A METHOD TO IDENTIFY SYSTEMS BASED ON RANDOM BINARY EVENTS

Petre JUNIE[1], Mihai TERTIŞCO[2], Cristian EREMIA[3], Gabriel ENE[4]

Prezenta lucrare prezintă rezultatele cercetării ştiinţifice personale privind modelarea şi identificarea experimentală a sistemelor cu evenimente binare aleatoare (EBA). Structura modelelor acestor sisteme este de tipul regresiilor logistice. Pentru identificare se propune utilizarea criteriului verosimilităţii maxime [4]. În vederea estimării parametrilor modelului a fost concepută şi testată o metodă de tip Monte – Carlo. Caracteristicile statistice ale estimaţiilor parametrilor modelului sunt precizate.

The present paper presents results of personal scientific research on modeling and experimental identification of systems with Random Binary Events .Models of these systems is the type of logistic regression. For identification we propose to use maximum likelihood criterion [4]. In order to estimate the model parameters was designed and tested a method of Monte – Carlo. Statistical analysis is of model parameters estimates are given.

Keywords: Modeling, identification, Logistic Regression Model, Monte Carlo method, maximum log likelihood criterion

1. Introduction

Classical methods of systems identification primarily refer to processes whose dynamic behavior is described by either differential equations or difference equations. The most elaborate methods of identification aim at constant parameter linear systems that satisfy the requirements imposed by the applications specific to the domain of industrial processes automation described by such models [1] For systems with discrete events characterized by discrete streams of operations and discrete activities accompanied by phenomena of blocking, non-synchronization and conflicts new modeling formalisms have been developed [2] Classic models covered by conventional identification methods describe the

[1] Eng., Department of Automatics and Systems Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: enegabrielcristian@yahoo.com
[2] Prof., Department of Automatics and Systems Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: tertisco_mihai@yahoo.com
[3] Eng., Department of Automatics and Systems Engineering, University POLITEHNICA of Bucharest, Romania
[4] Eng., Department of Automatics and Systems Engineering, University POLITEHNICA of Bucharest, Romania, email:junpetre2000@yahoo.com

dynamic behavior of a single object from a collection of similar objects in which processes that are subject to physical and chemical laws occur. In the present paper we are concerned with models of systems with binary independent random events. Unlike traditional models, this particular type of models describes a homogeneous lot M of cardinality N consisting of two distinct entities. These entities can be separated into two classes. Each entity in this population is characterized by a dependent variable Y (output) and one or more independent variables (input) x. Variable Y can take only logical values: 1 or 0, yes or no, sick or healthy, etc. The independent variable can take logic values or can take values in the set of real numbers. In most applications encountered in the literature of expertise these independent variables take logic values, 1 or 0. Based on experimental testing of each entity (from the N, of set M) entities can be divided into two classes: entities class with Y = 1 and entities class with Y = 0. The model in which we have only one independent variable x is called the logistic model SISO (single - input - single - output). In the case of several independent variables the model is called MISO model (MULTY - input - single - output). For simplicity we refer in particular to the type SISO logistic models.

Table 1

Data on the analysis of a probable cause of the Challenger shuttle disaster

x	Y	X	Y	x	Y
Temperature	Defect	Temperature	Defect	Temperature	Defect
66	0	57	1	70	0
70	1	63	1	81	0
69	0	70	1	76	0
68	0	78	0	79	0
67	0	67	0	75	1
72	0	53	1	76	0
73	0	67	0	58	1
70	0	75	0		

In the case of the identification theory, the model that expresses the probabilistic interdependence between the dependent variable Y, binary type, and one or more independent variables x, is called *logistic regression*. For example, the experimental data are disclosed in Table 1 regarding the analysis of probable causes of the Challenger space shuttle disaster (1986) which shows the various temperatures at which the damage of a specific mechanical bond along the N = 23 tests occurred or not. In this case, the set of entities consists of the N trials of the

shuttle that are divided into two classes: the trials class which results in the occurrence of the defect Y = 1 and the trials class in which the defect does not occur Y = 0. Given the results of these N trials of the shuttle, we could build a logistic model that would help us answer the question: "what is the probability that the defects should occur (Y = 1) at a given temperature x?" In this example the class of events Y = 1 contains the following six categories of random events caused by independent variable values x:

Y=1, x=75); (Y=1, x=70); (Y=1, x=63); (Y=1, x=58); (Y=1, x=58); (Y=1, x=57)

2. Logistic Regression

The regression equation obtained in this case is of a type different from other known regressions, such as continuous, single dimensional, multidimensional, linear and nonlinear etc. Tree variants of the logistic model structure for a SISO (single input single output), found in the literature of expertise [3]. In the first variant the continuous size "p" is a nonlinear function of x and of two unknown parameters: β_0 and β_1. If the event Y = 1, then this event's occurrence takes place with the probability ***P(Y=1|x) = p***. This type of regression provides information about the importance of variables x in the differentiation of classes, and about the classification of one observation into one of the classes. Unlike classical linear regression, in the case of logistic regression, instead of dependent variable Y, which may take the binary value Y = 1 → *"success" or* Y = 0 → *"failure"*, it is used a continuous variable ***p***, which takes values ranging from 0 to 1. A value of ***p*** is interpreted as the probability of obtaining a "success" (Y = 1), subject to the independent variable value *x*. Then the opposite event Y = 0 has a probability of occurrence P (Y = 0) = 1-p.

The SISO logistic regression model is [3]:

$$P(Y=1 \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)} = p \tag{1}$$

The MISO logistic model with k independent variables, will be

$$p = (Y=1 \mid x_1 ....x_k) = \frac{\exp(X^T \theta)}{1+\exp(X^T \theta)}. \tag{2}$$

where, $X = [1 x_1 x_2 ... x_k]^T$: is the vector of categorical variables, and $\theta = [\beta_0 \beta_1 ... \beta_k]$ is the parameters vector.

3. Log likelihood function for SISO logistic system

In order to identify a SISO type logistic process there are used the ***n*** pairs of input – output data experimentally obtained. These data are direct successive observations of that particular set of n entities in which each entity i is characterized by the pair of values (Yi and xi).

Based on these n pairs of experimental data those values of vector parameters θ need to be determined so that the model obtained can best describe the experimental data and to ensure a high level of generality, in the sense of being able to correctly describe the specific logistics process behavior in other points too *(y, x),* which are not part of the original set of *n* points of the experimental data. Among these points from the experimental data set there are some in which *Y = 1* and others in which *Y = 0.* Since the output of the process is a logistic variable which within the experiment takes the values *Y1, Y2 ,..., Yn* then the output of the model in the n experimental points is expressed by the probabilities sequence, *(p (Yi = 1 | xi) or p (Yi = 0 | xi) = 1 - p (Yi = 1 | xi)).* The probabilistic description of the entire set of *n* independent random events of logistic type is expressed by the product of *n* random probabilities related to observed binary random events:

$$P_n = \prod_{i=1}^{n} p_i \tag{3}$$

Within this product there are two types of terms: terms corresponding events for which Yi = 1, *pi = p = Pr (Yi = 1, xi, parameters)* and terms related to the events for which Yi = 0, *pi = 1-p.* Under these conditions the relation (3) becomes:

$$P = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i} \tag{4}$$

In [5], the probability function (4) is marked *L (data, parameters)* and is called the ***likelihood function [3]*** of SISO logistic regression. The likelihood function depends on the logistic regression *parameters* and *experimental data* with the following expression for the logistic model for the binary random events SISO type:

$$L((\beta_0, \beta_1); Data) = \prod_{i=1}^{n} \left(\frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}} \right)^{Y_i} \left(\frac{1}{1+e^{(\beta_0+\beta_1 x_i)}} \right)^{1-Y_i} = \prod_{i=1}^{n} \frac{(e^{(\beta_0+\beta_1 x_i)})^{Y_i}}{(1+e^{(\beta_0+\beta_1 x_i)})} \tag{5}$$

If the case of logistics processes identification the problem is to find those values for model parameters that will ensure the maximum likelihood function. These values, in the case of a SISO model logistics are noted: $\hat{\beta}_0$ *şi* $\hat{\beta}_1$ and constitute the so-called model parameter estimates for the purposes of maximum likelihood. The problem of maximum likelihood estimates for a logistic SISO regression model is:

$$\hat{\theta} = \arg\max_{\beta_0,\beta_1} \prod_{i=1}^{n} \frac{[\exp(\beta_0 + \beta_1 x_i)]^{Y_i}}{1 + \exp(\beta_0 + \beta_1 x_i)} \tag{6}$$

where, $\hat{\theta} = [\hat{\beta}_0 \ \hat{\beta}_1]^T$ is the parameters estimation vector of the SISO. logistic model .

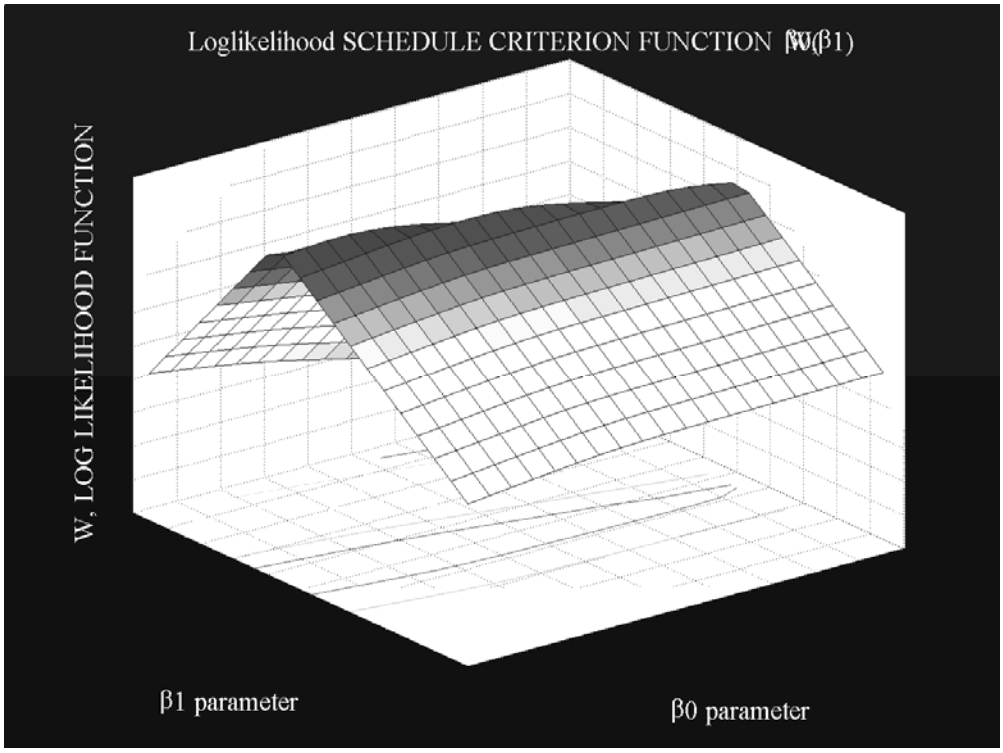

Fig.1. Graphic image of the surface described by the logarithmic likelihood function when experimental data logistics table 1

Applying the natural logarithm of the likelihood function (6) events results in the function log likelihood (LL) binary logistic model with random shit. This function denoted LL (βo, β1) has the expression:

$$LL\ l(\beta_o, \beta_1) = \ln[L(\beta_o, \beta_1)] = \sum_{i=1}^{n} Y_i(\beta_o + \beta_1 X_i) - \sum_{i=1}^{n} \ln[1 + e^{(\beta_o + \beta_1 X_i)}] \tag{7}$$

The functions *L ((βo, β1)* and LL *(βo, β1)* were maximum in the plane parameters βo, β1, at the same point coordinates: which is the maximum likelihood estimates of logistic SISO model parameters estimate.

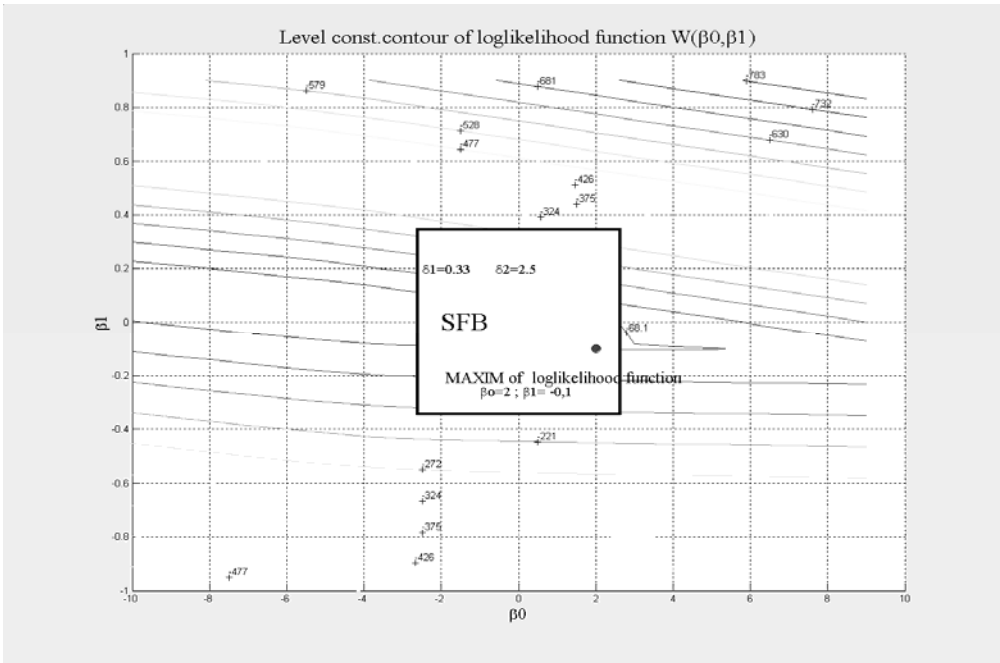

Fig. 2. Izolevel contour lines of the log likelihood function surface shown in Fig. 1 and SFB (Search Field Borders)

In order to solve the problem of maximum value we have available a set of *n* pairs of experimental data, observations input - output,

$$\text{data} = \{(X1, Y1), (X2, Y2), ..., (Xn, Yn)\} \tag{8}$$

which, for example, for the specific case of Table 1 these data are:

$$\text{data} = \{(X1 = 66\ Y1 = 0), (X2 = 70\ Y2 = 1), ..., (X23 = 76, Y23 = 1)\} \tag{9}$$

Using these data there has been developed a MATLAB program that built the graphic in Fig.1 of the logarithmic likelihood function (6) for the example in
Using these data there has been developed a MATLAB program that built the graphic in Fig.1 of the logarithmic likelihood function (7) for the example in Table 1. The plateau in the extreme area of the surface can be seen in Fig.1, which makes it more difficult to find the point of coordinates $\beta_0^{\wedge} and \beta_1^{\wedge}$ of the plane of the model parameters, corresponding to maximum function (7). This maximum point was also highlighted Fig.2 containing the image of the izolevel lines of the same log likelihood function. These contour lines were drawn using the same MATLAB program mentioned above. Given the issues mentioned on the log likelihood surface geometry for experimental data, we used a Monte Carlo method for searching the maximum point [5].

4. Experimental Determination of the Search Field Borders (SFB)

Classical Monte-Carlo algorithm(CMCA) is random testing of the Log Likelihood surface, using two test sequences of random numbers S1 and S0 of finite length, (one sequence for each parameter β_1 and β_0). These sequences are cut from infinite strings of random numbers uniform probability distribution in the band- plan under investigation in the two parameters area.

Step 1*: Generate a pair of random numbers [S0 (k = 1), S1 (k = 1)] with these values and existing experimental* ***data*** *(12) is calculated*
log Likelihood(1)=LL(1) and is stored in memory M:

LL (S0 (1), S1 (1),) = LL (1) →M

Step 2: *increment by one count variable k = k +1 a number of tests and generates a new pair of random numbers that are calculated*

LL (S0 (k), S1 (k), data) = LL (k)

Step 3*: Compare the L (k) with M from the previous step:*

IF,
LL (k)> M
THEN
replaced the old content is LL (k)→M and return to Step 2
OTHERWISE
return to Step 2, M preserving the previous value.

Fig. 3. CMCA for random search of the maximum log likelihood in logical SISO model case.

The two random sequences obtained from two random number generators in Matlab CMCA algorithm of random search of the maximum log likelihood LL (βo, β1) simulation, in the SISO case, involves the execution of three steps[5] described in Fig.3 The three steps described above are performed within SFB represented in fig.2. The condition to put to a stop the CMCA is expressed either by setting the maximum number Kmax test or by imposing a successful consecutive number of Ks steps in the search area experimentally delineated.
Limit values of the parameters (b0min, b0max, b1min, b1max) determines SFB. These limits are settled by means of pre explorations of the LLF values, made in a 9-node network shown in Fig. 4.

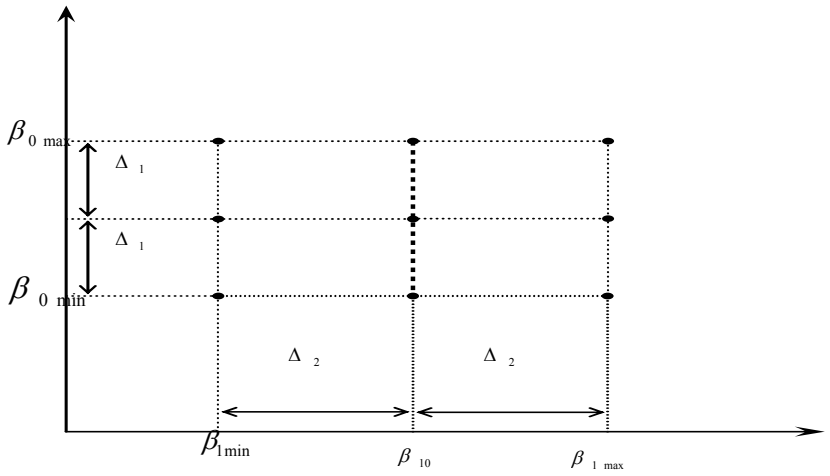

Fig.4. Experimental determination of the SFB: Δ_1, Δ_2 -limits of variation of parameters in the search process; β_{10}, β_{00} - SFB center coordinates in the plane parameters

The node from the network centre coincides with the SFB center only if LLF values in the 8 peripheral nodes are lower than the LLF values in the centre of the network. In the opposite case, when in one of peripheral nodes the SFB value is greater than the value of SFB in the center then the network is moved placing the node in the center in the point with the highest value of LLF. The search continues in the same manner until the greatest value of LLF is in center network. In the case study shown in table 1 the parameters variation limits $\Delta_1 = .5$ respectiv $\Delta_2 = 2$ and the initial coordinates dorm the centre of the network $(\beta_{10}, \beta_{00}) = (0,0)$ are arbitrary (Fig. 2).

5. The CMCA test results

CMCA Testing was performed for various N lengths of sequences of random numbers N: N = 50, N = 500, N = 1000 N = 1500 N = 10000. These variants were with CMCA. CMCA applying shown in Fig.3 for the two variants presented above, the results of the maximum LLF search process, shown in Fig. 6 were obtained. Furthermore there are presented two main observations, drawn from the analysis of experimental test results stated above.

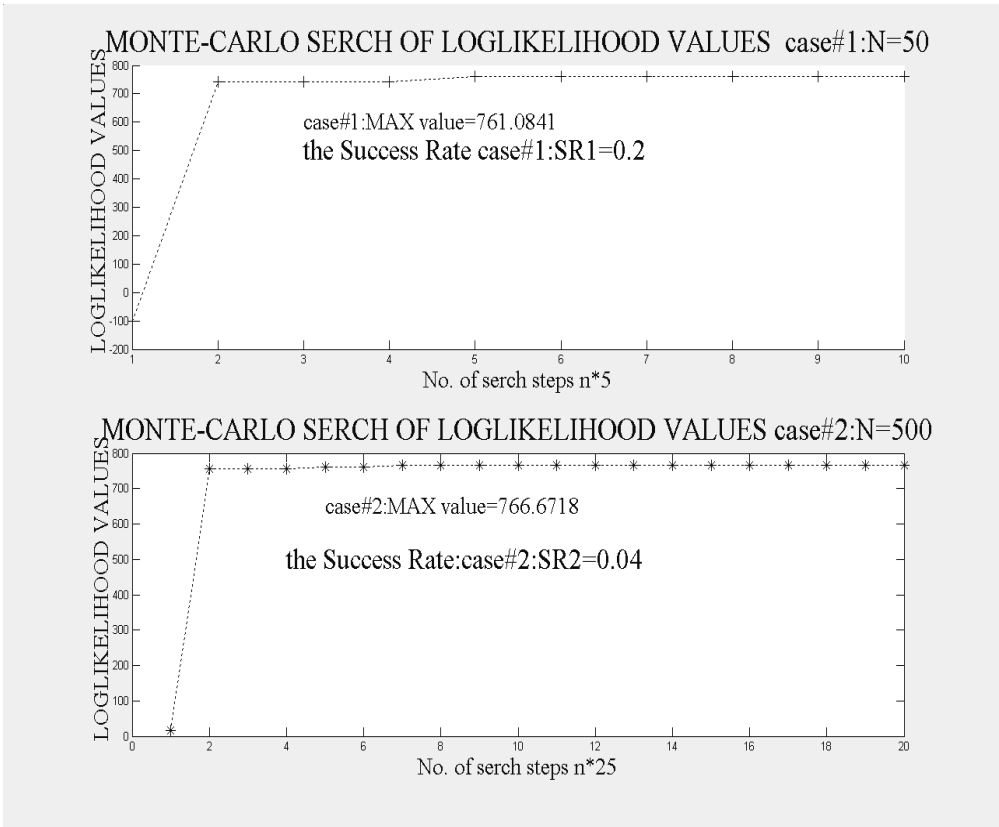

Fig. 5. Evolution of the maximum LLF search process in a choice N = 50 and N = 500
• Remarkable observation # 1 related to the modification of search efficiency depending on the length N sequence search

CMCA application results in case 1 Fig. 5 (for N = 50 steps for searching and data in Fig. 1) illustrates how the maximum LLF point in the parameters plane (point coordinates, β0 =- 0.10127, β1 =- 3.043 ') was found after only 20 practical steps. And in the second case for N = 500, the coordinates point, β0 =- 0.103, 3.003 =- β1 was found after 25 steps. *To assess the CMCA efficiency in all these cases, an indicator called "The Success Rate" was introduced, equal to the ratio between the number of successful steps and the total number of steps N.*

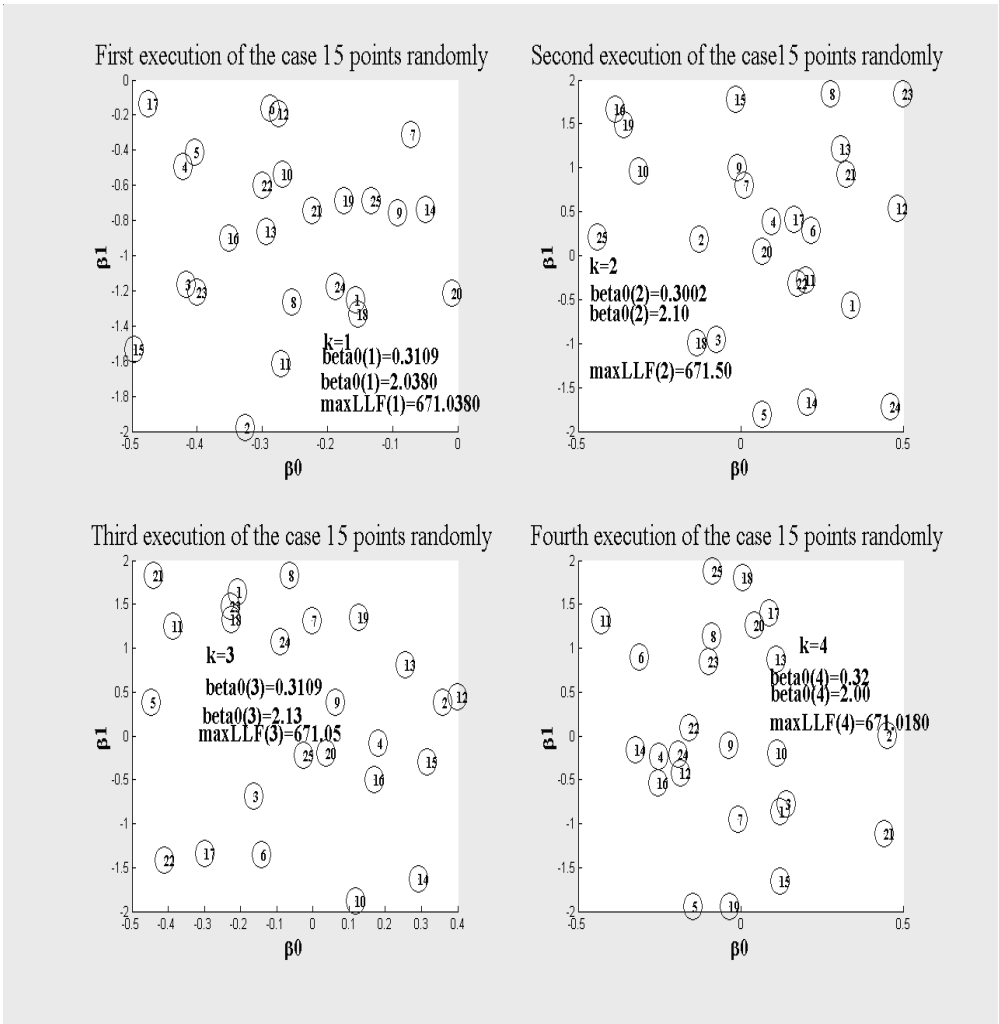

Fig. 6. Topography of the 15 search points by applying maximum LLF CMCA.

• Remarkable observation # 2 regarding the random character of the segments topography of finite length random numbers

If the sequence of 1500 random numbers can be imagined as consisting of 100 consecutive segments of the same length N = 15 random numbers. Each of these segments can be used for repeated searching of he maximum LLF with CMCA.

Fig. 6 shows both the positioning of the steps to be taken within the search (topography) and parameter values in the case of the first four sequences out of the 100 in which the string of 1,500 random numbers has been segmented. The apparent random nature of the topography as well as of the estimates can be noticed in all 100 cases:

$$\tilde{\beta}_0(1),:\tilde{\beta}_0(2),\tilde{\beta}_0(3),.....,\tilde{\beta}_0(100):\tilde{\beta}_1(1),:\tilde{\beta}_1(2),.....,\tilde{\beta}_1(100),$$

and, (10)

$$\tilde{\beta}_i(k)=\beta_i+\varepsilon_i(k)\ \ ;\mathrm{i}=0,1\ ,\mathrm{k}=1,2,...,100$$

where ε - random-noise with zero average value and dispersion σ_ε.

These observations lead us to the simple idea that instead of applying the CMCA only once on a long string of N random numbers we should apply CMCA by N repeatedly, = N / No of times on shorter sequences formed of No random numbers. Then, based on the results (10) we determinate the parameter estimates through mediation. In the case of a MISO type logistics model with *n*+1 parameters and Nr, we have short sequences repetitions:

$$\hat{\beta}_i=\frac{1}{N_r}\sum_{k=1}^{N_r}\tilde{\beta}_i(k)=\beta_i+\frac{\sum_{k=1}^{N_r}\varepsilon_i(k)}{N_r}\qquad , \tag{11}$$

where i = 0,1,2 ,..., n
From (11) result,

$$M\,(\hat{\beta}_i-\beta_i)^2=\frac{N_r{}^2\sigma^2{}_{\varepsilon i}}{N_r}=\frac{\sigma_{\varepsilon_i}{}^2}{N_r}=\sigma_i{}^2\qquad , \tag{12}$$

where M is mathematical expectancy operator and σ_i is the dispersion estimates of the parameters. For N, >10 the noise dispersion can be approximated by the following relation:

$$\hat{\sigma}_{\varepsilon_i}\approx\left[\frac{1}{N_r-1}\sum_{k=1}^{N_r}(\tilde{\beta}_i(k)-\hat{\beta}_i)^2\right]^{1/2} \tag{13}$$

If samples of length n from a population are extracted, then for values of n> 10 the sample averages are distributed (approximately) normally (according to the central limit theorem [6]). Given (11) it results that the distribution of random values probabilities of the estimates $\hat{\beta}_i$ are asymptotically Gaussian. Thus, you can apply the well known rule of the "three sigma" for determining the estimate probability $P(|\beta_i-\hat{\beta}_i|)$:

$$P\left(| \beta_i - \hat{\beta}_i | < 3\frac{\sigma_{\varepsilon_l}}{\sqrt{N_r}} \right) \cong 0.997 \tag{14}$$

From (14) results that the probability value is very close to one, for the inequality to be fulfilled,

$$\left(| \beta_i - \hat{\beta}_i | < 3\frac{\sigma_{\varepsilon_i}}{\sqrt{N_r}} \right) \tag{15}$$

6. Conclusions

The paper highlights the models of logistic processes particularities with random binary events and presents a technique for identifying these processes. In order to estimate the logistic model parameters, it is necessary to apply the statistical criterion maximum likelihood. Original Contributions:

1) A Monte Carlo method is put forward in order to estimate logistic model parameters using the maximum log likelihood criterion;

2) Statistical analysis of parameters estimate. The estimates of parameters, thus obtained, are not deviated and the estimates variances can be approximated by (15). In conclusion we consider that the theory of systems modeling and identification [3] should be extended to the field of random binary events systems.

R E F E R E N C E S

[1] C. *Penescu., M. Tertişco, E. Ceangă,* Identificarea experimentală (Experimental Identification- in Romanian), Editura Tehnică, Bucureşti, 1971

[2] *F. Bădulescu, F. Gorunescu,* Informatica oncologică: (Oncologicals Informatics – in Romanian), Ed. Didactică şi Pedagogică, Bucuresti, 2003

[3] *D. Ştefănoiu , J. Culiță, P. Stoica,* Modelarea şi identificarea sistemelor (Systems Modeling and Identification – in Romanian), Editura Printech, Bucureşti, 2005

[4] *M. Tertişco, P. Stoica,* Identificarea şi estimarea parametrilor sistemelor (Sistems Identification and Parametrs etimation – in Romanian), EdituraAcademiei, Bucureşti, 1980

[5] *M. Tertişco, P. Stoica,* Identificarea asistată de calculator a sistemelor (Computer-aided identification system – in Romanian), Editura Tehnică, Bucureşti, 1985

[6] *V.S. Antyufeev, A.L. Marshak,* Monte Carlo method and transport equation in plant canopies, *Remote Sens. Environ.* 31:183-191 USA, 1990

[7] *R.Y. Rubinstein, D.P. Kroese,* Simulation and the Mont Carlo Method, New York: John Wiley & Sons, 2007

[8] *P. Ojeda, M. Garcia, A. Londono, N.Y. Chen,* "Monte Carlo Simulations of Proteins in Cages: Influence of Confinement on the Stability of Intermediate States", Biophys. Jour. (Biophysical Society) **96** (3): 1076–1082. (Feb 2009).