The Technical Process Properties Monitoring Based on the Data Normalization Method

Tatiana I. Lapina[1]

Abstract— Some non-conventional methods of model identification based on data normalization are presented In the article. The named methods are used for tendencies approximation and forecasting, and moments of the stochastic process properties divergence identification in case of providing stability of a technical process.

Keywords— data mining

Introduction

Urgency of development of methods for analysis and processing of statistic information is explained by the fact that previously developed and successfully user methods for solving technical problems in many cases are completely unfit for use in economy, medicine and social sphere because of much greater complexity of investigated object and boundary conditions, or they require for larger costs of computational resources, that does not correspond to received result. Thus, development of effective methods for analysis of statistic data is one of the important task while constructing models adequate to the object of the investigation.

Among the problems, which have to be solved in the process of statistic analysis of data, one may select the following:
- Definition of similarity of sampling data, received as a result of the experiment or corresponding to different time references.

- Definition of the moment change of stochastic processes properties according to sampling data.
- Definition and description of distribution functions class within which search of the model type will be carried out.
- Construction of the model adequate to the object of investigation.
- Evaluation of precision of definition of statistic dependence slass.

Methods, based on the use of order statistics (G.Daivid), rank criterions (Ya.Gaek, Z.Shiduck), rank correlations (M.Kendal), robust methods (P.Khiuber), graphic methods of analysis (J. Tiuky). Nontraditionally used while solving problems of identification.

However, received theoretical results do not always meet the requirements of practice, especially in conditions of limited (n<50) volume of data and high a priori uncertainty on probability characteristics of phenomena being investigated, Particularly this is relative to the problems

1 Kursk State Technical University, Kursk, 305040, ul.50 Let Oktyabrya, 94, lapinati@mail.ru

of distribution forms identification according to sampling data since they are solved by using limited set of analytical models. Real processes may not always be referred to well-known classes of models with sufficient degree of adequacy.

New direction in statistic methods of analysis based on the use of information criteria at decision making on the form of distribution of sampling data (Ivchenko B.P., Martyshenko A.A., Grigorovich V.D., Yudin S.V.) is being developed over the last years.

This article deals with the questions of models forms identification based on the data normalization method and using information criteria. Recommended methods give saitable tools for processing some types of statistic experimental data and definition of selected model correctuess.

The Method of Data Normalization

Quality of economic system control is defined in many respects by accuracy of its description or by the quality of the used model, that is by the quality of identification. Quality of of identification, in its turn, is characterized by adequacy of the model to the real object. Mapping of real objects into models space is always connected with loss of information. That is why it is impossible to speak about absolute adequacy of models to real objects. Degree of inadequacy is defined by:

- error of initial premises (theoretical statements about observed object) in the task of function type and structure;
- approximate properties of models;
- error of measurements (instrument error).

Use of inadequate models in practice may lead to faulty theoretical and practical conclusions.

Among the tasks, which are necessary to solve while analyzing statistical data in order to obtain the correct description of the object it is possible to state the following ones:

• the determination of the selected data, obtained as a result of experiment.
• determination of the moment when properties of stochastic processes change according to the selected data.
• creating of a model which will be adequate to studies process.
• prediction of an object or a process state.

The object of this article is to represent a method of statistical information analysis based on data normalization. The named method presents a wide range of possibilities for experimental data processing and model identification.

Data normalization method helps to unite the traditional statistical methods which represent data as a function belonging to different distributions with probability density functions and quantitative estimations based on scarce amounts of initial information combined with visual methods, which let to understand a problem, describe it in a formal language, determine criteria for comparison of alternatives and determine a rational formalization of the problem.

Methods of normalized statistics make it possible to obtain mathematical description of the probability density function of the distribution via the composition of the distribution, normalized to the interval [0,1]:

$$f_{\hat{y}_x}(y) = \sum_{i=1}^{n} p_i f_{\hat{x}_i}(x),$$

where

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \in [0,1].$$

The known distributions can be used as distributions or elementary functions, and by the linear transformation of function and argument they can be lead to interval [0,1]:

$$z = a \cdot x + b,$$

$$y(z) = c \cdot f(z) + d;$$

so as to they would satisfy the requirements of the probability density distribution.

$$y_1(z) \geq 0 \qquad z \in [0,1]$$

$$\int_0^1 y_1(z)dz = 1$$

Thus, any continuous and non-negative in a certain interval function can be considered as a probability density function, after pressing or extending its curve at a chart with respect to the weight coefficient.

№	Data Value
1	0.74
2	0.416
3	0.431
4	0.725
5	0.436
6	0.401
7	0.422
8	0.454
9	0.408
10	0.434

***Table 1**. Data value controls parameter*

In order to automate methods of data analysis based on normalized statistics the automated system was created. It helps to carry out studies of distributions structural properties and to obtain their quantitative characteristics, to carry out comparative data analysis, etc.

Let us illustrate an example by applying the method or normalized distributions to a sample.

The sample data, were preliminary normalized to an

interval [0, 1] according to the formula:

$$x_i = \frac{x_i - x_{min}}{x_{max}}$$

Let us calculate the mean and the central moments which represent "statistical indicators" of the probability density function of the distribution:

$$m_x = 0.495,\ \sqrt{\mu_2} = 0.377,\ \sqrt[3]{\mu_3} = 0.186,\ \sqrt[4]{\mu_4} = 0.415$$

These estimators are taken as the coordinates of the certain point P in the four-dimensional space. In the same space we will add points, which correspond to the existing data from the relative points of standard distributions stored in the automated system. Then, select from the points from the space which has a minimal distance to P.

This point became point F_{min} with coordinates

$$(0.498, 0.392, 0.190, 0.406)$$

$$(|PF_{min}| = 0.018).$$

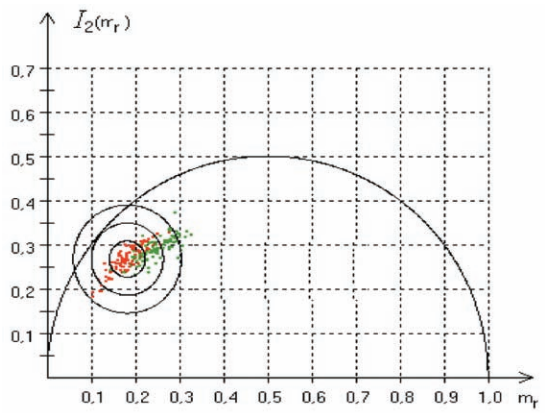

Figure 1. The indicators received from the sample

The following density function of distribution corresponds to it:

$$f(x) = 0.5834 \cdot \left|\sin\left((0.6 - x)\cdot\frac{\pi}{3}\right)\right|^{2\cdot\sin(2\cdot(0.8-x)\cdot\pi)}$$

The curve of this density probability function is given below on Figure 2.

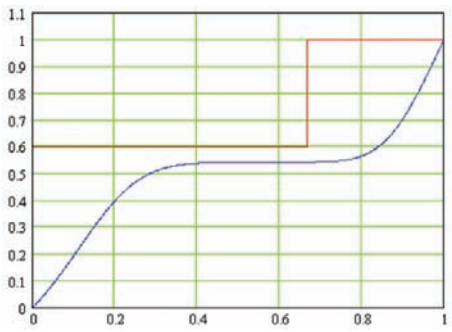

Figure 2. The indicators received from the sample

To confirm a hypothesis of the distribution law, we used the Cholmogorovs criteria. The divergence between the experimental data and data the approximated with the model is:

$$D = \max\left|F_n(x) - F(x)\right| = 0.6$$

A variable **λ** can be calculated as

$$\alpha = 0.05 \quad \lambda_\alpha = 1.36 \geq \lambda$$

As for $\alpha = 0.05 \quad \lambda_\alpha = 1.36 \geq \lambda$, than the model can be accepted as an adequate to the data set.

The data normalization method helps to decrease significally the number of calculations and to receive the adequate result when working with short data sets.

The problem of the divergence moment identification is one of the key ones in dataset structural analysis. It can be solved with one of the following approaches: first, by segmentation of the curve, which simulates probability density function of the distribution i.e., boundaries between the homogeneous fragments of the curve can represent moments of divergence of the stochastic process properties; second, the position of statistical estimators in the possible values space indicate the moments when the properties of the stochastic process change. Thus, it is possible to obtain description of process structural properties before and after the moment of divergence, and to receive the quantitative characteristics of the process.

An example of the divergence moment identification is represented on Figure 3.

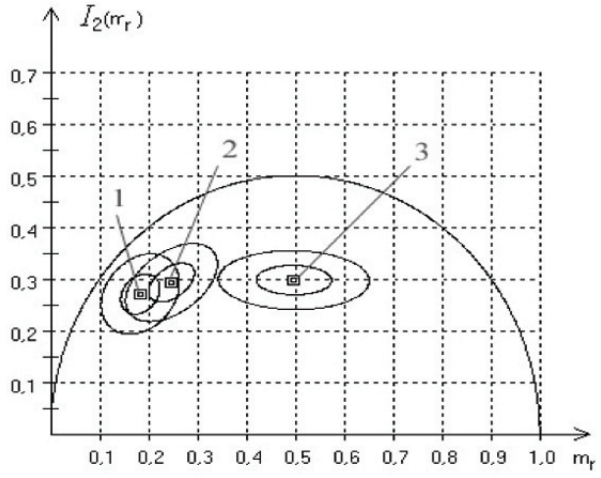

Figure 3. The statistical indicators grouping at the divergence moments

For definition of quantitative characteristics of change of properties of casual parameter it is possible to use information criterion.

Informational criteria can be used to define quantitative characteristics of random variable properties

changes.

To receive the distribution density function of chance variable we should brake value space of this variable into finite number k of intervals A: (x_{i-1} x), where x_i {i = 1,k) are frontier points of intervals, P_1• P_2,....,P_n- probability of chance variable X values hitting the intervals A_1,A , ...,A :

$$Pi = P(x \in Ai) = \int_{x_i=1}^{k} f(x)dx \quad (1)$$

It is known from the information theory that the entropy independent chance variable X is defined as

$$h = -\int_{-\infty}^{x} f(x)\ln f(x)dx \quad (2)$$

It is shown in [1] that for the entropy H of discrete variable received from continuous variable

$$H = h - \ln \Delta x \quad (3)$$

where Δx is the width of interval.

Formula (3) involves the logarithm of dimensional quantity Δx From physical point of view the logarithm of dimensional quantity is not specified, so we have to introduce chance variable $y = x/\sigma$ then

$$H = h - \ln \Delta x/\sigma \quad (4)$$

Lets consider a case when the dispersal of controllable parameter X values is defined by the borders of control a and b (a<b).

Let $p_i = P(x < a)$ be the probability of the value of random parameter X is less than a; $p_i = P(x > b)$ is the probability of the random parameter X value is more than b, $p_i = P(a \le x \le b)$ the probability of the random parameter X value lies in the interval from a to b.

The condition entropy of such random process is defined as

$$H = -p_1 \ln p_1 - p_2 \ln p_2 - p_3 \ln p_3$$

According to (3) and (4) the condition of a random process is defined by the entropy:

$$H = h - \ln(b-a)/\sigma \quad (5)$$

If measure of inaccuracy of controllable parameter can't go beyond an allowed value Tн, it means that the centre of a dispersal meaning of a parameter (a math expectation) field coincides with the centre of a tolerance band, and the mean-square deviation corresponds to an equation.

$$\sigma_0(x) = [T_H - \Delta x]/U_{q/2}$$

Where $T_н$ is a low limit bound of a measurable parameter, Δx is the centre of tolerance band (a math expectation of a reference distribution of measurements inaccuracy), $U_{q/2}$ – a fractile of a preassigned order of obtained distribution, $\sigma_0(x)$ - mean-square deviation of reference measurement inaccuracy distribution.

Let the measurement of parameter in a point of time to correspond to reference distribution of inaccuracy with mean-square deviation $\sigma_0(x)$. In the point of time t_1 mean-square deviation changed value to σ_1. Then the alteration of random parameter state during Δt= t_1 - t_0 is equal to entropy difference ΔH:

$$\Delta H = |H_1 - H_0| = \ln\sigma_1 / \ln\sigma_0 \quad (6)$$

Due to a logarithm property it is equivalent to

$$\sigma_1^2 = \sigma_0^2 e^{2|H_{t1} - H_{t0}|}$$

Thus the alteration of random parameter value can be controlled by random parameter dispersion $\sigma_1^2(x)$ estimation.

In practice, the estimation $\hat{H}_1$ is used instead of entropy value H_1:

$$\hat{H}_0 = -\sum_{i=1}^{3} \frac{f_{i0}}{n} \ln \frac{f_{i0}}{n}$$

$$\hat{H}_1 = -\sum_{i=1}^{3} \frac{f_{i1}}{n} \ln \frac{f_{i1}}{n}$$

Where fi is the frequency of controlled parameter X measurements hitting the i field of measured parameter value space; *n* is the amount of sampling.

In economics processes monitoring, when it is rather difficult to get the sufficient quantity of experimental data, data valuation methods and informational and statistical methods application helps to get information about regularities and tendencies of objects behavior, define similarity or difference of objects or processes structure properties, their quantitative and qualitative estimations, helps to interpret experimental data statistically and to analyze data real-time.

Conclusions

Quality of economic system control is defined in many respects by accuracy of its description or by the quality of the used model, that is by the quality of identification. Quality of of identification, in its turn, is characterized by adequacy of the model to the real object. Mapping of real objects into models space is always connected with loss of information. That is why it is impossible to speak about absolute adequacy of models to real objects. Degree of inadequacy is defined by:

- error of initial premises (theoretical statements about observed object) in the task of function type and structure;

- approximate properties of models;

- error of measurements (instrument error).

Use of inadequate models in practice may lead tofaulty theoretical and practical conclusions.

The method of Data Normalization can be effectively used in those fields of industry and technology, which suffer from financial and factual difficulties while gathering experimental data. The named method helps to receive information about some tendencies and regularities in process and objects dynamics, identify similarities and differences of their structure, their quantitative estimations, to provide the statistical explanation of experimental data and real-time analysis.

Bibliography

• Urazbahtin I.G., Lapina T.I., Random process from modeling based on statistical indicators for led distributions. // News of KSTU, KSTU publishing, №4, 2000.
• Lapina T,I., Urazbahtin I.G.
• Data Normalization as a basis for Time Series forecasting THE SOCIETY OF PHOTO-OPTICAL INSTRUMENTATION ENGINEERS (SPIE)
• Central Univ./City Branch, 9876 Light Ave., Philadelphia, PA USA, 2005.
• Efron B. Nontraditional methods of multidimensional statistical analysis. – Collection of articles. – Moscow, Finance and Statistics, 1988.
• Principal components of Data Series: Caterpillar method. Under supervision of Dalilov D.L., Zhigliavsky A.A., S-Peterburg University, 1997.

Author: Tatiana I. Lapina
Born in Russia, Kursk, In 1985 graduated from the Engineering Department and in 1998 obtained her Ph.D. in " Management in social and economic systems " from the Kursk State Technical University. Now she is a associate professor of the Information systems and technologies Department of the Kursk State Technical University.

Tatiana I. Lapina is interesrted in aspects of the creation of a system for objects monitoring based on informational statistical approach, processing of one-dimensional and many-dimensional signals.