is a total number of observed values, p is a vector of parameters of a model, $x_j \in R^n$, y_j^{ob}, $j = 1,...,k$ are the statistical data reflecting observation over corresponding input values and output values of parameters of process, $[a]$ is an integer part of number a.

By default, parametrical identification of a model is carried out using a subset of observation, representing first 3/4 of all observed values (this parameter (3/4) can be changed in interactive mode), other $[1/4 * k]$ of observed values are used later on for estimation of the quality of the constructed model forecast.

If the user selects the second criterion – Chebyshev's min-max criterion, the model parameters are calculated by minimization non-smooth function:

$$p = \arg\min_{p} \max_{1 \le j \le m} \left| f(x_j, p) - y_j^{ob} \right|. \quad (2)$$

As a formal criterion for choosing the best mathematical model from any set of models we use the least value of root-mean-square deviation of actual values of the functions from the calculated values.

For example, if root-mean-square criterion of parametric identification is selected, then the model with minimal value of

$$R_1 = \sum_{j=1}^{m} \left[f(x_j, p) - y_j^{ob} \right]^2 \quad (3)$$

is chosen to be the most adequate model.

The quality of model in case of Chebyshev's min-max criterion is defined by value

$$R_2 = \max_{m+1 \le j \le k} \left| f(x_j, p) - y_j^{ob} \right| . \quad (4)$$

For modeling one-dimensional processes $(n = 1)$, in particular, for time series the following functions are included in the worked out system:

1) Linear function $y = p_1 + p_2 t$, $l = 2;$

2) Logarithmic-linear function $y = p_1 + p_2 \ln t$, $l = 2;$

3) Exponential function $y = p_1 + e^{p_2 t}$, $l = 2;$

4) First Tornquist function $y = \dfrac{p_1 t}{p_2 t + p_3 c}$, $l = 3;$

5) Power function $y = p_1 t^{p_2}$, $l = 2;$

6) Root $y = p_1 + p_2 \sqrt{t}$, $l = 2;$

7) Hyperbola $y = p_1 + \dfrac{p_2}{t}$, $l = 2;$

8) Linear-hyperbolical combined function $y = p_1 + p_2 t + \frac{p_3}{t}$, $l = 3;$

9) Kinetic function $y = p_1 e^{p_2 t} t^{p_3}$, $l = 3;$

10) Logarithmic parabola $y = p_1 + p_2 \ln t + p_3 \ln^2 t$, $l = 3;$

11) Hyperbola II $y = p_1 + \frac{p_2}{t} + \frac{p_3}{t^2}$, $l = 3;$

12) Parabolic function $y = p_1 + p_2 t + p_3 t^2$, $l = 3;$

13) Logarithmic-hyperbolic combined function $y = p_1 + p_2 \ln t + \frac{p_3}{t}$, $l = 3;$

14) Exponential-power function $y = p_1 e^{p_2 t} (\ln t)^{p_3}$, $l = 3;$

15) Root II $y = p_1 + p_2 \sqrt{t} + p_3 \ln t$, $l = 3;$

16) Root III $y = p_1 + p_2 \sqrt{t} + \frac{p_3}{\sqrt{t}} + p_4 \ln t$, $l = 4;$

17) Logarithmic hyperbola $y = p_1 + p_2 \frac{\ln t}{t}$, $l = 2;$

18) Logarithmic hyperbola $y = p_1 + p_2 \frac{\ln t}{t} + p_3 t$, $l = 3;$

19) $y = p_1 + p_2 \sin t$, $l = 2;$

20) $y = p_1 + p_2 \sin t \cos t$, $l = 2;$

21) $y = p_1 + p_2 \frac{\sin t \cos t}{t} + p_3 t$, $l = 3;$

22) $y = p_1 + p_2 \frac{\sin t \cos t}{t}$, $l = 2;$

23) $y = p_1 + p_2 \sin t + p_3 \cos t$, $l = 3;$

24) $y = p_1 + p_2 \sin t \ln t$, $l = 2;$

25) $y = p_1 + p_2 \left(\frac{\ln t}{t} \right)^2 + p_3 \frac{\ln t}{t}$, $l = 3.$

After parametrical identification the developed system allows calculate the following characteristics: average arithmetic - $\bar{y} = \frac{1}{m} \sum_{j=1}^{m} y_j^{ob}$; variance - $S_{var}^2 = \sum_{j=1}^{m} (y_j^{ob} - \bar{y})^2 / (m-1)$; residual variance - $S_{rv}^2 = \sum_{j=1}^{m} (y_j^{ob} - y_j^c) / (m-n-1)$; RMS deviation - $S_{rv} = \sqrt{S_{vr}^2}$; correlation coefficient - $R = \sqrt{1 - S_{rv}^1 / S_{var}^2}$; coefficient of determination - $\mathrm{D} = \mathrm{R}^2$; Fisher's test - $F = \frac{S_{var}^2}{S_{vr}}$; Durbin-Watson test - $d = \sum_{j=1}^{m} (l_j - l_{j-1})^2 / \sum_{j=1}^{m} l_i^2$, $l_j = y_j^{ob} - y_j^c$, $j = 1,..,m$; average relative error - $\varepsilon = \frac{1}{m} \sum_{j=1}^{m} (y_j^{ob} - y_j^c) \cdot 100\,\%$, where l is a number of estimated parameters, k is the length of a series, y_j^{ob}, $j = 1,...,k$ is actual values of output parameters, y_j^c, $j = 1,...,k$ is calculated values of output parameters.

The interactive system was developed in VISUAL BASIC for WINDOWS; database was developed in DBMS ACCESS. Figure 1 displays a fragment of the interactive system.

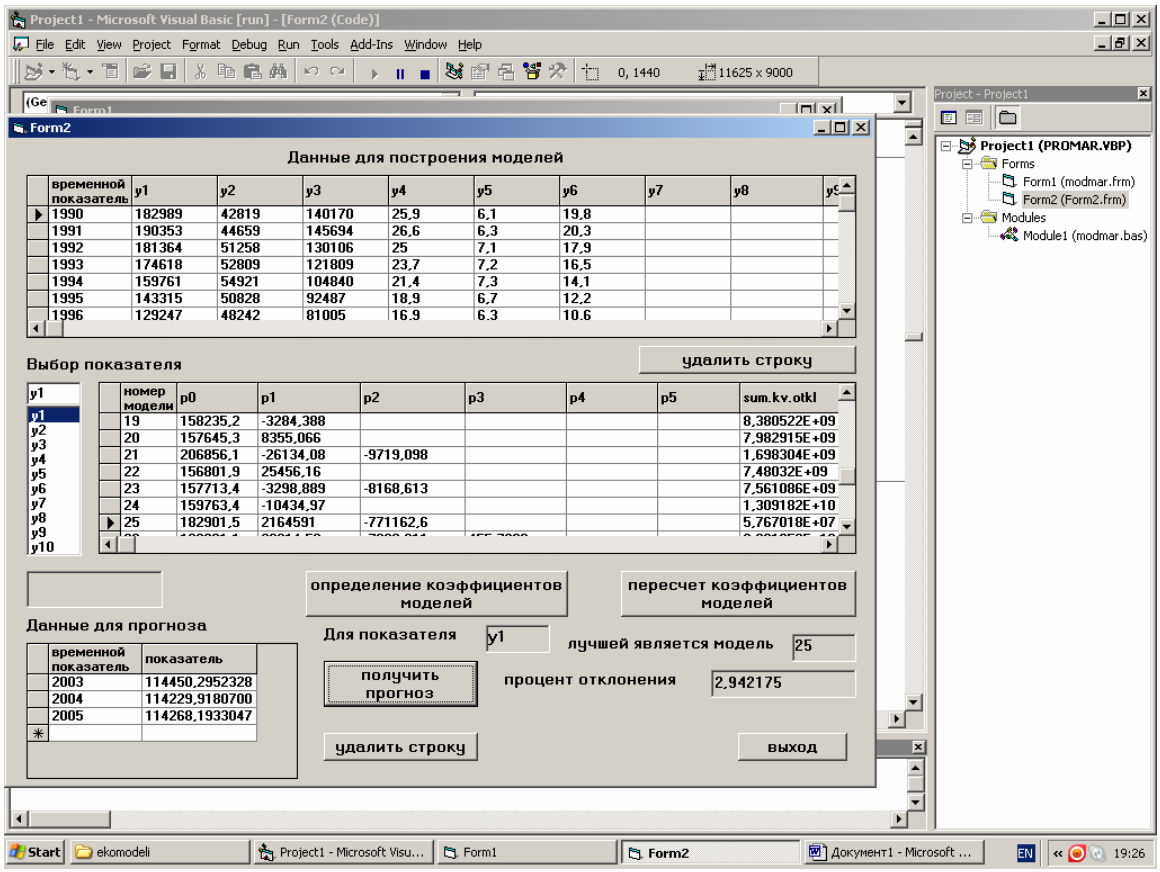

Fig.1

The interactive system was tested on constructing mathematical models for time series describing social and economic phenomena in Azerbaijan Republic (birth rate, death rate etc.). The testing was conducted with the aim of studying and predicting these phenomena.

References

1. K.R.Aida-zade, M.N.Khoroshko. An approach to mathematical modeling and control technological processes. J. «Electronic modeling», Kiev, Vol.30, № 5, 2008, pp. 37-48 (in Russian).
2. S.Z.Guliyev, M.N.Khoroshko. On combined stages of parametrical identification and optimization for dinamical processes. Transactions of Azerbaijan National Academy of sciences, series of physical-technical and mathematical sciences, Baku, Azerbaijan, 2009, Vol. XXIX, № 3, pp. 10-16 (in Russian).
3. Rzaev T.N. Identification of System. Elm, Baku, 2007, 518 p. (in Azeri).