EIDOS, 4
Marzo-agosto 2011, 43-49
ISSN:1390-499X
eISSN:1390-5007

TIME SERIES FORECASTING BASED ON DATA NORMALIZATION METHODS

T. Lapina[1], South-West State University, Russia

Recepción/Received: 2011-03-06
Aceptación/Accepted: 2011-07-12
Publicado/Published: 2011-08-05

Resumen -La utilización de series de datos es de gran importancia cuando analizamos procesos de diferente naturaleza. En el artículo se presentan algunos métodos no-tradicionales de análisis de la dependencia, aproximación y predicción, basados en la normalización numérica.

Palabra Clave -Predicción de procesos casuales, indicadores estadísticos

Abstract -Using Data series is of great importance when examining processes of different nature. In the article some nontraditional methods of dependence analysis, tendencies approximation and forecasting techniques based on Data normalization are represented

Keywords -Forecasting of casual processes, Statistical indicators

[1] Tatiana Lapina, lapinati@mail.ru, Universidad del Sur-Oeste de Rusia

INTRODUCCIÓN

There is a number of methods dealing with making forecasts, which can effectively be used for predicting the statistical evaluations. Any standard procedure of forecasting includes such necessary elements as making a preforecast orientation, i.e., the determination of purposes, tasks, period of prevention, collection and data analysis in the interval of retrospection and identification of a forecast model. The most important procedure here is a construction of the forecast model, which can be executed by one of the known methods, for example, heuristically, by constructing an analytical model in case if there are known general laws which determine development of process, common structure, most important analytically expressed functional connections, and also control sample, which allows to verify the fitness of the model, obtaining the statistical model of forecast based of statistical data set, that characterize the period of retrospection. The application of traditional analytical models of stochastic processes is connected with the a priori search for the structure of these models and often characterized with limited information about the nature of the process dynamics. The parameters determination of the statistical model of a process and extrapolation accuracy estimation will require statistical data, which characterize process in the period of retrospection. So, all the spoken above indicate decrease of the conclusions authenticity in process extrapolation. So, it seems of great importance to examine the approach without connections to a rigid structure of the process model and limited requirements for the information range.

Collection and information processing (a semantic content and quantity of which is determined by the special features of each period) is performed at all the periods of the control process simulation in the social and economic systems, beginning from the gathering the general information about the system and concluding with creating a model.

The interferences, which appear while obtaining and information processing, lead to the fact that between the real data and the results of characteristic measurement of process there is only an stochastic connection, that is why parameters of the process carrying information about the identified object are of probabilistic nature. Because of the special features of social and economic processes, model can be created only with statistical data.

Models and methods of control can be adequate to real processes under conditions of limited data sets only in case of appropriate identification methods.

An experimental data analysis, as a rule, is connected with processing of time series processing, which includes analysis, forecasting, current analysis of prediction and data processing automation.

An applied statistics is intended for data analysis in the presence of stochastic actions. Among the tasks, which it is necessary to solve statistical data analysis to obtain the correct description of the object it is possible to state the following steps:

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

In spite of the large number works of on the indicated directions, the tasks, especially prediction and analysis of the effectiveness of the forecast of time series is studied insufficiently.

In the article some nontraditional methods of dependence analysis, tendencies approximation and forecasting techniques based on Data normalization are represented.

Data normalization methods are oriented to information obtaining while processing small amounts of experimental data.

The idea of the approach is based on presentation of different classes of distributions led to the interval [0; 1]. The approach allows to examine structural properties of distributions and to find the quantitative evaluation of some of their characteristics.

Normalization gives a possibility to combine the classical methods of statistical data analysis taking data series as special sets belonging to different distributions. The sets are determined by probability density function and some 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. To solve the problem, we need to fit a mathematical description of the probability density function by composing different distributions normalized to the interval [0, 1].

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]:

$$f_{\hat{y}_x}(y)=\sum_{i=1}^{n} p_i f_{\hat{x}_i}(x), \text{ where } \sum_{i=1}^{n} \mathrm{p_i}=1, \quad \mathrm{p_i}\in[0,1] \quad (1)$$

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.

$$y_1(z)\geq 0\,, \quad z\in[0,1], \qquad (2)$$
$$\int_0^1 y_1(z)dz=1$$

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. All these procedures are to be made in an order represented on a figure 1.

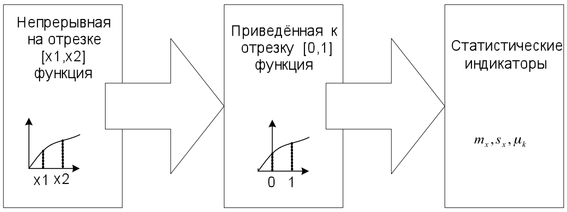

Непрерывная на отрезке [x1,x2] функция
x1 x2
Приведённая к отрезку [0,1] функция
0 1
Статистические индикаторы
m_x, s_x, μ_k

Figure 1. The stochastic functions transformation order continuous function on interval [x1, x2] normalized to the interval [0, 1] function Statistical indicators

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. The process of obtaining a model with the automated system includes the following rules for stating distribution:

- possible indicators values of stochastic distributions are limited to zero and functions for central moments of the Bernoulli distributions are symmetrical to the value the mean, which is equal to 0.5.
- the method of linear transformation of function

$$z=a\cdot x+b\,, \;\; y_1(z)=c\cdot f(z)+d\,; \qquad (3)$$

and argument suppose normalizing any function with assigned space of definition to the interval [0, 1], by the linear transformation of argument and function.

$$1. \;\; y_1(z)\geq 0\,, \quad z\in[0,1], \qquad (4)$$
$$2. \int_0^1 y_1(z)dz=1\,.$$

the method of distributions composition suppose forming a distribution law as composition of the distributions, normalized to the interval [0, 1]:

$$f_{\hat{y}_x}(y)=\sum_{i=1}^{n} p_i f_{\hat{x}_i}(x), \quad \sum_{i=1}^{n} p_i=1, \;\; p_i\in[0,1]\,. \qquad (5)$$

These methods let to obtain “own indicators” of the known classes of distributions, different elementary functions and their compositions, which satisfy probability density distribution.

The automated system let us calculate “statistical indicators” of the analyzed stochastic process, compare them with existing “own indicators” of distributions and select the form of model.

Let us apply the method or normalized distributions to a sample. With sample data, which were preliminary normalized to an interval [0, 1] according to the formula:

$$\hat{m}_x=\sum_{i=1}^{n}\hat{x}_i/n\,, \;\; \hat{I}_{\text{к}}=\sqrt[\text{к}]{\sum_{i=1}^{n}(\hat{x}_i-\hat{m}_x)^{\text{к}}/n} \qquad (6)$$

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 can be taken as the coordinates of the certain point P in the four-dimensional space. In the same space we will add points, which corresponds to the existing data from the “library” of density function. Let us 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)($|PF_{\min}|=0.018$).

The following density function of distribution corresponds to it (fig2):

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

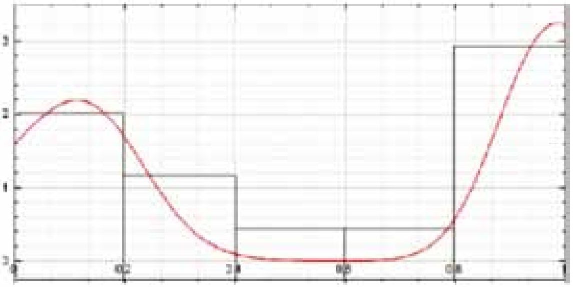
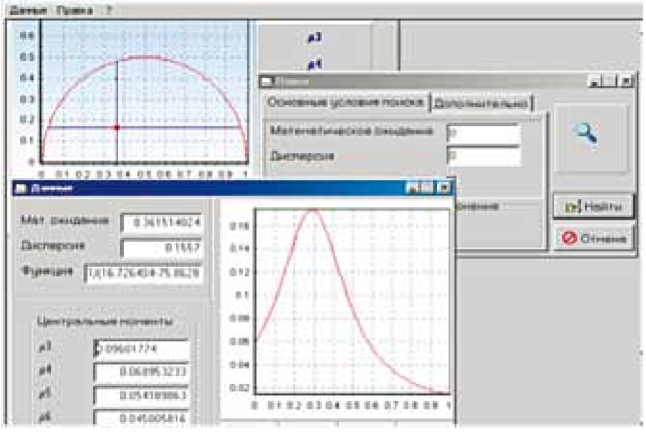


Figure 2. Curve of given density probability function

Having tested a hypothesis about the distribution, we found that the obtained function with significance level $\alpha = 0.05$ is the distribution density function of the studied process. An example of work is represented on figure 3.



Математическое ожидание
Дисперсия

Мат. ожидание 0.361514024
Дисперсия 0.1557
Функция 1,0/16.726404*Z5.0628

Центральные моменты
μ3 0.09601774
μ4 0.060953233
μ5 0.054189863
μ6 0.045005816

Figure 3. An example of using automated system for statistical data analysis.

Comparison of different distributions is a basis for creating models, which are represented with compositions of moments of their "own indicators" normalized to interval [0, 1] form and in finding "statistical indicators". These "statistical indicators" are calculated as the closest from their geometrical positions for moments data sets, which are also calculated for the sample.

"statistical indicators" for a big number of possible probability density functions are saved at a data base. The automated system uses the data base in order to find closest distributions.

On the Fig.4 is shown an example of obtaining a distribution for experimental time series using "statistical indicators".

The process of obtaining a model from an analyzed sample using "normalized distributions" method is described at work [1].

While analyzing such types of data usually we suggest that observed curve is an observed realization of a certain stochastic process. Frequently it is necessary to carry out the comparative analysis of two realizations of stochastic process or two data sets, obtained from different objects. On the basis of statistical data normalizing it is possible to analyze the structural properties of distributions positions of the statistical indicators of distributions in the space of possible values.

If we observe similar or close estimations for statistical indicators in the space of possible values, we can make a conclusion about similarity of their density probability functions, as they are described with the same elementary functions.

The divergence of the position of statistical indicators indicates different structural properties of the objects being investigated. The problem of realization of structural analysis of the selected data is closely related indication moments of stochastic process properties changing, which can be solved as follows:

first, by segmentation of the curve, which simulates density of the selected data distribution i.e., boundaries between the homogeneous fragments of the curve can represent identification of one or more moments of changing (or divergence) of the stochastic process; second, the position of statistical indicators in the space of possible values speaks about change of the properties of stochastic process as a whole. In this case it is possible to obtain description of process structural properties before and after the moment of discord, and it means calculation of the quantitative characteristics of process (figure 5).

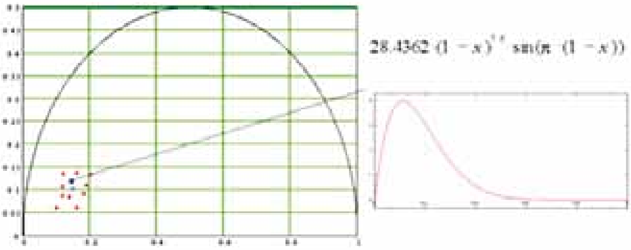


Figure 4. An example of distribution form creating for experimental data by method of "statistical indicators".

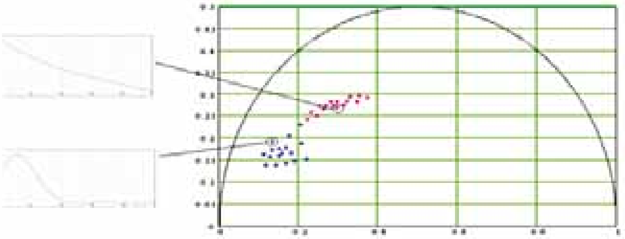


Figure 5. Analysis of structural divergence between two data samples

Methods of Data Normalization are of great importance for those cases then it is hard to obtain experimental information because of its high costs or other reasons. In such situation mentioned methods make it possible to get information about tendencies in object behavior, to determine similarities and differences in object (or process) structural properties, calculate qualitative and quantitative estimations, provide real-time statistical descriptions of experimental data.

Now solution of many problems in social and economic systems connected with analyzing time series and solving such problems as time-series analysis, forecasting. To solve problems dealing with time series analysis with stochastic noise methods of applied statistics are used. These methods also include statistical forecasting as a part.

Forecast can appear ineffective in case if stochastic process changes its properties. That is why forecast's ineffectiveness (divergence) determination is of great importance and helps to avoid errors while dynamic of time series changes.

As a "divergence moment" we understand such a moment when occurs change in average values of observed time series of errors between simulation and forecast.

In some cases it is possible to extract data sets, which describe stochastic process as composition of homogeneous fragments. In that case we can make a structural data analysis, which includes extracting homogeneous fragments on an experimental curve and their description.

Structural approach to data analysis is closely connected with identification of moments when stochastic process changes its properties. It is possible to find boundaries between homogeneous fragments by finding one or several moments of divergence of the stochastic process. On the other hand, if we do not know parameters of the stochastic process before or after divergence moment, or do not know them exactly, then identification of moments of stochastic process properties changing includes evaluation of parameters, i.e. description of experimental data before and after divergence.

In paper a new approach to evaluation of statistical forecasting reliability based on normalized statistics is proposed.

Using traditional analytical models for stochastic processes is connected with the a priori identification of the structure of the models, which mostly takes place with scarce information about the nature of the process dynamics.

Parameters of statistical model of the process and extrapolation accuracy estimation require that there is statistical data characterizing process during previous periods presented. But this may reduce reliability of the results of extrapolation. That is why it is seems to de of practical and theoretical importance to examine an approach, which don't use fixed model structure and do not have requirements to sample size.

The method represents dynamic sample (which is used for extrapolation) as an oriented process, for which it is necessary to know how process dynamics changes in time. For any process we can use the following equation to describe its dynamics:

$$f(t) = f_T(t) + f_n(t) + \varepsilon(t),\ t \in [0,\mathrm{T}],\ \ t \in [0,\mathrm{T}], \qquad (7)$$

where:

fT(t) – a component of a signal, which changes slowly, usually called as trend,

fn(t) – a cycling component;

□(t) – a stochastic noise, which can be described by stochastic process.

Dynamics of any process can be determined by functions, consisting from several component.

Forecasting of stochastic processes requires that we extract the main component (no stochastic) from a sample during time-series analysis. approximation methods and harmonic Fourier analysis are used most frequently for the solution of this problem. However, these methods proved to be ineffective for short samples and in case if there is no a priori information about the trend component frequencies.

In work it is proposed to use a method "caterpillar" to extract main component from an initial data series [3].

Let us break sample F=(f1, f2, f3, … f$_N$) in accordance with "caterpillar" method into additive components

$$F = \sum Fi$$

In other words, let us realize the singular decomposition of initial data series. The matrix, whose elements are equal to $x_{ij} = x_{i+j-1}$, we will consider as M-dimensional sample which obtain k elements or M-dimensional data series, which is represented by M-dimensional trajectory (broken line in M-dimensional space of k-1 components). So, we will obtain the matrix X, where "line - individual, column - property".

Every vector of the matrix can de treated as a trend, cycle or noise. To extract the main component of data series or trend we will use methods of normalization statistics, which let us to obtain a mathematical description for probability density function by composing of different distributions led to the interval [1] in accordance with its statistical indicators in a space formed by statistical indicators of standard distributions.

$$\hat{m}_x = \sum_{i=1}^{n} \hat{x}_i / n,\ \ \hat{I}_{\kappa} = \sqrt[\kappa]{\sum_{i=1}^{n} (\hat{x}_i - \hat{m}_x)^{\kappa} / n} \qquad (8)$$

As a trend is a component of a data series which changes slowly, then all samples belonging to a certain trend are to have similar structures. Is the samples are brought to the same scale and calculate their statistical estimations (mean square error, variance, etc.), then values of the estimators are to be almost the same (Fig. 6a).

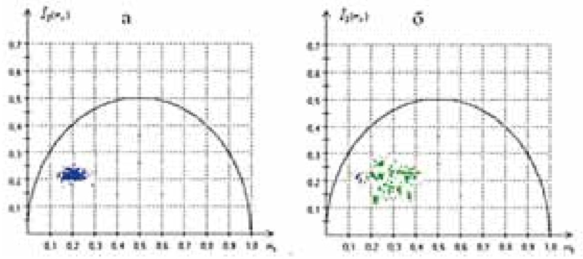


Figure 6. Possible estimation values and graphical presentation of statistical indicators for normalized distribution

Let us led vectors to the interval [0, 1] in accordance with the following expression:

$$\widetilde{F}_j^i = \frac{F_j^i - \min(F^i)}{\max(F^i) - \min(F^i)} \qquad (9)$$

Using bystrep- method, will create N adequate samples $\widetilde{F}_j^i$ from every vector $\widetilde{F}_j^i$ Then calculate mean square error m_x and standard error of the mean s_x for every sample and will accept these two estimations as a certain point in a two-dimensional space.

The closer value of the component to a data series dynamics, the closer value of ρ^i to 0. having extracted a cycling component using such methods like Fourier analysis and, adding its value with the trend, we determine no stochastic component of data series dynamics.

The value of changing variable of the process for every retrospective moment can be represented like:

$y_l=y_{l-1}+e_{l-1}$, i=l, ...,N,

where:

y_i - value of data series dynamics at i step of the retrospective period;

y_{i-1} - value of data series dynamics at the previous moment of time;

e_{i-1} - increment of the variable at i step;

N- number of points of dynamic data series.

As values of increments are stochastic, there is possible to determine a distribution and its characteristics for them. We will also need to take into account dependence between the previous and posterior increments.

Suppose, that during the forecasting period dynamic data series change the same way, then it is possible to use characteristics of increments and statistical testing in order to modell increments change for the forecasting period.

For every i step for the forecast period we will have the following:

$x_j=x_{j-1} + e_{j-1}$, j=l,...,M,

where:

j – number of a step at the forecasting period;

M - number of the steps at the forecasting period;

x_j - value of a variable at the previous step;

x_j - a modelled value at the j step.

Now we can obtain a formula for process variable:

$$x_{\text{пр}} = y_N + \frac{1}{k}\sum_{j=1}^{M} e_j$$

Summing all the spoken above, we can say that the proposed method let us simply extrapolate a stochastic process, but we have to notice that there is some inconvenience caused by using of initial data normalization.

As selection a model's form in normalized distribution method is based on indicators positions in space of indicators $S_k = m_r \times \sqrt[k]{|\mu_k|}, k = 2,3,...$, then if indicator's position is changed, it means, that the model's structure also changes. A simulation function is selected agree with criteria of minimal divergence between theoretical and sample distributions. In other words, the simulation function must have minimal divergence between its "own indicators" and standard statistical indicators for sample in space of moments.

Thus, indicator's position changing is related to model's structure changing. If statistical indicators do not significantly change their values in space of indicators while experimental data processing, it is possible to conclude, that structural properties of the model do not change with time (Fig.7a). If statistical indicators change significantly their values in space of indicators while experimental data processing, then structural properties of the model

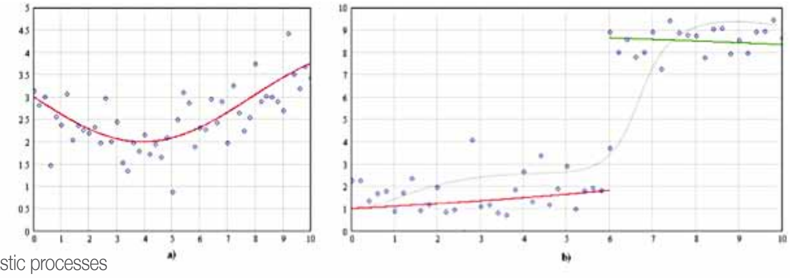


Figure 7. Examples of stochastic processes

change with time and stochastic properties dynamics changes also (Fig.7b).

Initial data of stochastic processes, which was used for creating models, are represented with time series. The received model is a basis for process changes monitoring and forecasting of its dynamics as the model can calculate indicators values for future time periods.

For example, general tendency of parameter's statistical observation states a linear character of changes. Though, detail study of the process based on creating density probability function of distribution reveals that there are essential divergences in simulating functions for different periods of observations, which can indicate stochastic process structural properties changing.

Thus, at moment t6 a change of average value of observed simulation error sample occurs and forecast based on all the data from the sample can lead to error.

COCLUSIONS

Thus, the examined methods of statistical data analysis represented with time series are able to determine distribution of general sample using only limited data set of observations. Methods also can be used for comparison of two sample distributions, forecasting stochastic process, data set extractions, which can describe stochastic process as composition of homogeneous fragments. The proposed approach gives possibility to realize structural time series analysis, which includes extracting homogeneous fragments on experimental curve, and which help to avoid error in making forecast. The method of Data Rationing can be effectively used in those fields of engineers and technology processing of signals. The monitoring and control system of designs of objects makes possible to shift from analysis of particular threats to systemic and interdisciplinary viewpoint on emergencies. As a result it should foster abilities to forecast and prevent emergency situations. The monitoring system of designs of objects is one of the instruments needed to accomplish this goal.

BIBLIOGRAFIA

1. Efron B. Nontraditional methods of multidimensional statistical analysis. – Collection of articles. – Moscow, Finance and Statistics, 1988.

2. Principal components of Data Series: Caterpillar method. Under supervision of Dalilov D.L., Zhigliavsky A.A., S-Peterburg University, 1997.

3. Urazbahtin I.G., Urazbahtin A.I.. Properties of distributions of casual be-masks, Set in the limited interval, Telecommunications, 2005, № 5, c.5-9.

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.

Cooperante en Investigación con el área de Posgrados de la UTE.