Občianske združenie VZDELÁVANIE–VEDA–VÝSKUM
Civil Association EDUCATION–SCIENCE-RESEARCH
Некоммерческая организация ОБРАЗОВАНИЕ-НАУКА-ИССЛЕДОВАНИЕ
Andrusovova 5, 851 01 Bratislava, Slovakia
www.v-v-v.sk

Príspevky v časopise sú recenzované, neprechádzajú jazykovou redakciou.
Contributions in the journal have been reviewed but not edited.

Názov časopisu (Journal Title)
Aplikácie informačných technológií
Information Technology Applications

Šéfredaktor (Editor in chief)
doc. Ing. Martin Šperka, PhD. Faculty of Informatics, Paneuropean University in Bratislava

Výkonný redaktor (Executive editor)
Ing. Michal Grell, PhD. Civil Association EDUCATION–SCIENCE-RESEARCH in Bratislava

VORONEZH INSTITUTE OF HIGH TECHNOLOGIES

Information Technology Applications / International Scientific Journal – 1/2015

Print:
Publishihg and Printing Centre "Nauchnaya Kniga"
32e, ul. Srednemoskovskaya, Voronezh, Russia, 394030
www.n-kniga.ru
E-mail: zakaz@n-kniga.ru

Časopis vychádza dvakrát do roka.
The journal is published twice times a year.

Časopis je recenzovaný, číslo 1/ročník IV.
The Magazine is reviewed, volume 1/ IV.

Address of the Editorial Office:

Civil Association
Education-Science-Research
Andrusovova 5
851 01 Bratislava,
Slovakia
vvv.esr@gmail.com

Pan European University, n.o.
Tematínska 10
851 05 Bratislava,
Slovakia
Phone: +421 2 68 203 639
martin.sperka@paneurouni.com

Predplatné (Subscription):
Subscription orders must be sent to the editorial Office.
The price is 20 EUR a year. It is possible to order older issues only until present supplies are exhausted (15 EUR an issue).

Vydáva (Published by)
Pan-European University, Tomášiková 20, 821 02 Bratislava, Slovakia, IČO: 36077429 in collaboration with the Civil Association EDUCATION-SCIENCE-RESEARCH, Andrusovova 5, 851 01 Bratislava, Slovakia, IČO: 42255180

Výroba (Production)
Voronezh Institute of High Technologies

Number of copies: 100 pieces, **ISSN:** 1338-6468 (print version), **EAN** 9771338646000 51

Registration No.: EV 4528/12

(Elektronická verzia časopisu) Electronic version of Journal: http://www.v-v-v.sk

Zadané do tlače (Delivered to the press): February 2015

Содержание / Contents

………………………………………………………………………..

Dear Readers!

This paper of the journal is devoted to the issues of data mining, the use of methods of mathematical modellling, optimization and decision making support in the tasks of management of complex socio-economic and technological objects. The issue mainly includes the articles written by the research scientists of Pan-European University (Bratislava), in collaboration with the scientists of Voronezh Institute of High Technologies (Russia).

The research area of the presented works is related to the solution of a number of current scientific and practical problems, among which are:

- Building of the information database for the purpose of handling the problems of data mining and predictive modelling;
- Building of parametric and semi-parametric quantile-regression models and their use in the data analysis problems;
- Identification of objects with inhomogeneous characteristics;
- Algorithmization of decision-making in management of objects with inhomogeneous characteristics;
- Development and use of the expert virtual resource of corporate intellectual capital in the managerial decision-making environment.

We hope that the published articles will attract the attention of many researchers working in the field of information technology, and will be interesting for the young researchers as well as for the recognized authorities in the field of use of information technology to meet the challenges of data mining, modelling and management in various fields.

O. N. Choporov, D.Sc. Voronezh Institute of High Technologies Voronezh, Russian Federation, took part in the release of of the collection.

President of Voronezh Institute of High Technologies
D. Eng. Sc. Prof. Y. E. Lvovich

Special aspects and principles of identification of objects with inhomogeneous characteristics

M. Sperka

Abstract:

The analysis of the ways of the efficiency improvement of management of technological objects is carried out. It is suggested that there are different inhomogeneity in the objects in the majority of tasks, and from the perspective of the specific characteristics of technological objects, the inhomogeneity measurement is one of the sources to increase management effectiveness. The technology of development of identification and management principles aimed at improvement of efficiency and quality of object functioning based on the analysis of the diversity category in relation to objects with inhomogeneous characteristics is proposed. The scheme of an algorithm of testing of objects for inhomogeneity is developed. For identification of objects with inhomogeneous characteristics the procedure of decomposition of their mathematical description of homogeneous components with the use of current information obtained as a result of the experiment, archival information and expert assessments, is proposed.

Key words:

Principles of identification, cybernetic assesments, optimization and efficiency, inhomogeneous characteristics, technological objects, management, homogeneous components.

ACM Computing Classification System:

Control structures, Development frameworks and environments, Software development techniques.

Introduction

The control systems entails their ever more complex topology, i.e. they become more diverse, which calls for developing more efficient techniques of their design [1].

One of the sources of such diversity in designing control systems, being one of the most important concept in cybernetics, is the diversity of controlled entities. A separate class of controlled entities that feature these differences resulting in their wider diversity is entities with inhomogeneous properties. Even to now, when controlling them differences in their characteristics (their inhomogeneity) are only taken into account partially and obliquely. No system approach based on assessing the effect of such inhomogeneity on the quality of control processes is followed. Respectively, their control algorithms are not too effective. Inhomogeneity of controlled entities of coinciding physical nature and use and its manifestations give birth to the problem of optimal choice, which means offering the best possible option among the many admissible.

When accounting for the inhomogeneity of controlled entities, their objective functions and control actions are formulated for uniform series of entities or – if the level of inhomogeneity is really high – for each such entity separately, to optimize their control, improve the efficiency and quality of their functioning. The task of identification of control of objects is closely related to the problem of their inhomogeneity.

1. The concept of identification and management with inhomogeneous characteristics

The concept of effective high quality control of entities hinges on developing the techniques of optimizing control systems. The task of optimization emerges in the course of planning and controlling production and technological processes, controlling economic systems, identifying entities and systems, while planning experiments, developing and implementing CAD systems, etc. Solutions of the optimization problem in various areas of technology is the topic of numerous studies by both domestic and foreign scientists. Meanwhile advances in optimization techniques remain one of the most important and topical issues in cybernetics.

With upgrades in control systems their topology becomes ever more complex, i.e. diverse, and that results in the need to find ways to improve the efficiency of their design, e.g. using sensitivity functions [1, 2]. Another source of diversity faced when designing control systems, itself being a core concept in cybernetics, is differences between controlled entities. The role and importance of differences and diversity in cybernetics were studied in detail by William Ross Ashby [1] and later in [2]. One of the classes of controlled entities that feature differences resulting in their higher diversity is formed by entities with inhomogeneous properties. The concept of "inhomogeneity" is found in the description of many physical phenomena and processes.

The principles of identifying and controlling entities with inhomogeneous properties have found their development in [3-5]. The task of classifying inhomogeneities that separate classes of entities with inhomogeneous properties has arisen from practical problems of controlling production processes in radio and electric machinery components manufacturing, microelectronics, capital construction design, and healthcare [5, 7]. The task of optimal control has called for accounting the inhomogeneity of the characteristics of entities. Sources of inhomogeneity had to be classified, techniques to identify and assess the level of such inhomogeneity developed, the uniform control problem formulated.

1) develop classifications of sources and techniques to assess the degree of inhomogeneity of entities;

2) design techniques to identify such entities;

3) develop the techniques to simplify rationally mathematical models, separating and describing their homogeneous components to reduce time and scope of identification experiments;

4) develop the techniques to set and solve the problem of multiple criteria optimization and of numerical assessment of qualitative characteristics;

5) develop assessment techniques to provide for efficiency, reliability and quality when control is affected by subjective factors due to man-technology interaction;

6) design algorithmic procedures that take care of optimal distribution of functions between man (DMP) and machine;

7) develop heuristic preparation techniques for decision making, for drafting and adopting decisions that resist formalization;

8) design simulation systems that combine expert assessments with formalized models;

9) streamline algorithmic support for decision making.

10) Consider the inhomogeneity of entities' characteristics and the way it affects control efficiency.

11) Such inhomogeneity precipitates differences in problem setting and in mathematical description of optimal control of technological entities of one and the same physical nature and use. When formulating and solving the new cybernetic problem of principles of identifying and controlling entities with given properties, one has to evaluate the economic efficiency of control systems constructed on the basis of those principles [8].

For such valuation we analyze possible approaches to the task of control in the following two situations:

1) principles of identifying and controlling entities with inhomogeneous properties are absent;

2) principles of identifying and controlling entities with inhomogeneous properties are fully developed.

The first situation corresponds to hypothesis H1: the technological entity has an integrated mathematical description and a unique setting of the problem of its optimal control for every mode it may function in. In that case one may recourse to single-time identification: collecting the information terminates at a prescribed moment, and a unique entity model is constructed once the experiments are finished, the entity inhomogeneity disregarded completely [59]. The structure of control algorithm is predefined by the setting of optimization problem and its parameters, i.e. by the structure and parameters of its mathematical description.

That hypothesis disagrees with the actual features of technological entities. Following such a description results in poorer production quality and entails extra expenditures on restructuring the control system. Attempts are taken to reduce economic losses due to inadequate mathematical description and algorithmic support of optimal control. Systems based on relatively "rough" models are designed and launched, to be gradually tuned while control quality is improved due to continuous replenishment of experimental data [10, 11]. Such an approach is justified in case the entity parameters change continuously, e.g. they drift while the respective characteristics remain unknown. In case the entity is inhomogeneous, its mathematical model – built for certain conditions shall eventually differ in either its structure or discrete values of its parameters from actuality. Restructuring such models and algorithms requires accumulating new experimental data. The respective period is aggravated with economic losses, since the probability of having adequate products diminishes at that time. Even after the model tuning is over it may remain far from the optimal level of quality. We call such losses "the losses of first type".

From the point of view of control, the features making our entities different are significant in case they affect formation of the objective function and the quality of processes going in the entity. Hereunder we shall call differences in several significant features (characteristics) between controlled entities their inhomogeneity. Inhomogeneity is a category related to the diversity of entities of identical physical nature and use.

There exist well known and developed techniques of identification and control of entities with uniform characteristics [9, 12, 13]. However inhomogeneity in their characteristics is also a feature of well-tuned technological processes. In case inhomogeneity is insignificant, one may use regression analysis technique [14] to retrieve the characteristics of the respective technological process. If, on the other hand, the inhomogeneity of the managed entity is significant, its statistical samples become inhomogeneous and it becomes impossible to use probabilistic techniques. Besides, mathematical description of such entities becomes hardly adequate (if possible at all) in case one does not account for the significant inhomogeneity of their characteristics, and quite complex and poor in its performance.

Meanwhile it is known that information from the DMP may be brought in to reduce ambiguity in selecting the objective function, hence to progress to adaptive techniques [7].

One needs a clear description of the scope of problems solved by the DMP when identifying and choosing the tactics to run the process going on in an entity with inhomogeneous properties. Issues should be reviewed that the DMP faces when interacting in dialogue mode with digital and analog computers in the course of decision making. These issues of DMP functions have only found their reflection in some publications [15, 16]. Commonly most entities with inhomogeneous properties do not accept any special impacts that might be used to identify such entities and search for their optimal control. These entities have to be tested either in their normal operational mode or via computer simulations [15, 17].

Accounting for the inhomogeneity of entity characteristics in either partial or oblique way improves control efficiency somewhat. It may be considered "rough" tuning. To achieve "fine" tuning in controlling such entities one needs to introduce certain criteria and account for the significant inhomogeneity of their characteristics.

When identifying such entities one needs to consider their specifics in dependence of control tasks. Problems of such identification in the conditions of inhomogeneity is the topic of monograph [3]. However it deals with one task only: reducing the error margin when planning experiments. Inhomogeneity sources are classified there with respect to entity nature only, disregarding the physical nature of processes or the degree of inhomogeneity manifestations. Inhomogeneity sources are divided into discrete and continuous. First, variability is assessed of raw materials, equipment, performers, i.e. inhomogeneity descriptors belonging to qualitative input variables of the process. Moreover, the level of such qualitative descriptors is not considered at all, as having no role in the issue [3]. Inhomogeneity source of continuous type is the drift of some entity output descriptor with time or some other coordinate (change in the catalyst activity, equipment aging, variations in the composition of raw materials, etc.). All the inhomogeneities are demonstrated to be of quantitative and qualitative nature [3]. The authors of [4] undertook a theoretical study of α,τ-uniform processes, i.e. they only considered certain uniform components of discrete dynamic processes without analyzing other sources of inhomogeneity. Studies are practically absent that would offer exhaustive classification of inhomogeneities of controlled entities, quantitative assessment of their degree, account for them in the course of forecasting and control, or describe the techniques of selecting and describing uniform components.

We shall consider a managed entity, fully identifiable in case one can pinpoint its inhomogeneous properties, find a mathematical description linking that entity inputs and outputs, and formulate a single optimization criterion for it. In actual conditions one may only expect some – larger or smaller – degree of identifiability. This degree of identifiability for the class of entities with inhomogeneous properties consists in assessing the level of identification and description of their uniform components for each type of inhomogeneities.

The analysis of a priori information and Substantive description of entities with inhomogeneous properties [1, 18-21] has demonstrated that alongside their inhomogeneous properties entities of that class feature [4]:

1. Stochastic character of processes run in them.
2. Constraints of varying level on active experiments.
3. Control aggravated by incomplete a priori information and ambiguities.
4. Numerous input and output variables.
5. Lack of exact quantitative relationships between their input and output variables.
6. Dynamic nature of processes run in them.

The characteristics of controlled entities being inhomogeneous, such features affect the selection of techniques chosen for their identification; in many cases special approaches have to be developed to separate uniform components and shorten the time needed for such identification.

Identifying entities with inhomogeneous properties makes it possible to describe uniform components mathematically in the simplest way via their significant inhomogeneities. It yields the basic economic effect at the initial stage of developing systems to manage complex technological processes, improves the efficiency of mathematical models that describe the processes adequately and are quite capable to solve the optimization problem [1, 2]. In its turn, simplifying the models results in simpler structure of systems to manage technological processes.

When accounting for inhomogeneities, the objective function and control actions are formed for separate uniform series of entities or for individual entities in case their inhomogeneities are high. Thus formed, they help to run the processes optimally and improve control efficiency.

It is also advisable to take inhomogeneity of the controlled entities into account when these entities differ significantly from each other. It affects the choice of control techniques and the decision making process that boils down to the type and level of control actions. As of now, there are no theoretically proven techniques available to assess the inhomogeneity of entities by their aggregate features; therefore the problem arises of quantitative assessment of the degree and level of inhomogeneity. The number of mathematical models of uniform components grows for higher inhomogeneity levels, so that the scope, time and cost of experiments needed to identify entities with inhomogeneous properties grow too. Hence one faces the task of shortening identification time, i.e. of developing mathematical instruments to accelerate identification of uniform components. That would simplify mathematical description and increase both accuracy and efficiency of the respective adequate models, economically being more feasible from the control point of view: simpler efficient adequate models require less computer resources for decision making [23].

2. Specifics of identifying entities with inhomogeneous properties

Identifying entities with inhomogeneous properties should be preceded by the analysis of inhomogeneity of the sub-group of entities $\Omega_v (v = \overline{1, q})$ according to the variance of their most informative descriptors $Q_j (j = \overline{1, m})$ and their entropy based on observational results. In case these inhomogeneity prerequisites are met, one proceeds to identify the characteristics of entities, i.e. decomposes their mathematical description into uniform components (see definitions 4, 5 [23]).

Following [88] and the classification of inhomogeneity sources for entities, the number of models describing the population of uniform components is given by the expression [23]:

$$C_{yi} = ls \sum_{j-1}^{s} \left(m_{ja} \sum_{i=1}^{mq} P_i \omega_i e_i \right), \tag{1}$$

$i = \overline{1, m}, \qquad j = \overline{1, s},$

where l is the number of series of entities in sub-group $\Omega_v (v = \overline{1, q})$;

S is the number of stages in the process run in the entity;

m_{jq} is the number of output variables, y_i characterizing the j-th stage of the process for q-th series of entities;

p_i is the number of control ranges for the i-th output variable;

ω_i is the number of inhomogeneity types of uncontrolled qualitative input variables for the i-th output variable;

e_i is the number of quality levels of uncontrolled input variables for the i-th output variable;

m_q is the number of output variables, $y_i (i = \overline{1, m})$ for the q-th series of entities.

Uncontrolled input variables include the parameters of entity design and state, of raw and input materials for the process run, and descriptors of technological environment that cannot be controlled due to technical and economic reasons.

Physical and mathematical models are used to identify design inhomogeneities. It is impossible to employ mathematical models during the initial stage of the study because of a lack of respective analytic and experimental data. Applying physical models is limited by the very complexity of entities so they are idealized in a way. It is feasible to take a series of direct observations on the normally functioning entity during the first stage of analysis to study the inhomogeneity of its physical fields.

Depending on the task of studying the sources of design inhomogeneity and the availability of *a priori* information we will use the following approaches to identification:

1) conduct a series of direct observations of the industrial entity in its normal operational mode and process the data obtained using variational analysis techniques;

2) study preliminarily the entity in its normal operational mode modeling its physical fields on an analog or physical installation.

Production inhomogeneities stem from variable quality of raw and other materials, from results of previous treatment and level of production. These inhomogeneities are affected by random descriptors and are of stochastic character.

In case production inhomogeneities and uncontrolled input variables affect the output variables of processes run in the entity strongly, one needs to identify the parameters of vectors describing the inhomogeneity of uncontrolled input variables for mathematical description of uniform components. Accounting for design and production inhomogeneities, the model of uniform components for the *i*-th output variable belonging to the *q*-th series of entities has the form [1]:

$$y_{iq} = f(X, U, Z, V, W), \tag{2}$$

where $X=\{x_1...,x_d\}$ is the vector of significant controlled input variables;

$U=\{u_i,...,u_n\}$ is the vector of control actions;

$V=\{v_i,...,v_e\}$ is the vector characterizing production inhomogeneity in the quantitative uncontrolled input variables;

$W=\{\omega_1,..\omega_f\}$ is the vector characterizing production inhomogeneity in uncontrolled qualitative input variables; -*is*-

$Z=\{z_1,...,z_p\}$ is the vector characterizing design inhomogeneity.

To construct the model of uniform components for the *q-th* series of entities, one has to introduce quantitative descriptors of the levels of vectors *Z* and W into equation (2) or separate parameters $z_1,\ldots,z_p$ into uniform groups.

To identify static characteristics under the conditions of inhomogeneity, one uses current information obtained from experiments, archive information and expert assessments [24-26]. In dependence of means used to accumulate statistical data to identify static characteristics, one may differ between the active and passive experiments.

Following the split of sub-group $\Omega_v(v = \overline{1, q})$ into series of entities $\Omega_y(v = \overline{1, l})$ that are significantly inhomogeneous in their descriptors Q_j, a set of significantly controlled input and most informative output variables is formed for each *q-th* series of entities. Depending on the possibility to stage active experiment, one chooses the technique to identify static characteristics and designs and implements the experiment. Initially it disregards the inhomogeneity of uncontrolled qualitative input production variables. Experiment results are then used to assess statistical significance of the regression coefficient and the adequacy of *i-th* model $(i = \overline{1, m})$ for the *q-th* series of entities; the respective multiple correlation coefficient is calculated using either experimental data or archive material. If the model describes adequately the process run in the entities of the separate *q-th* series, the respective multiple correlation coefficient $R_M < 0.80$, the *i-th* model thus retrieved is considered to be poorly efficient [24]. Then the variational analysis technique [27] is used to estimate the effect of inhomogeneity in uncontrolled qualitative input production variables and these are accounted for when constructing the following series of models for entities of the *q-th* series in sub-group

$\Omega_v(v=\overline{1,q})$ Finally one defines which other types of inhomogeneities in process characteristics have to be taken into account to identify sub-group entities, separates uniform components for each inhomogeneity type and forms the *a priori* model structure for the series $\Omega_y(v=\overline{1,l})$.

In case uncontrolled input variables may be estimated quantitatively, identification techniques based on the current or archive information are used to describe uniform components. The plan of active experiments ***E*** may be implemented in case one manages to select the values from among the population of these variables that correspond to prescribed variability levels $\Delta V_{jk}(k=\overline{1,N}, j=\overline{1,n}$, where ***N*** is the number of experiments in plan *E*, and *n* is the number of input variables.

A separate group of uncontrolled quantitative input variables is formed by individual parameters of entity that remain constant or change with time. To account for their inhomogeneity in the model describing uniform component for the *q-th* series of entities, quantitative descriptors are introduced or the model of uniform component is constructed separately for each individual entity in the course of control [4]. Sometimes the latter appears difficult because of lack of statistical material, while setting active experiments on the entity remains impossible [4]. It is more feasible to separate and describe uniform components at the level of individual parameters, when a variable is introduced into the model of uniform components that characterizes individual features of the population of entities in the *q-th* series with respect to a certain descriptor *Qj* [3]. Then the mathematical model is constructed implementing the plan of a passive experiment [3].

The issue of reducing experiment duration becomes quite important for identifying entities with inhomogeneous properties when their static characteristics have to be identified for each uniform component separately, their number increasing for higher degree of inhomogeneity. The number of uniform components growing, the scope of experiment needed to identify them also grows, and that affects the time spent on designing control systems for entities with inhomogeneous properties. There arises the problem of designing techniques for accelerated identification of uniform components [4].

In many papers the practical applications for the study of objects with heterogeneous characteristics are discussed [28-35].

Conclusion

The proposed approaches and principles can be used in order to solve a wide range of tasks of identification and control of technological processes with varying inhomogeneity in production and technological processes, in control of economic systems, in identification of entities and systems, in planning experiments and CAD systems developing and implementing etc.

The concept of an effective high quality control of entities hinges on the development of the techniques of optimization of control systems. The task of optimization emerges in the course of planning and controlling. The solution of the optimization problem in various areas of technology is the subject of numerous studies by both domestic and foreign scientists. Meanwhile advances in optimization of the techniques remain one of the most important and topical issues in cybernetics.

References

[1] Shermergor, T.D. Elasticity Theory for Microinhomogeneous Media, M., 1977.

[2] Eykhoff, P. Foundations of Identification and Control Systems. M., 1975.

[3] Frolov, V.N. The Techniques of Constructing a Model of Processes with Inhomogeneous Properties. – In: Proc. Voronezh Polytech. Inst.: "Mathematical and Technical Issues of Medical Cybernetics", Sci. Ed. Prof. Bykhovsky, M.L., Voronezh, 1978.

[4] Frolov, V.N. Controlling the Treatment of Chronic Diseases. – Abstracts. VII All-Union Conference on Problems of Control, Vol. 2. M.-Minsk, 1977.

[5] Frolov, V.N. Problems of Optimizing the Treatment of Chronic Diseases. – In: Proc. Voronezh Polytech. Inst.: "Automation and Computer Technology". Voronezh, 1977.

[6] Balakirev, V.S., Dudnikov, E.G., Tsyrlin, A.M. Experimental Retrieval of Dynamic Characteristics of Controlled Industrial Entities. M., 1967.

[7] Frolov, V.N. Stochastic Approach to Treating Chronic Diseases. – In: Proc. Voronezh Polytech. Inst.: "Automation and Computer Technology". Voronezh, 1977.

[8] Barannikov, N.I., Lvovich, Ya.E., Raikhel, N.L., Frolov, V.N. Mathematical Leakage Current Model of Forming High Voltage Anodic Aluminum Foil. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[9] Polyak, B.T. Convergence of Iterative Stochastic Algorithms and their Convergence Rate. I. General Case. "Automatics and Telemechanics", 1976, No. 12.

[10] Barannikov, N.I., Lvovich, Ya.E., Raikhel, N.L., Frolov, V.N. Mathematical Leakage Current Model of Forming High Voltage Anodic Aluminum Foil. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[11] Batishchev, D.I. On Experimental Comparison of Some Techniques to Search for Functions of Many Variables. – In: "Extremum Search". Tomsk, 1969.

[12] Frolov, V.N., Lvovich, Ya.E. Principles of Identification and Control of Entities with Inhomogeneous Properties: Monograph. Voronezh: PPC "Science Book", 2010.

[13] Vavilov, A.A., Imaev, D.Kh., Poshekhovov, L.B. Some Issues in Synthesizing Complex Technological Control Systems Using Sensitivity Functions. – Proc. All-Union Workshop School "Sensitivity, Optimization, Problem Solution". Voronezh, 1978.

[14] Borodyuk, V.P., Letsky, E.K. Statistical Description of Industrial Entities. M., 1971.

[15] Modeling Installation for Simulating the Selection of Optimal Tactics for Treating Chronic Diseases. Information Leaflet No. 327-76. Compiled by: Apalkov, V.A., Bala, Yu.M., Stolpovskaya, L.N., Frolov, V.N. Voronezh Inter-Industry Center for Sci.-Tech. Information and Propaganda. Voronezh, 1976.

[16] Zakheim, L.N., Electrolytic Capacitors. M., 1963.

[17] Technical Guide (Rank Correlation), Issue 3. Compiled by: Dyakova, I.S., Krug, G.K. Spec. Automatics Design Bureau, Moscow.

[18] Vorobyev, N.N., Kleinfeld, D.S. A Study of the Initial Collector Current Drift in Planar Silicon Transistor. – Electronic Technology, Series 2. Semiconductor Instruments. M., 1969.

[19] Keizman, V.B., Frolov, V.N. Analysis of the Process of Synthesis of Lithium Ferrites with Rank Correlation Techniques. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[20] Nalimov, V.V., Chernova, N.A. Statistical Techniques for Planning Extreme Experiments. M., 1965.

[21] Polyak, B.T., Tsypkin, Ya.Z. Potential Capabilities of Adaptation Algorithms. – "Issues of Cybernetics. Adaptive systems". Cybernetics Scientific Council, USSR Ac. Sci.. M., 1976.

[22] Rastrigin, L.A., Markov, V.A. Cybernetic Models and Cognition. Riga, 1976.

[23] Frolov, V.N., Chilyakov, A.S. On the Issue of Optimal Composition and Temperature of FL-98 Impregnation Lacquer. – In: Proc. Voronezh Polytech. Inst.: "Electric Drive and Automation of Industrial Installations". Voronezh, 1973.

[24] Bala, Yu.M., Frolov, V.N. Some Issues of Algorithmic Treatment of Therapeutic Diseases. – In: Proc. Voronezh Polytech. Inst.: "Mathematical and Technical Issues of Medical Cybernetics". Voronezh, 1977.

[25] Kaplinsky, A.I., Krasnenker, A.S. On Formulating Stochastic Optimization Algorithms. – Preprint. Inst. Industrial Economics, Ac. Sci. Ukraine, Donetsk, 1974.

[26] Lvovich, Ya.E., Raikhel, N.L., Frolov, V.N. Mathematical Models of Foil Treatment in Industrial Electrochemical Etching Units. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[27] Lvovich, Ya.E., Stupachenko, A.A., Fomin, K.B., Frolov, V.N., Shishkov, B.A. Modeling in Problems of Studying Optimization of Complex Processes. Teaching Aid. Voronezh, 1974.

[28] Preobrazhensky Y. P. Evaluation of the effectiveness of the system of intelligent decision support / Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 116-119.

[29] Zyablov. E. L. Creating object-semantic model of the control system / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 029-030.

[30] Panevin R. Y. optimal control of a multi-stage technological processes / R. Y. Panevin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 77-80.

[31] Zazulin A. V. Peculiarities of building a semantic domain models / A. V. Zazulin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 026-028.

[32] Zyablov. E. L. Development of the linguistic means of intellectual support simulation-based-semantic modeling / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 024-026.

[33] Lvovich Ya. E. Adaptive control of Markov processes in a conflict situation / Ya. E. Lvovich, Y. P. Preobrazhensky, R. Y. Panevin // Herald of the Voronezh state technical University. 2008. Vol. 4. No. 11. P. 170-171.

[34] Zyablov. E. L. Markov decision processes of the first type with multiple absorbing States / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 68-71.

[35] Lvovich Ya. E. Decision making in expert-virtual environment / Ya.E.Lvovich, I.Ya.Lvovich // Voronezh: Publishing house "Science book", 2010, p. 139.

. . . .

Doc. Ing. Martin Sperka, PhD.
Paneuropean University, Bratislava, Slovakia

Algorithmization of decision-making in management of objects with inhomogeneous characteristics

F. Schindler

Abstract:

The question of algorithmization of decision-making in identification and management of objects with inhomogenious characteristics is considered. The analysis of features of objects with inhomogenious characteristics is carried out. The decision-making basis in management of objects with inhomogenious characteristics is defined. The necessity for the development of human-machine procedures of decision-making in management of objects with inhomogenious characteristics is proved. General principles of management of objects with inhomogenious characteristics are presented. Inhomogeneity of management results in a variety of variants of management, each of which is characterized by the set of indicators. The three approaches for turning of indicators which depend on the information sources are offered. They are aprioristic, dialogue and adaptive. The classification signs which are the basis for a choice of the optimum decision in management of objects with inhomogenious characteristics are offered.

Key words:

Decision making, management, objects with heterogeneous characteristics, optimal solution.

ACM Computing Classification System:

Control structures, Development frameworks and environments, Software development techniques.

Introduction

The development of techniques of optimization of control and decision-making systems is closely connected with the concepts of effectiveness and high quality control of objects. The problem of optimization and decision-making emerges in planning and controlling of production and technological processes, in management of economic systems, in identification of objects and systems, in planning of experiments, in CAD systems development and implementation, etc. The solution of the optimization problem in various areas of technology is the subject of numerous scientific papers. Meanwhile the development of optimization techniques is one of the most important and topical issues in cybernetics.

1. The particularities of controlling entities with inhomogeneous properties

With upgrades in control systems their topology becomes ever more complex, i.e. diverse, and that results in the need to find ways to improve the efficiency of their design, e.g. using sensitivity functions [1, 2]. Another source of diversity faced when designing control systems, itself being a core concept in cybernetics, is differences between controlled entities. The role and importance of differences and diversity in cybernetics were studied in detail by William Ross Ashby [1] and later in [2]. One of the classes of controlled entities that feature differences resulting in their higher diversity is formed by entities with inhomogeneous properties. The concept of "inhomogeneity" is found in the description of many physical phenomena and processes.

In many cases one is forced to solve the task of optimization and decision-making while *a priori* information remains incomplete, stochastic characteristics of controlled entities are only known partially and various ambiguities are present. That is why adaptive control systems find more and more use [3, 4]. While *a priori* information remains incomplete and disturbance is present, optimization techniques develop towards seeking for "the best" optimization algorithms [5, 6], developing a single approach to analyze and synthesize optimization algorithms [7, 8], and build a "man-machine" interaction dialogue in automated systems that manage technological processes, and working on computer-aided design [9-11]. Simulation techniques [12] find more and more use for complex entities and systems.

The task of optimal control has called for accounting the inhomogeneity of the characteristics of entities. Sources of inhomogeneity had to be classified, techniques to identify and assess the level of such inhomogeneity developed, the uniform control problem formulated.

Consider the inhomogeneity of entities' characteristics and the way it affects control efficiency.

Such inhomogeneity precipitates differences in problem setting and in mathematical description of optimal control of technological entities of one and the same physical nature and use. When formulating and solving the new cybernetic problem of principles of identifying and controlling entities with given properties, one has to evaluate the economic efficiency of control systems constructed on the basis of those principles [5]. For such valuation we analyze possible approaches to the task of control in the following two situations:

1) principles of identifying and controlling entities with inhomogeneous properties are absent;

2) principles of identifying and controlling entities with inhomogeneous properties are fully developed.

That hypothesis disagrees with the actual features of technological entities. Following such a description results in poorer production quality and entails extra expenditures on restructuring the control system. Attempts are taken to reduce economic losses due to inadequate mathematical description and algorithmic support of optimal control. Systems based on relatively "rough" models are designed and launched, to be gradually tuned while control quality is improved due to continuous replenishment of experimental data [13, 14]. Such an approach is justified in case the entity parameters change continuously, e.g. they drift while the respective characteristics remain unknown. In case the entity is inhomogeneous, its mathematical model – built for certain conditions – shall eventually differ in either its structure or discrete values of its parameters from actuality. Restructuring such models and algorithms requires accumulating new experimental data. The respective period is aggravated with economic losses, since the probability of having adequate products diminishes at that time. Even after the model tuning is over it may remain far from the optimal level of quality. We call such losses "the losses of first type".

The second situation corresponds to hypothesis H_2: the technological entity features inhomogeneous properties and there are diverse mathematical descriptions and settings of the problem of optimal control available for various conditions of functioning of that entity. In that case we have extra economic expenditures to identify the models for all the uniform components and set respective optimization problems. Also, larger computer storage space is needed. We shall call these losses "the losses of second type". Losses of first type are absent then, since we have optimization models constructed and control algorithms defined in advance for each set of conditions characterizing the inhomogeneity.

From the point of view of control, the features making our entities different are significant in case they affect formation of the objective function and the quality of processes going in the entity. Hereunder we shall call differences in several significant features (characteristics) between controlled entities their inhomogeneity. Inhomogeneity is a category related to the diversity of entities of identical physical nature and use.

Despite the existence of wide class of entities with inhomogeneous properties, they are commonly managed using classical techniques. Moreover, only some selected sources of inhomogeneity are considered, and no systemic approach is pursued. The latter should be based on assessing the level of inhomogeneity and its effect on the quality and efficiency of processes occurring in such entities. It is attempted instead to improve control efficiency via automatic control and stabilization of the number of controlled entry variables without any account of their own inhomogeneities and inhomogeneity of uncontrolled entry variables. In other cases the best option is chosen from those already available [15] or the level of control actions is calculated using the data from preceding steps of control and the weighted average control that describes the changing trajectory of output variables. However, when *a priori* information remains incomplete, the algorithms proposed to manage the process with inhomogeneous properties disregard the information from the decision making person (DMP) at each consecutive step of control [16].

Inhomogeneity of certain features and their manifestations define the way the control task has to be set. Together with differing sets of output descriptors and differing forms of objective function it all results in setting the problem of optimal selection: the best option has to be chosen among the many admissible, given the ambiguities. Publications are available [17] that interpret the problem of optimal selection as searching for an optimal solution on a discrete set. However, the approach outlined in it focuses on developing respective algorithms while offering no methodological foundations for the problem. The principles of identifying and controlling entities with inhomogeneous properties have found their development in [16, 88].

It is particularly the practical problems of "fine" control that have called for accounting the inhomogeneity of entity characteristics, for developing a classification of inhomogeneity sources from the point of view of control needs, ways to assess the degree of such inhomogeneities, the degree of identifiability of entities, designing algorithms to choose identification techniques that are still lacking from any publications.

Eventually one has to solve the problem of decision making and selecting control actions to run the process going on in an entity in the most efficient way. Therefore, the control system has to be built as an "automated system of decision making" (ASDM). To control entities with inhomogeneous properties, the ASDM structure is defined by the features typical for that class of controlled entities, various sources of information, and possibilities and conditions for its active search.

Moreover, inhomogeneous technological processes have to be managed on the basis of incomplete information while ambiguities remain numerous and sources of information are scanty. When controlling such processes decision making is driven by deterministic, even random preferences of the DMP and these shape man-machine procedures that are followed. That is why still another problem arises: to formalize decision making procedures and develop algorithms to select the current control objectives and levels of control actions on the basis of online information coming from the DMP. Most inhomogeneous technological processes preclude any active search for optimal control options among the set of those admissible, or remain limited in possibilities to do so and that makes computer simulations necessary. The task consists in developing methodological foundations for such simulations, staging them for various situations, selecting control algorithms and their parameters and forecasting control outcome. Since the final decision consists in choosing the values of control actions, man-machine procedures have to be implemented using ASDMs. In that case one needs to combine rationally the functions of DMP with digital and analog computers to improve the efficiency and reliability of decision making and quality of running inhomogeneous processes.

2. General principles and methods of decision-making during control of entities with inhomogeneous properties

Inhomogeneity of entities in certain descriptors and its manifestations define the setting of the problem of control, and the differing set of output descriptors and result in varying form of the objective function (Definition 1, [18]). It all results in the problem of optimal selection related to defining the best option among the many admissible [16].

If one reaches the level of mathematical description while identifying the characteristics of the entity, commonly a uniform optimization model is developed to control that entity, and that model is used to select the optimal solution $U^*=\{U_1^*,...,U^*_K\}$, where U^* is the optimal vector of control actions. The optimization model is understood to be the mathematical formulation of the objective function and its constraints on the basis of static and dynamic models of the entity.

In case the control problem is non-uniform, selecting the tactics to control the entity has to follow several optimization models. The DMP is brought in for selection, and he/she prefers this or that option on the basis of subjective information. This option may be updated using the current information obtained in the entity normal operational mode or remain unchanged.

Heterogeneity stemming from diverse options of control problem manifests itself as the need to account for various sets from the total population of entity descriptors and as the differing significance of descriptors when selecting optimal controls for separate sub-groups, series, and separate entities with inhomogeneous properties. Heterogeneity of controls generates diverse control options, each of them characterized by its own set of descriptors. The procedure of selecting optimal control includes forming a set of descriptors, defining their weights $\alpha_i\{i=1,...,m)$ in the course of ranging them, convoluting the descriptors into a global criterion and selecting the type and level of control action to reach a preset control goal.

In dependence of sources of information three approaches may be used to convolute descriptors:

a) the *a priori*, according to estimated descriptor weights obtained in advance of searching for optimal solution [19];

b) the dialogue, making it possible to construct man-machine procedures for assessing descriptor weights from the current results of searching for optimal solution [3, 18];

c) the adaptive, making it possible to tune the probability of controlling this or that descriptor by the current information and information coming from the DMP [20].

In the result of identifying controlled entities with inhomogeneous properties one separates uniform components, each of them characterized by its own set of descriptors. If that set is determined, their weights $\alpha_i\{i=1,...,m)$ are set *a priori* of control, and their values and relations remain constant. Descriptors are ranged by their significance via sequential solution of the extremum problem: problem solution $U_i^*=\{U_1^*,...,U_n^*\}$ belonging to the Pareto set is found for the area S_i [19]. Eventually, searching for the optimum on the consecutive descriptor in the ranging series results in the area S_i degenerating into a point. That is why it is recommended [19] to limit this search to a compromise that only differs from the optimal U_i^* within acceptable limits, using the concessions technique.

In case the experts that take part in ranging descriptors disagree on concessions, one proceeds to synthesize a global criterion as a function of the initial descriptors [19]. The solution of multi-criteria problem is reduced to an unconventional optimization [19]:

$$W(Q) = W(Q_1,Q_2,...,Q_m) \rightarrow \min_{\bar{u} \in S}. \quad (1)$$

Accounting for the requirements on the global criterion [21], the latter may be presented as:

$$W(Q) = \sum_{i=1} \alpha_i \frac{Q_i - Q}{Q_i^{**} - Q^*} = \sum \alpha_i \tilde{Q}_i(u), \quad (2)$$

where Q_i^{**} is the minimum value of i-th descriptor in the area S;

Q_i^* is the maximum value of i-th descriptor in the area S;

α_i is the weight of i-th descriptor obtained via expert assessment with its further ranging [21].

$$\alpha_i = \frac{\alpha_i}{\sum_{i=1}^{m} \hat{\alpha}_i} \quad (3)$$

he values of α obtained yield constant weights of descriptors of the process and their relations.

When choosing the tactics to run the process, descriptor weights may change, and if the decisions are taken with incomplete *a priori* information and ambiguities in the assessed quality, the decision making person (DMP) is brought in at certain stages of solving the problem of optimal control. In that case the dialogue approach [1] is used to retrieve descriptor weights.

To convolute criteria, a non-parametric procedure is followed in dialogue mode with the DMP, the convolution algorithm proposed in study [1]. The DMP may help to define the set $\bar{a} = \{a_1,..., a_m\}$.

Since entities are heterogenic, each of them or each separate series of entities in the sub-group needs its own set of descriptors to be formed. Meanwhile, descriptors in the set may contradict each other, while their contradiction traits may be different. A leading (core) descriptor has to be distinguished in each group of contradictory descriptors. A set of descriptors $F=\{F_j\}$, $j=1,2,...\kappa$, where κ is the number of contradictory groups is formed from such leading descriptors. Identifying the leading (core) descriptor in the group belongs to DMP functions.

There is a number of techniques available for grouping descriptors. Studies [9, 30] suggest two algorithms for extreme grouping when random values $f_e^2=1(l=1,...,K)$, $A_1,A_2,...,A_K$ are split into groups in the best possible way via minimizing the functionals [22, 23]:

$$J_1 = \sum_{Q_i \in A_1} (Q_i, f_1)^2 + \sum_{Q_i \in A_1} (Q_i, f_2)^2 + \cdots + \sum_{Q_i \in A_K} (Q_i, f_K)^2; \quad (4)$$

$$J_2 = \sum_{Q_i \in A_1} |(Q_i, f_1)| + \sum_{Q_i \in A_1} |(Q_i, f_2)| + \cdots + \sum_{Q_i \in A_K} |Q_i, f_K|; \quad (5)$$

Using these algorithms, crossed groups $A_1...,\ A_\kappa$ are separated from the total population of descriptors $Qi,..,Q_m$. The problem of defining "the best number of groups" may be a task for DMP, while searching for the number of descriptor groups may be an iterative problem. We consider it possible to use the technique of *a priori* ranging [24] to split descriptors into groups preliminarily while executing extreme grouping to define the number of descriptor groups. That method may be applied in case the ranging diagram features an uneven distribution of the sum of ranks. Study [18] suggested an adaptive approach to convoluting descriptors; with it probabilities of involving a given criterion may be tuned using the current information and the information coming to the DMP, while the goal remains indefinite [16].

Formulating the optimization problem does not yield any conclusion on the techniques to solve it. Such techniques are simply decision making algorithms for choosing the best possible option of running a process. Same as in formalized presentation of the optimization problem, to develop its solution technique one needs to account for the specifics of the controlled entity, e.g. the degree of its identifiability and the characteristics of the set of its descriptors.

When the population of control options $V=\{V_j\}$ contains a single V_j with its selection probability $P_{vj=}1$, the optimal control is chosen following DMP's deterministic preferences.

If there is a probability $P_i \neq 0$ of bringing in some optimization problem or selecting an *i*-th control option, while DMP indicates the preferred control option stochastically, the selection of optimal control follows DMP's stochastic preferences.

Selecting controls for entities with inhomogeneous properties depends on such entities' heterogeneity, on the presence of ambiguities and on conditions of active search for optimal control options [25].

Below we introduce classification indicators to encode the choice of techniques of optimal control [25].

Identifiability indicator, *a*. Since entity identification depends on control goals [69], the specifics of identification while controlling inhomogeneous properties consists in selecting uniform components and in constructing their models to improve the efficiency of entities' functioning. We interpret the degree of identifiability [16, 18] as the level of selecting and describing uniform components $y_i = f_i^{r_k}(u)$ for each inhomogeneity type. The level of separation of uniform components depends on the initial information and the possibility to account for sources of inhomogeneity when identifying the characteristics of entities.

Depending on the degree of their identifiability, entities with inhomogeneous properties are separated into those fully and partially identifiable [3, 25].

When selecting optimal control options for partially identifiable entities, *a priori* information remains incomplete and there is a number of ambiguities, so adaptation techniques are needed [3, 25]. To select optimal control options for fully identifiable entities mathematical models are available for every uniform component, the optimization criterion is defined in a unique way, its probability $P_j = 1$, and the constraints are set, so the optimization problem may be formulated as [25]:

$$\left.\begin{array}{c} \Psi(x) \to \max \\ x \in G \\ G = \begin{cases} x \,|\, f_i(x) \le b, i = 1, \dots, m; \\ x_j \ge 0, j = 1, \dots, n. \end{cases} \end{array}\right\} \quad (6)$$

Depending on the form of objective function and constraints on the solution of the problem of selecting optimal tactics to run a fully identifiable process mathematical programming techniques are used [3].

We set the identifiability indicator $a = 1$ for fully identifiable controlled entities and $a = 0$ in the opposite case [3, 16].

Selection stochasticity indicator, *b*. Three basic types of ambiguities are faced when selecting control options [14]:

1) process output variables are subject to random perturbations, their characteristics unknown;

2) functions setting control goals are not single extremum (the desired values of output variables depend on the entities' inhomogeneity). They are defined for a discrete set of possible types of control actions and there is no explicit analytical expression via control actions;

3) there is no criterion to assess the efficiency of running the process in general.

Depending on the degree of identifiability of entities some ambiguities are removed in the course of identifying their inhomogeneous characteristics. Ambiguities of selecting control options depend on individual inhomogeneous characteristics of entities, the inhomogeneity of uncontrolled input variables and non-uniformity of the control problem. Ambiguities are defined using current information, and information from the DMP is brought in next [3, 4].

Ambiguities remain present in the course of selecting optimal control; the choice is assessed stochastically and the criterion for the preferred alternative features some probability. Therefore the stochasticity indicator is set as $b = 1$. In case a single optional control problem is indicated, the stochasticity indicator $b = 0$. In the first case selecting the optimal (efficient) control option follows stochastic preferences of the DMP, in the second it is deterministic [16, 18].

Selection activity indicator, *c*. Not every entity with inhomogeneous properties tolerates active selection of optimal control option. Search duration may be limited due to economic, technological or technical reasons (C = 0). To solve the non-trivial control problem (when a set of alternative control options is available, possibility is lacking for active search, one needs to adjust control parameters often, etc.) it is advisable to use simulations [12]. This simulation approach to the task of designing search algorithms for optimal control depends generally on the following factors:

1) limited possibilities for active experiments;
2) time constraints of decision making for extended processes;
3) random interferences and perturbations due to changes in characteristics that defy quantitative assessment;
4) systematic changes in control actions in the course of control;
5) strong influence of DMP on the process as the decisive element of the control system.

To conduct simulations one needs reasonably simple imitational models describing the studied process adequately [27].

Selecting the technique of optimal control goes according to the values of classification indicators $Z = [a, b, c]$ [3, 16].

The specifics of entities with inhomogeneous properties defines what combination of known techniques and special man-machine procedures should be used (Table 1.1) [16, 18]. Beside the *a, b, c* indicators one needs to account for the indicators of entity type: its dynamism and discreteness [19].

Some applied problems are considered in different papers [28-35].

Conclusion

The offered principles and methods of an optimum choice can be used for the solution of a wide range of problems of rational management of objects with inhomogenious characteristics including productions and technological processes, economic, social and other systems.

References

[1] Shermergor, T.D. Elasticity Theory for Microinhomogeneous Media, M., 1977.
[2] Eykhoff, P. Foundations of Identification and Control Systems. M., 1975.

[3] Frolov, V.N. Problems of Optimizing the Treatment of Chronic Diseases. – In: Proc. Voronezh Polytech. Inst.: "Automation and Computer Technology". Voronezh, 1977.

[4] Ashby, W. Ross. Introduction to Cybernetics. M., 1959.

[5] Barannikov, N.I., Lvovich, Ya.E., Raikhel, N.L., Frolov, V.N. Mathematical Leakage Current Model of Forming High Voltage Anodic Aluminum Foil. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[6] Stragovich, V.G. Adaptive Systems Theory. M., 1976.

[7] Foundations of Silicon Chip Technology, Oxidation, Diffusion, Epitaxy. 1969.

[8] Petrovsky, A.M. Systemic Analysis of Certain Medico-Biological Problems in Treatment Control. "Automatics and Telemechanics", 1974, No. 2.

[9] Davies, J.A. The Migration of Metal and Oxygen during Anodic Film Formation. J. Electrochem. Soc., 1965, V. 112, No. 7.

[10] Ksiezjk Marek, Danek Adam. Studies of Absorption and Elimination of Drugs, Part IV. Computer Evaluation of Pharmacokinetic Parameters of the Two Compartment Open Model, Pol. J. Pharmacol., Pharm, 1975, V. 27, No. 5.

[11] Gaskarov, D.V., Golinkevich, G.A., Mozgalevsky, A.V. Forecasting the Technical State and Reliability of Radioelectronic Instrumentation. M., 1974.

[12] Modeling Installation for Simulating the Selection of Optimal Tactics for Treating Chronic Diseases. Information Leaflet No. 327-76. Compiled by: Apalkov, V.A., Bala, Yu.M., Stolpovskaya, L.N., Frolov, V.N. Voronezh Inter-Industry Center for Sci.-Tech. Information and Propaganda. Voronezh, 1976.

[13] Barannikov, N.I., Lvovich, Ya.E., Raikhel, N.L., Frolov, V.N. Mathematical Leakage Current Model of Forming High Voltage Anodic Aluminum Foil. – In: Proc. Voronezh Polytech. Inst.: "Automatics, Automation of Measurements", Issue 4. Voronezh, 1973.

[14] Batishchev, D.I. On Experimental Comparison of Some Techniques to Search for Functions of Many Variables. – In: "Extremum Search". Tomsk, 1969.

[15] Bykhovsky, M.L. Selection of Optimal Treatment Plan. – In: "Computer Diagnostics and Information Search in Medicine". M., 1969.

[16] Frolov, V.N. The Techniques of Constructing a Model of Processes with Inhomogeneous Properties. – In: Proc. Voronezh Polytech. Inst.: "Mathematical and Technical Issues of Medical Cybernetics", Sci. Ed. Prof. Bykhovsky, M.L., Voronezh, 1978.

[17] Zakheim, L.N., Electrolytic Capacitors. M., 1963.

[18] Frolov, V.N. Controlling the Treatment of Chronic Diseases. – Abstracts. VII All-Union Conference on Problems of Control, Vol. 2. M.-Minsk, 1977.

[19] Polyak, B.T. Convergence of Iterative Stochastic Algorithms and their Convergence Rate. I. General Case. "Automatics and Telemechanics", 1976, No. 12.

[20] Frolov, V.N., Chilyakov, A.S. On the Issue of Optimal Composition and Temperature of FL-98 Impregnation Lacquer. – In: Proc. Voronezh Polytech. Inst.: "Electric Drive and Automation of Industrial Installations". Voronezh, 1973.

[21] Raikhel, N.L., Frolov, V.N. Applying Experiment Planning Techniques to Retrieving the Characteristics of the Process of Forming Anodic Aluminum Foil (1968-71 Studies Review) – In: Proc. Voronezh Polytech. Inst.: "Material Sci.-Tech. Conf. (1972)". Voronezh, 1972.

[22] Bala, Yu.M., Frolov, V.N. Applying Mathematical Techniques to Selection of Tactics of Treating Chronic Diseases. – Materials on Mathematical Software and Use of Computers in Medico-Biological Studies. Obninsk, 1976.

[23] Gaskarov, D.V., Golinkevich, G.A., Mozgalevsky, A.V. Forecasting the Technical State and Reliability of Radioelectronic Instrumentation. M., 1974.

[24] Markova, N.E., Chervinskaya, O.Z., Gribkov, S.P., Fedorova, M.A. Electrochemical Etching of Capacitor Aluminum Foil. – Issues in Radioelelctronics, Series III, No. 5, 1962.

[25] Tsypkin, Ya.Z. Adaptation and Training in Automatic Systems. M., 1968.

[26] Rastrigin, L.A., Markov, V.A. Cybernetic Models and Cognition. Riga, 1976.

[27] Vavilov, A.A., Imaev, D.Kh., Poshekhovov, L.B. Some Issues in Synthesizing Complex Technological Control Systems Using Sensitivity Functions. – Proc. All-Union Workshop School "Sensitivity, Optimization, Problem Solution". Voronezh, 1978.

[28] Preobrazhensky Y. P. Evaluation of the effectiveness of the system of intelligent decision support / Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 116-119.

[29] Zyablov. E. L. Creating object-semantic model of the control system / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 029-030.

[30] Panevin R. Y. optimal control of a multi-stage technological processes / R. Y. Panevin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 77-80.

[31] Zazulin A. V. Peculiarities of building a semantic domain models / A. V. Zazulin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 026-028.

[32] Zyablov. E. L. Development of the linguistic means of intellectual support simulation-based-semantic modeling / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 024-026.

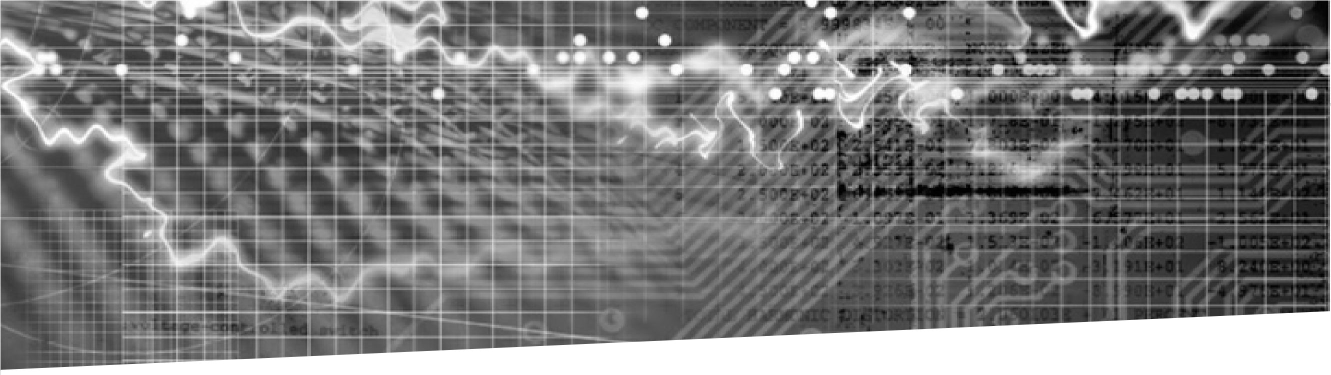
[33] Lvovich Ya. E. Adaptive control of Markov processes in a conflict situation / Ya. E. Lvovich, Y. P. Preobrazhensky, R. Y. Panevin // Herald of the Voronezh state technical University. 2008. Vol. 4. No. 11. P. 170-171.

[34] Zyablov. E. L. Markov decision processes of the first type with multiple absorbing States / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 68-71.

[35] Lvovich Ya. E. Decision making in expert-virtual environment / Ya.E.Lvovich, I.Ya.Lvovich // Voronezh: Publishing house "Science book", 2010, p. 139.

....

Prof. RNDr. Frank Schindler, PhD.
Paneuropean University, Bratislava, Slovakia
frank.schindler@paneurouni.com



Structurization of an expert and virtual resource of the corporate intellectual capital in the environment of adoption of administrative decisions

I. Y. Lvovich

Abstract:

The new approach to the building of systems of managerial decision making support based on the integration of expert evaluation techniques and algorithmic procedures of computer (virtual) components functioning that are viewed as the virtual expert resource of corporate intellectual capital, is considered.
The concept of the two types of virtual expert resource is introduced: the procedural and the knowledge resources. Algorithms of interaction between real and virtual experts for the purpose of generation of prospective managerial decisions via methods of optimization - search and variational modeling, are considered.

Key words:

Expert and virtual resources, corporate intellectual capital, adoption of administrative decisions, expert estimation, optimization, variation modeling.

ACM Computing Classification System:
Control structures, Development frameworks and environments, Software development techniques.

Introduction

The modern stage of the society development is characterized by stronger competition between different manufacturers, globalization of the markets, and introduction of technological novelties. The speed of decision making comes to the foreground, that is the process aggravated by the need to process large streams of inhomogeneous information. Therefore the role of intellectual resources of both separate operating entities and whole industries becomes more important. Intellectual resources stand apart as an independent object of economic and managerial relations and turn into an important strategic resource of every organization that controls its competition and development capabilities [10].

1. The concept of virtual expert resource as a form of intellectual capital information resource

Considering the development of intellectual resource in the corporate information environment one needs to treat separately expert and virtual components in the structure of intellectual capital and multi-alternative presentation of its elements in the course of managerial decision making [2, 3]. By 'decision making' we mean a three-stage procedure that includes analyzing the initial information, preparing to make decision and selecting a decision generated in the course of interaction between the expert (experts) and a computer system. We shall call the combination of expert and computer resources 'the virtual expert resource for decision making' and treat it as a component optimizing the management of corporate social (economic) system.

Effective corporate management is defined by a set of indicators (criteria) that stem from the task set for the functioning system. As the system evolves, it accumulates the information on both itself and the mechanisms of its development, such information presented as expert knowledge (human capital) and computer information resources. Together they form the corporate intellectual capital system within their common information space. Using the algorithm of interaction between the components of intellectual capital we form a new integrated virtual expert resource and develop management procedures to optimize system development, as shown in the structural flowchart presented in fig. 1.

As follows from the structural flowchart, virtual expert resource serves as the basis for choosing managerial decisions. The leading role is played by expert, a decision making person (DMP), designer and manager. However virtual expert resource is oftentimes as valuable as the human resource in shaping the intellectual capital needed for managerial activities.

Despite such an understanding of importance of integrated intellectual resource in the computer component of decision making, it is still attributed an auxiliary role only, as first postulated in the 1970-80's. That point of view finds its concentrated expression in [4]: information processing automats may serve as perfect assistants in decision making, but never more than that. Meanwhile the authors of [5] point to certain weaknesses of experts who are incapable to explain how they arrive at specific decisions and recourse to fuzzy and conclusions difficult to understand. The authors suggest using artificial expert systems to produce an equivalent of human expert. Study [6] indicates a more promising path forward: it is only the human-computer alliance, a well designed system of man-machine procedures (man-machine system, MMS) that may improve reliably the quality of decisions taken.

The choice between the human and the computer resource is usually dependent on whether decisions are taken in formalized vs. non-formalized ways [7]. The first approach means solving highly structured problems with sufficiently clear algorithms, the choice of the solution explained using formal mathematical techniques and computer resources. The second approach relies on DMP cognition, i.e. is informal.

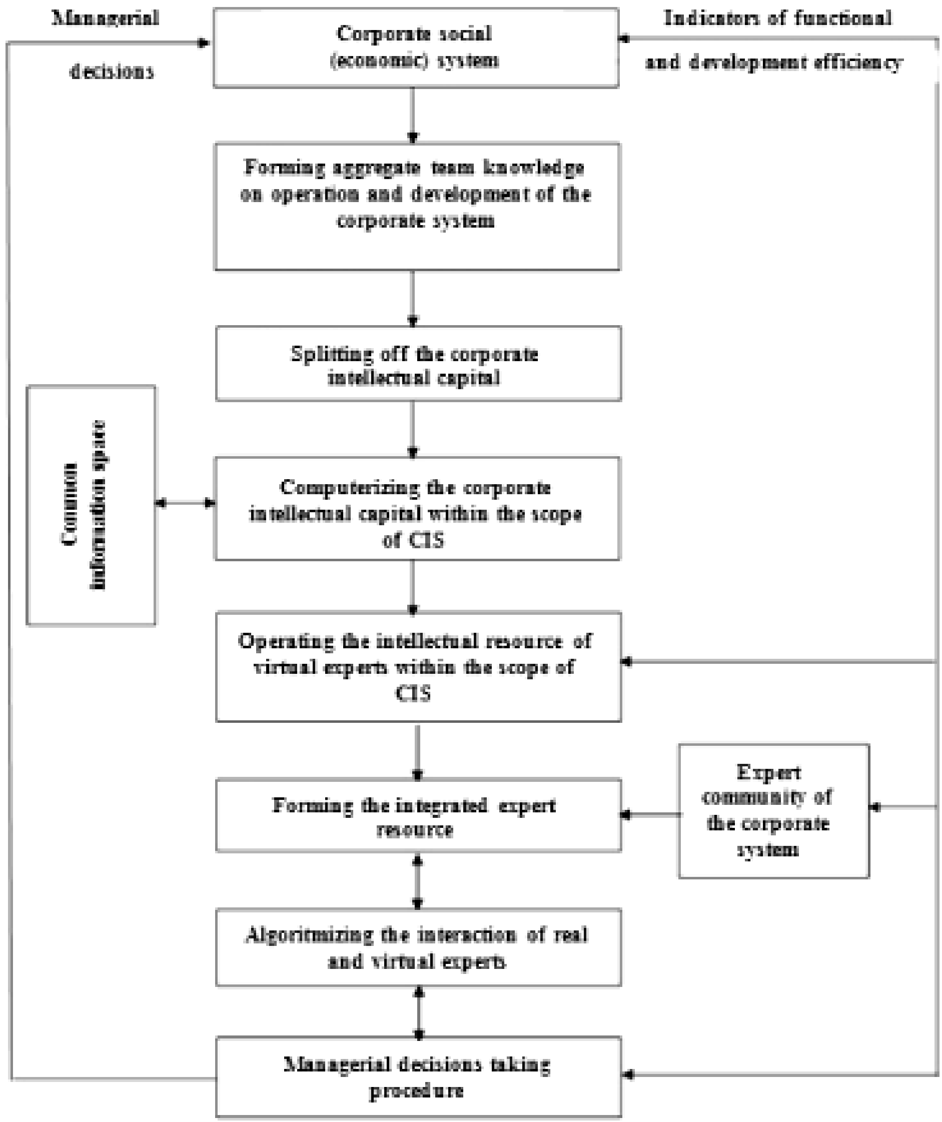


Figure 1. Optimizing and managing corporate systems with virtual expert resource of intellectual capital. A structural flowchart

Currently, partially formalized solutions are predominant; the computer resource viewed as intellectual support for DMP in decision making and the basic question is how to organize the man-machine dialogue optimally.

Combining formal and informal techniques to substantiate decisions assumes wider use of expert assessments and man-machine procedures to prepare for taking decisions. It is suggested to consider the set of such procedures as ‘virtual’ expert while the DMP commanding the necessary scope of knowledge, intuition and experience is a ‘real’ expert.

Then selecting the best (optimal, rational) decision from a set of alternative options depends on the integral assessment based on objective analysis by virtual expert and on subjective understanding of preferences on the part of real experts. The virtual expert resource provides the necessary conditions for equivalent interaction of real and virtual experts to form an integrated intellectual resource.

It is suggested to identify two types of information resource in the corporate intellectual capital that supports interaction between the virtual and real experts during managerial decision making: procedural and knowledge-based. The first virtual expert resource provides intellectual support for decision choosing, the second supports all three stages of decision making.

The procedural virtual expert resource processes current information on the operation and development of corporate intellectual capital applying modeling, optimizing and multi-agent search techniques to structured knowledge-type information.

2. Formalizing the interaction of components of virtual expert resource of procedural type via optimized search modeling

Virtual expert resource of procedural type is meant to intellectualize support for managerial decision making in cases when the corporate knowledge environment is weakly structured, features fuzzy links and a multi-level subordination hierarchy.

The basic components of virtual expert resource are real and virtual experts, the basic principles of their interaction proposed in studies [3, 8].

Either an individual real expert (IRE) or a team of real experts (TRE) is brought in to make decision.

In their turn, virtual experts may be divided into the following types according to the functions they execute during decision making:

– imitational prognostic virtual expert (IPVE);

– multi-alternative virtual expert (MVE);

– multi-agent virtual expert (MAVE).

Each type of experts listed above has its own formalized presentation with the help of the following components:

$x = (x_1, ..., x_j, ..., x_J)$ is the vector of values of varied parameters;

$y = (y_1, ..., y_i, ..., y_I)$ is the vector of values of quantitative indicators;

$\nu = (\nu_1, ..., \nu_{i^1}, ..., \nu_{I^1})$ is the vector of linguistic values from subjective assessments of qualitative indicators;

$F = (F_1, ..., F_{i_1}, ..., F_{I_1})$ is the set of criteria for selecting the best option; a single criterion is used in the extreme case;

$\varphi = (\varphi_1, ..., \varphi_{i_2}, ..., \varphi_{I_2})$ is the set of constraint functions;

$W = (W_1, ..., W_l, ..., W_L)$ is the set of alternative optional managerial decisions (of at least two);

W^* is the optimal (rational) decision;

W_∂ is the set of dominating options close to W^*;

f is the expert preference function including both objective criteria from set F and personal subjective assessments;

Φ(f) is the group preference function depending on the vector of individual preferences of expert team members, $f = (f_1, \ldots, f_d, \ldots, f_D)$, where D is the number of members in expert team;

Γ is the mechanism of group expert valuation and selection of decision.

The sequence of forming the listed components is shown in detailed structural scheme of three-stage decision making procedure (Fig. 2).

Of particular importance for formalized presentation of interaction between the components of virtual expert resource is the multi-alternative virtual expert (MVE).

Analysis of initial information
1. Collecting the data on the coordinates of vectors *x*, *y*. 2. Defining vector ν. 3. Selecting the technique for designing the imitational prognostic model. 4. Building the imitational prognostic model.

↓

Preparation for decision making
1. Forming sets *F* and φ. 2. Selecting the real expert (experts). 3. Selecting the virtual expert. 4. Selecting the mechanisms for group expert valuation, $d = \overline{1, D}$. 5. Forming the set *W* (in case of MVE it is W_∂).

↓

Decision selection
1. Selecting the mode of interaction for real and virtual experts and the respective algorithm. 2. Virtual expert valuation of alternatives. 3. Selecting the final decision.

Figure 2. Structural scheme of three-stage decision making procedure with virtual expert resource of procedural type

To describe MVE, an optimization model is formed, its optimized variables being the coordinates of vector $x = \{x_1, x_2, \ldots, x_n\}$, plus the specially introduced alternative variables:

$$Z_m = \begin{cases} \text{1 in case the alternative at } m\text{-th level of aggregation of } W_m^* \text{ is chosen as} \\ \text{a prospect for forming } W^* \\ \text{0 in the opposite case, } m = \overline{1, M} \end{cases}$$

that form vector $Z = \{ Z_1, \ldots, Z_{m1}, \ldots, Z_M \}$, $m = \overline{1, M}$

Besides, quantitative measurements of subjective assessments are taken into account: $V = \{v_1, \ldots, v_{m2}, \ldots, v_{M2}\}$.

Then the optimization model for MVE takes the form:

$\Psi_{i1}(z, y(x, \nu)) \rightarrow$ extr, $i_1 = i_1 = \overline{1, I_1}$,

$\varphi_{i2}(z, y(x, \nu)) \leq \varphi^*_{i2}$, $i_2 = \overline{1, I_2}$,

$x_j^{мин} \leq x_j \leq x_j^{макс}$, $j = \overline{1, J}$

$$z_m = \begin{cases} 1, & \\ 0, & \end{cases} \qquad m = \overline{1, M}$$

where y (x, ν) denotes the possibility of using the model of dependence of y on (x, ν) while calculating y.

This optimization model serves to form the algorithms for the following functions of MVE [8, 9]:

– forming the set of prospective options;
– aggregating the set of prospective options;
– maintaining the interaction with real expert.

Each function has several implementations. They all hinge on the basic variational procedure over the components of vector z.

Integrating the procedures of variational modeling with MVE functions requires combined use of fuzzy systems, neuron networks and genetic algorithms, i.e. intellectual modeling defined by the concepts of Computational Intelligence.

If MVE optimization model only accounts for vector z:

$\Psi_{i1}(z) \rightarrow$ extr, $i_1 = \overline{1, I_1}$,

$\varphi_{i2}(z) \leq \varphi^*_{i2}$, $i_2 = \overline{1, I_2}$,

$$z_m = \begin{cases} 1, & \\ 0, & \end{cases} \qquad m = \overline{1, M},$$

then the set of prospective options is formed using the multi-alternative optimization model [10].

In case vectors z and x are taken into account:

$\Psi_{i1}(z, y(x)) \rightarrow \underset{z,x}{extr}$, $i_1 = \overline{1, I_1}$,

$\varphi_{i2}(z, y(x)) \leq \varphi^*_{i2}$, $i_2 = \overline{1, I_2}$,

$x_j^{мин} \leq x_j \leq x_j^{макс}$, $j = \overline{1, J}$,

then the set of prospective options is formed using the following models:

– parametric optimization models

$\Psi_{i1}(x) \rightarrow \underset{x}{extr}$,

$\varphi_{i2}(x) \leq \varphi^*_{i2}$, $i_2 = \overline{1, I_2}$,

$x_j^{мин} \leq x_j \leq x_j^{макс}$;

– imitational model of mass service systems and Petri nets;

– model of image transformation.

Vector z is then included in the structure of the respective modeling procedure, making it possible to integrate the basic variational procedure into the algorithmic scheme.

When accounting simultaneously for vectors x and ν, variation of variables z_m, m = $\overline{1,M}$ is merged with forming the y(x, ν) model. That model cannot be constructed in its algebraic form, since the values of subjective assessments are defined by linguistic variables, and their quantitative grades are given by functions of belonging to fuzzy sets. In that case it is feasible to use the model of fuzzy rules in combination with the procedure of variational modeling.

Since each alternative belonging to the set of prospective options W_∂ is presented by components of level m = $\overline{1,M}$, aggregating these alternatives becomes possible via selecting the best combination of components. That function of MVE is executed via:

– mathematical description of the set of prospective options;

– euristic algorithms;

– genetic algorithms.

Upon executing that function, MVE should provide for its possible interaction with real expert. The following forms of interaction are possible:

– the use of linguistic variables;

– visualized imaging mechanisms of expert intuition;

– adaptive accumulation of expert information.

Then the formalized presentation of decision making for various types of experts takes the form [8, 11]:

– individual real expert:

$< x, y/\nu, F, \varphi, W, f, W^* >$;

– team of real experts:

$< x, y/\nu, F, \varphi, W, \Phi(f), \Gamma, W^* >$;

– imitational prognostic virtual expert:

$< x, y, \nu, W/F^{w_l}, \varphi^{w_l}, l = 1,L >$;

– multi-alternative virtual expert:

$< x, y, \nu, F, \varphi/ W_\partial >$;

– multi-agent virtual expert:

$< W/F^{w_l}, y^{w_l}, l = 1,L >$.

Selecting this or that optimization model and managerial decision follows the implementation of one of the possible modes of interaction for the components of virtual expert resource:

1. Virtual interaction mode (VM).

This option is based on interaction between virtual experts. Its first stage consists in Selecting the most effective Agent for Search and Assembly (SASA) of information from the structured information resource of intellectual capital, needed for joint operation of imitational prognostic and multi-alternative virtual experts. The respective formalized presentation looks as:

$< VM > = < MVE > < SASA > < MAVE > < IPVE > < MVE > \rightarrow W_\partial$,

2. Dual mode (DM).

This option is based on interaction between the IRE and VE. Since the activities of those experts are based on contradicting principles (see Table 1) [7], assessing the values of criteria Ψ_{i_1}, $i = \overline{1, I_1}$ for every alternative W_l should be presented as a (2 x J) matrix game (MG).

Table 1 – Activity principles for real and virtual experts during decision making

No.	Real expert	Virtual expert
1	Simplifies the situation in dependence of subject assessments, disregards certain alternatives or consequences thereof	Treats the situation following the accepted technique of model construction; simplification level stays the same for every alternative
2	Starts from subjective value of this or that optional decision called its utility	Uses a combination of objective assessments of quantitative indicators and subjective assessments of qualitative indicators
3	Overestimates probabilities of unlikely events while underestimating probabilities of extremely likely events	Attributes equal assessed probabilities
4	Bases one's decision on maximizing the linear combination of utility and subjective probability of reaching it	Bases one's decision exclusively on calculated values of criteria and constraints
5	Traditions of decision making and personal expert qualities are prevalent over the trend to maximize some criterion	Decisions are only assessed on the basis of imitational prognostic model and multi-alternative optimization

The respective formalized presentation looks like:

$$< \text{DM} > = < \text{IRE} > < > < \text{VE} > \rightarrow W^*_{МИ},$$

where $W^*_{МИ}$ is the optimal pure strategy of matrix game.

The second option is based on interaction between GE and MVE. It runs a matrix game of GE to select option W^* from the set W_∂ formed by MVE, i.e.:

$$< \text{DM} > = < \text{GE} > < > < \text{MVE} > \rightarrow W^*_{МИ}.$$

3. Team mode (TM) [8]

The first TM option is based on interaction between TRE and VE. To obtain a coordinated decision, a mechanism is used from the set Γ for TRE, followed by group assessment by real expert. It is then used to construct a man-machine procedure with VE which follows the imitational prognostic model to retrieve the values of criteria Ψ_i, $i = \overline{1, I}$:

$$< \text{TM} > = < \text{TRE} > < \Gamma > < \text{MMS} > < \text{VE} > \rightarrow W^*.$$

The second option prescribes the interaction of TRE and MVE. Now MVE starts forming the set of dominating options W_∂, and then the group preference by TRE is identified on that set with the use of one of the mechanisms from set Γ following the MMS principle:

$$< \text{TM} > = < \text{TRE} > < \Gamma > < \text{MMS} > < \text{MVE} > \rightarrow W^*.$$

4. Team mode with dominating expert (TMDE) [8]

A dominating expert (leader) is identified among the team of real experts.

The dominating expert (DE) sets the option that is reviewed via question-answer situations (QAS) to reach a decision coordinated with the team of real experts of equal rank. Interaction with virtual expert goes similar to the TM procedure:

$$< \mathrm{TMDE} > = \begin{cases} < \mathrm{DE} >< \mathrm{QAS} >< \mathrm{TRE} >< \Gamma >< \mathrm{MMS} > \\ < \mathrm{VE} > \rightarrow W^* \\ < \mathrm{DE} >< \mathrm{QAS} >< \mathrm{TRE} >< \Gamma >< \mathrm{MMS} > \\ < \mathrm{MVE} > \rightarrow W^* \end{cases}$$

Some of problems were considered in relatives on subject papers [12-20].

Conclusion

The suggested approach is based on the integration of the well-known tools of expert evaluation, the techniques of imitational prognostic modelling and multi-alternative optimization. They form instrumental and technical base for generation of managerial decisions in the virtual expert corporate information environment. That way the negative effect of ambiguity is diminished and the algorithmic procedures can be applied for selection of optimal managerial decisions on sets of alternative optional states of managed entities.

References

[1] Volkova, N.V. Corporate Information Environment as the Basis for Forming the Enterprise Intellectual Capital / Ya.E. Lvovich, N.V. Volkova // Informatics: Problems, Techniques, Technologies: Proc. X Int. Sci. Pract. Conf., Vol. 1, Voronezh: Publ.-Polygraphy Center, Voronezh State Univ., 2010. – p. 445-447.

[2] Volkova, N.V. The Role of Corporate Information Environment in Forming the Enterprise Intellectual Capital // Intellectual Information Systems: Proc. All-Russian Conf. Voronezh: FSBEU HPE “Voronezh State Technical University”, 2011. – p. 208-209.

[3] Lvovich, Ya.E., Lvovich, I.Ya. Taking Decisions in Virtual Expert Environment. – Voronezh: PPC “Science Book”, 2010.

[4] Evlanov, L.G. Theory and Practice of Decision making. – M.: Economics, 1984.

[5] Lvovich, Ya.E., Frolov, V.N., Podvalny, S.L. The Problem of Optimal in Applied Problems. – Voronezh: VSU Publ. House, 1980.

[6] Golubkov, E.P. Technologies of Managerial Decision Making – M.: “Business & Service” Publ. House, 2005.

[7] Engineering Support for Flexible Manufacturing of Radioelectronics Products / S.D. Kretov, V.M. Litvinov, Ya.E. Lvovich et al.– M.: Radio and Communications, 1989.

[8] Volkova, N.V. Formalized Presentation of Interaction of Virtual Expert Environment Components in Web-Oriented Systems of Corporate Management / Ya.E. Lvovich, N.V. Volkova // Bulletin Voronezh State Tech. Univ. – 2010. – Vol. 6, No. 2. – p. 6-9.

[9] Volkova, N.V. Algorithmizing Selection of Optimal Managerial Decisions on the Basis of Interaction of Real and Virtual Experts / Ya.E. Lvovich, N.V. Volkova // Modern Problems in Science and Education. – 2011. – No. 6; URL: http://www.science-education.ru/100-4917

[10] Intellectual Information Systems / Voronezh State Tech. Univ.; Sci. Proceed. Ed. Ya.E. Lvovich. – Voronezh, 2001.

[11] Volkova, N.V. Algorithmizing Selection of Optimal Managerial Decisions in Virtual Expert Environment / Ya.E. Lvovich, N.V. Volkova // Prospective Management Systems and Tasks: Collected References and Articles from the All-Russian Youth Conference. September 20, 2011. – Moscow: FSBEU HPE "G.V. Plekhanov Russian University of Economics", 2011. – p. 140-151.

[12] Preobrazhensky Y. P. Evaluation of the effectiveness of the system of intelligent decision support / Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 116-119.

[13] Zyablov. E. L. Creating object-semantic model of the control system / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 029-030.

[14] Panevin R. Y. optimal control of a multi-stage technological processes / R. Y. Panevin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 77-80.

[15] Zazulin A. V. Peculiarities of building a semantic domain models / A. V. Zazulin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 026-028.

[16] Zyablov. E. L. Development of the linguistic means of intellectual support simulation-based-semantic modeling / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 024-026.

[17] Lvovich Ya. E. Adaptive control of Markov processes in a conflict situation / Ya. E. Lvovich, Y. P. Preobrazhensky, R. Y. Panevin // Herald of the Voronezh state technical University. 2008. Vol. 4. No. 11. P. 170-171.

[18] Zyablov. E. L. Markov decision processes of the first type with multiple absorbing States / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 68-71.

[19] Lvovich Ya. E. Decision making in expert-virtual environment / Ya.E.Lvovich, I.Ya.Lvovich // Voronezh: Publishing house "Science book", 2010, p. 139.

[20] Glushanok T. M. Management of regional economic systems in the modernization of / by T. M. Glushanok, A. M. Luck, J. V. Komarov, I. V. Korolev, A. I. Korolev, I. Y. L., O. G. Pavlov, V. I. Panshin, A. P. Preobrazhensky, E. R. Safargaliev, N. G.Urvantsev, V. N. Filipova // monograph, Saratov, Publisher: limited liability Company "Center of professional management "Business Academy" (Saratov), 2013, P. 197.

Prof. Igor Lvovich, D. Sc.
Paneuropean University, Bratislava, Slovakia
office@vivt.ru

Formalization of user interaction with virtual expert resource of knowledge type via ontological modelling

E. Ruzicky

Abstract:

The question of creation of systems of support of adoption of administrative decisions on the basis of integration of formal (mathematical methods and a computer resource) and informal (mentality of a person) methods that make expert and virtual resource is considered. The description of an expert and virtual resource of the knowledge type which includes the knowledge base and the car of a logical conclusion is given. The ways of representation of knowledge are described, the prospects of use of productional rules or rules of substitution are introduced. The semantic description of basic components of an expert and virtual resource of the knowledge type with the use of ontologies and division of metadata on contextual and content is considered. The scheme of multialternative aggregation of components of an expert and virtual resource of the knowledge type is developed. The formalized procedure of interaction of users with a resource for making reasonable administrative decisions on the basis of optimizing models is offered.

Key words:

Decision-making support system, expert and virtual resource, representation of knowledge, ontologic modelling, expert estimation, optimization.

ACM Computing Classification System:

Control structures, Development frameworks and environments, Software development techniques.

Introduction

The growing importance of the activities for storing, transferring and reproducing of knowledge as the basic resource for operating subjects at various levels, as well as their networking interaction call for more advanced systems of management based on the modern information technologies. Making effective decisions for optimization of expenditures and increase of the reactivity of production in the environment of ever growing information fluxes of varying nature cannot be based only on expert assessments. Decision makers must be provided with modern techniques of analyzing, planning and forecasting based on optimization models and algorithms considering numerous parameters and criteria in the system of information acquisition, storage and processing.

In this regard, the considerable interest is represented by the designing of systems used to support managerial decision making based on integrating expert evaluation and algorithmic procedures for computer (virtual) components viewed as the virtual expert resource of corporate intellectual capital.

1. Expert and virtual resource of knowledge type

Combining formal and informal techniques to substantiate decisions assumes wider use of expert assessments and man-machine procedures to prepare for taking decisions. It is suggested to consider the set of such procedures as ‘virtual’ expert while the DMP commanding the necessary scope of knowledge, intuition and experience is a ‘real’ expert. Then selecting the best (optimal, rational) decision from a set of alternative options depends on the integral assessment based on objective analysis by virtual expert and on subjective understanding of preferences on the part of real experts. The virtual expert resource provides the necessary conditions for equivalent interaction of real and virtual experts to form an integrated intellectual resource.

It is suggested to identify two types of information resource in the corporate intellectual capital that supports interaction between the virtual and real experts during managerial decision making: procedural and knowledge-based. The first virtual expert resource provides intellectual support for decision choosing, the second supports all three stages of decision making.

In its structure the virtual expert resource of knowledge type splits into two basic components – the knowledge base and the logical conclusions machine. The base contains knowledge used by the machine to form logical conclusions. These conclusions are answers by the expert system to enquiries by users wishing to retrieve expert knowledge.

Apparently, the more information is stored in the knowledge base of intellectual helper, the more its actions will remind actions by an expert. Developing an intellectual helper may be an effective intermediate step before building a full-scale CIS to support managerial decision making [23]. Besides, the intellectual helper frees a lot of time for an expert, its use promoting accelerated problem solving.

Expert knowledge on solving specific problems is called the expert knowledge area. The expert system processes information (does its reasoning) in that area and makes logical conclusions (issues expert conclusions) that help reduce risks and shorten time expenditures by decision making persons (DMP).

Knowledge in expert systems may be presented in many ways. One of the wider used techniques consists of rules formulated in the IF-THEN form.

In a rule-based system, knowledge needed to solve problems is coded in the problem area in the form of rules and is contained in knowledge base. Undoubtedly, rules are used the widest for presenting knowledge.

Being the most important operational function, management aims at achieving goals set for each specific system and at forming conditions needed to reach it. Such conditions may include sustainability of some structure, its effective functioning, and support for a prescribed activity mode, saving or forming some qualitative feature in the system, execution of a prescribed program.

Consider the ontology aspect of functional mechanisms used by the virtual expert resource to retrieve and process data in support of decision making. The expert's task consists in analyzing the content of knowledge databases on a given subject area, or in studying facts and points of view on the problem, events and processes in the socio-economic sphere available from the open environment. The data accumulated in the system form knowledge space for subject area or studied problem, its basic form being the ontology.

The problem ontology is composed of classes of facts that reflect problem manifestation, patterns identified in facts contents, normative foundations for the studied problem, and assessments of the impact of that problem on various activities of socio-economic systems. To classify facts a special system is used of qualitative content indicators. To classify such indicators of fact importance as their "significance level" and "uniqueness" one uses numerical scales and logical values. Filter systems make it possible to characterize professional level of authors (sources) of data, their belonging to socio-economic or other sphere, the level of trust in fact content and its scalability, importance and utility.

The system of classes of ontologies makes it possible to model typical processes of knowledge formation presented as a series: single fact – generalized fact – empirical law – hypothesis – formal law. The work of expert forming the base of facts consists in searching new information resources that emerge in the newly opened or specially identified information environments [1]. Elements of the base are processed fragments of primary materials in their digital form with their attributed sets of values of semantic properties used to classify resources and manage the process of their analysis and use.

Separate components of ontology forms a system of rules (criteria) to assess systems of accumulated facts and conclusions stemming from them, described by certain formulas. These make it possible to recognize conditions for making typical decisions and executing special actions. An example of such criteria is the degree of completeness in presenting a typical structure of a process or a phenomenon that defines whether standards are met or special professional actions taken before making decision.

2. The structure of an expert and virtual resource of knowledge type

The virtual expert resource of knowledge type described under Section 1 consists of the following components:

- User interface. It is a mechanism for user interaction with expert system.

- Explanation tool. A component making it possible to explain reasoning by the system to the user.

- Operative memory. The global fact base used in rules.

- Logical conclusions machine. The SW component that shapes logical conclusions (decides which rules agree with facts or objects), prioritizes executed rules and executes the highest priority rule.

- Operative list of rules. Logical conclusions and list of prioritized rules formed by the machine, their templates corresponding to facts or objects that are kept in operative memory.

- Knowledge acquisition means. An automated means for entering knowledge into the system.

Many systems feature tools to acquire knowledge. In some expert systems these instrumental means are capable of self-learning, judging rules by induction: they use examples to develop rules automatically.

In a system based on rules, the logical conclusions machine defines which of the antecedent rules (provided there are such rules) follow the facts.

In dependence of the scope of problems solved by virtual expert resource, the logical conclusions machine of expert system takes either direct or reverse logical conclusion or both such conclusions together. Selecting the logical conclusions machine itself depends on the type of task set. Diagnostic problems are better solved using reverse logical conclusions, while problems of prognosis, current control and management fit direct logical conclusions better.

The logical conclusions machine functions in the mode of "recognition – processing of information – action" cycles until certain criteria are met, which terminates its cyclic activity and calls for "producing the decision".

Upon executing all the rules, management is returned to the interpreter of commands of the top level, so that the user may issue additional instructions to the command interpreter of expert system. Operating in the top level mode corresponds to the default mode in which DMP interacts with its environment via the virtual expert resource of knowledge type. This operation is called the task of "accepting new user command". Accepting new commands occurs particularly at the top level.

During the period when SW application of expert system is developed the top level is a user interface with command interpreter. However user interfaces of a more complex nature are developed capable of simplifying interaction with the expert system.

The main feature of virtual expert resource in support of decision making is the explanation tool envisaged in the system, which enables the user to ask how the system arrived at a specific conclusion. Certain information is needed for that. Rule-based system is capable easily to answer how a given conclusion was obtained, since the chronology of rule activation and the contents of operative memory may be saved to a stack. A developed explanation tool enables the user to ask questions of the type "what if" and study alternative pathways via hypothetical reasoning. In other words, DMP need not limit oneself to well-founded managerial decisions but can forecast various situations in dependence of changes in the external conditions.

The basic problem of traditional expert systems, their information presentation form based exclusively on the 'if-then' rules is their low speed in processing large data arrays from various DBs and multi-alternative user enquiries. Therefore, to implement full scale virtual expert resource of knowledge type in support of decision making one needs to introduce, besides the standard rules for data processing, algorithms and techniques that permit using any rule contained in operative list free of further sequential testing.

In that context it appears promising to use production rules or substitution rules [2]. Such rules find wider use in linguistics as a tool to define language grammar. Obviously, any mathematical or logical system is a set of rules that indicates the ways for transforming a string of symbols into a different set of sequential symbols. In other words, upon receiving the input string (the antecedent) production rule is capable to produce a new string (the consequent).

Such an idea is also true with respect to SW programs and expert systems where initial strings of symbols are input data and output strings result from certain transformations that the input data were subjected to.

That said, the structure of the proposed virtual expert resource of knowledge type in support of decision making looks as shown in fig. 1.

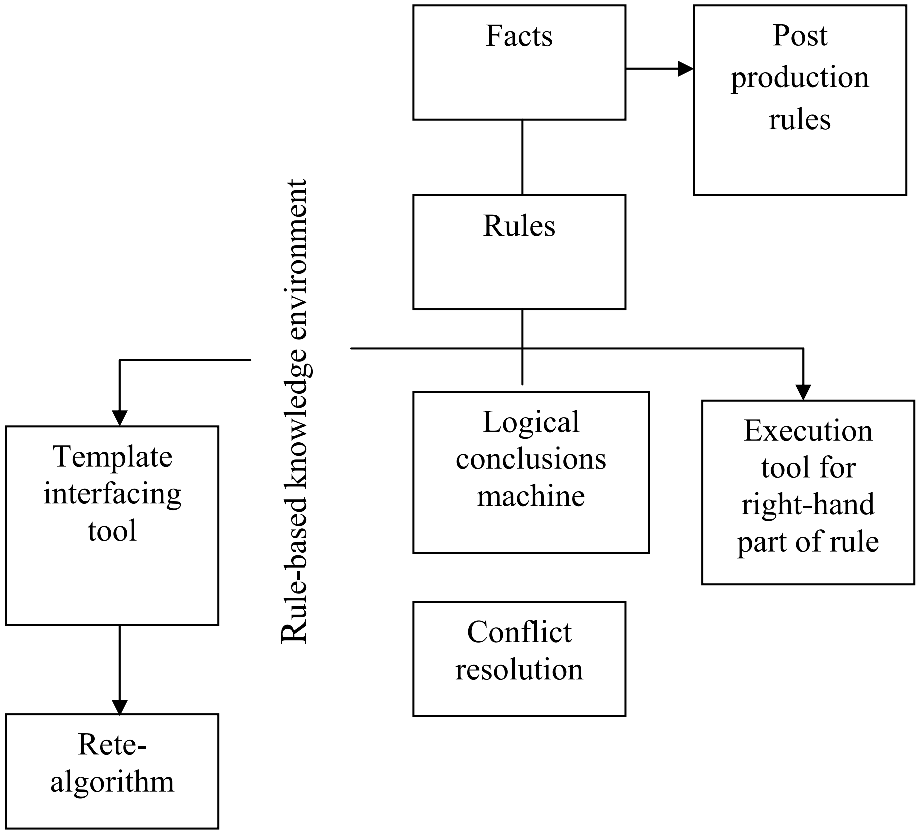


Figure 1. Structure of virtual expert resource for data processing and support in decision making

3. Semantic description of basic components of virtual expert resource of knowledge type that uses ontologies

Consider semantic description of basic components of virtual expert resource of knowledge type that uses ontologies [2], meta-data separated into contextual and content parts (fig. 2).

We introduce the following notations:

M_i are meta-data of i-th object;

O is the ontology containing the specimen (concepts) and predicates (relations);

$M_{ki}(O)$ are contextual meta-data that describe interrelations of concepts with other concepts or literals;

$M_{ci}(O)$ are content meta-data that describe knowledge contained in the specimen concept.

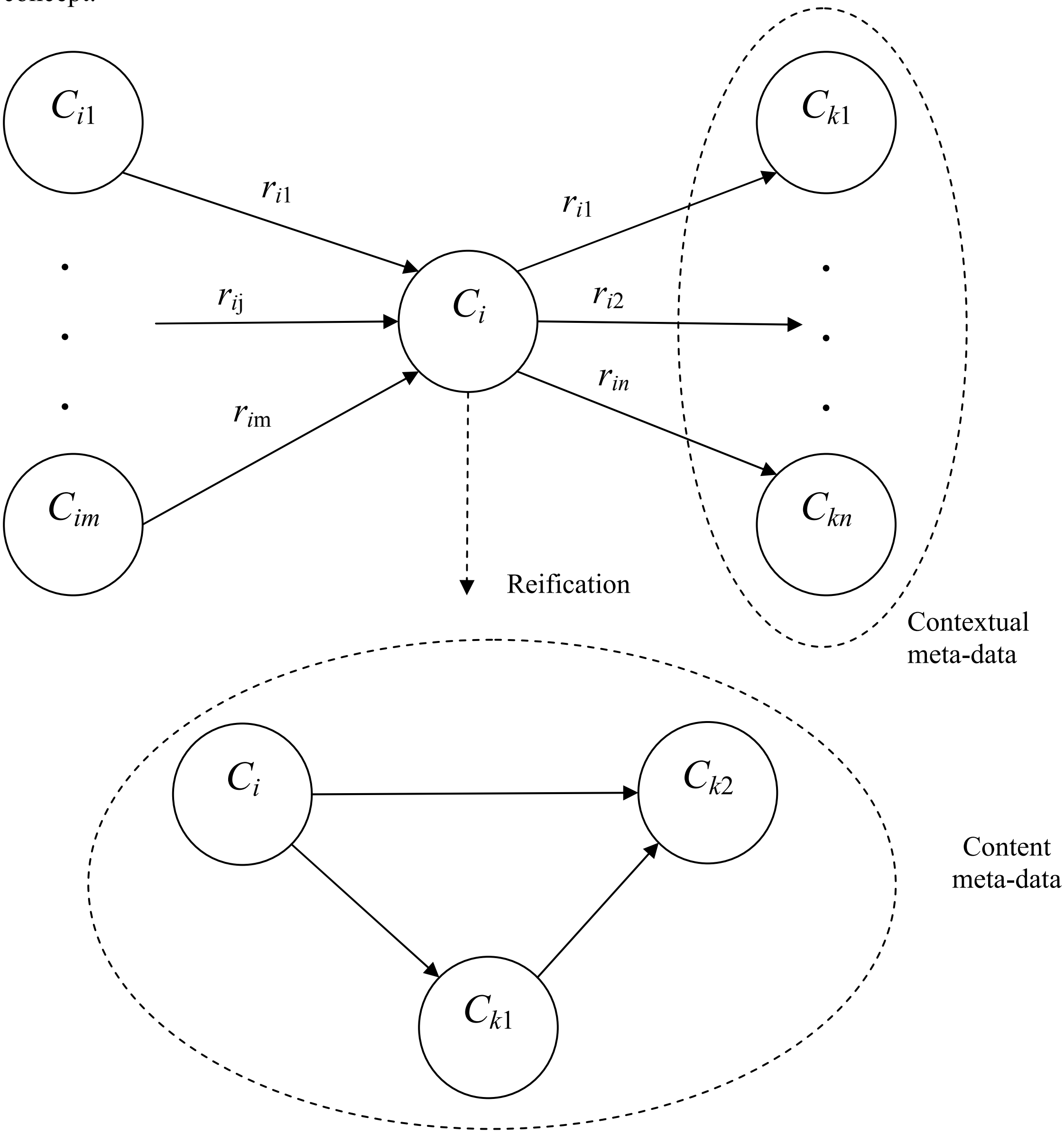


Figure 2. Meta-data structuring

Then:

$$M_i = \{ M_{ki}(O), M_{ci}(O) \},$$

where

$$M_{ki}(O) = (r_1(c_i, v_i) \wedge r_2(c_i, v_2) \wedge r_z(c_i, v_z)),$$
$$M_{ci}(O) = (\{r_1(c_1, v_1), k_1\} \wedge \{r_2(c_2, v_2), k_2\} \wedge \{r_k(c_k, v_k), k_k\}),$$

r_i is the predicate (relation) from ontology O,
c_i is the specimen or the concept of ontology O,
v_i is the specimen or the literal,
k_i are important given statements for object *i*.
Similar ontological description is used for web-services [2].

Therefore, virtual expert resource of knowledge type (*S*) is presented as an aggregation of ontologies used as the basis for constructing meta-data, web-services and technological operations, i.e. as a set of elements W_g $(g = \overline{1, G})$, each of them being some combination dependent on the set of ontologies *O* and the set of technological operations *T*:

$$W_g(O_g, T_g).$$

The basic techniques offered for constructing virtual expert resource are based on aggregating preliminarily chosen elements W_g into a combination of ontologies and technological operations (fig. 3).

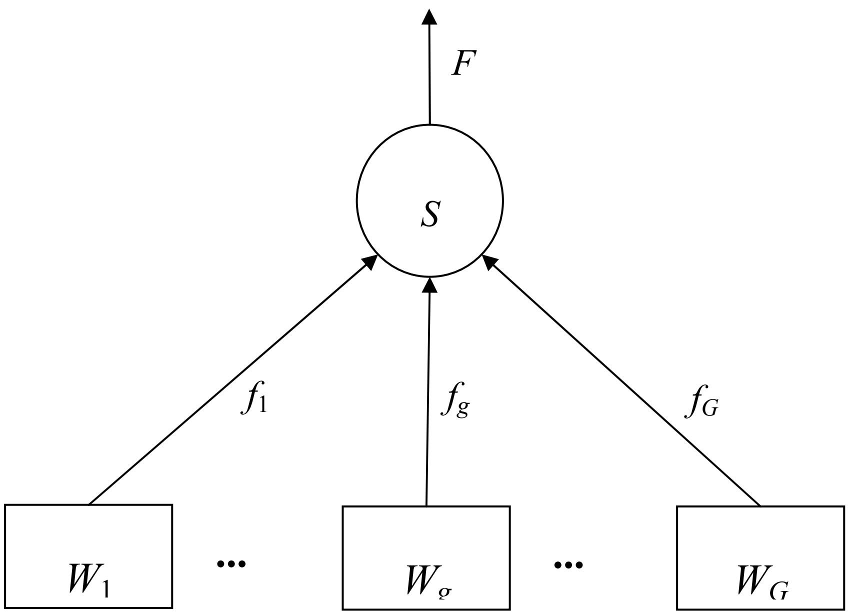


Figure 3. Aggregation scheme for the components of virtual expert resource of knowledge type

Such aggregation results in a certain level of vector criterion

$$F = \{F_1, \ ..., \ F_j, \ ..., \ F_J\}$$

where F_j, $j = \overline{1, J}$ are the criteria for selecting alternative options.

Note that vectors $f_g, g=\overline{1,G}$ include the values of indicator elements W_g, the values of indicator F_j, $j=\overline{1,J}$ for system *S* dependent on them.

Alternative implementation of ontologies and technological operations results in the diversity of the following sets:

set of ontologies

$$O \subset \times\{o_i : i \in I\},$$

where $\subset$ denotes relation; $\times$ denotes Cartesian product; $o_i = \overline{1,O_i}$ is the set of numbers of alternative optional implementations of ontologies of *i*-th type; *I* is the set of indices of the types of ontologies included in the CIS;

set of technological operations

$$T \subset \times\{t_k : k \in K\},$$

where $t_k = \overline{1,T_k}$ is the set of numbers of alternative optional implementations of technological operations of *k*-th type, *K* is the set of indices of the types of technological operations implemented in the system.

The diversity of elements O_i and t_k results in the diversity of combinations:

$$\omega_g = (O_i, t_k) - \omega_g = \overline{1,W_g}.$$

In that case forming the virtual expert resource follows the scheme of multi-alternative aggregating (Fig. 4) and results in a multi-optional character of the system itself

$$S_l, l = \overline{1,L},$$

where *L* is the total number of options.

Each option is characterized by its vector of criteria F_l. Selecting the best option is done with the multi-alternative optimization technique [3].

The first stage of optimizing consists in selecting implementation ontologies $O_i^*, i=\overline{1,I}$ and technological operations $t_k^*, k=\overline{1,K}$ according to vector criterion $F = \psi(f\omega_g(O_i, t_k))$ with the account of constraints of instrumental environment chosen to implement the components of virtual expert resource.

During the second stage those groups of concepts and relations are chosen from the ontology O_i^* that maximize semantic proximity to external meta-data if applied to the pairs of such meta-data operations t_k^* (fig. 5).

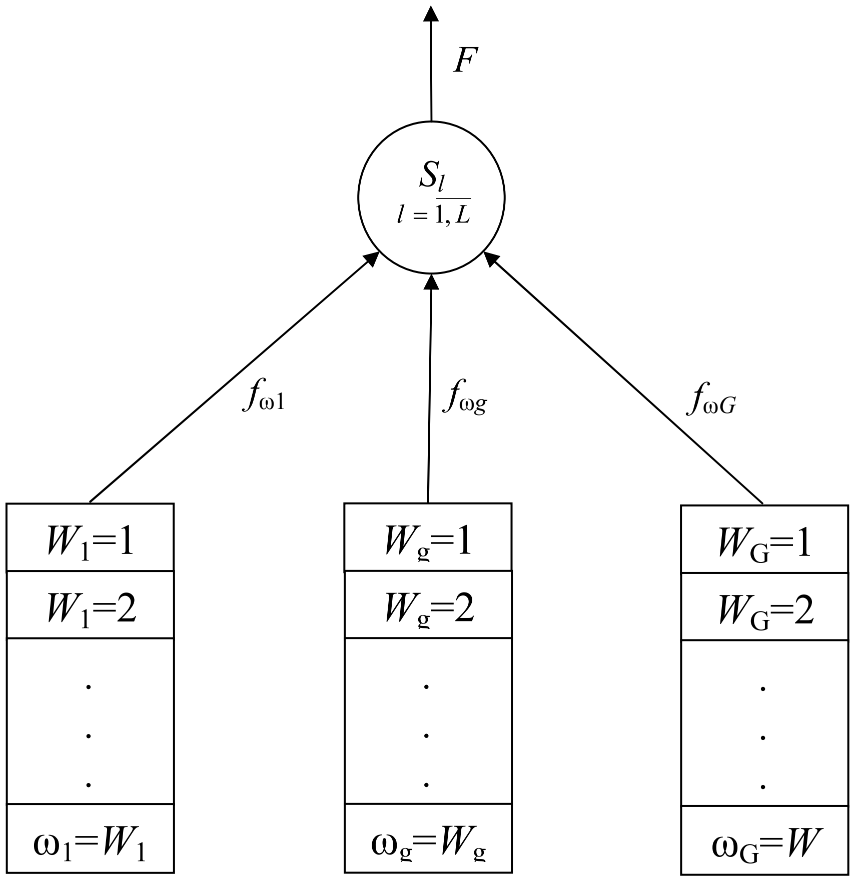


Figure 4. Scheme of multi-alternative aggregation of the components of virtual expert resource of knowledge type

Assessing semantic proximity B of the pair of meta-data M_n, M_p is done according to the formula:

$$B(M_k, M_p) = \sum\sum sim(m_i, m_j) . \sum_{m_i \in M_k} \sum_{m_j \in M_p} sim(m_i, m_j),$$

where m_i, m_j are sub-sets of meta-data, belonging to the sets of meta-data M_k and M_p, respectively

$$sim(m_i, m_j) = sim((c_i, r_j, i_k, k_l), (c_x, z_y, i_z, k_\omega) = (sim_c(c_i, c_x) + sim_R(z_j, z_y) + sim_I(i_k, i_z)) f(k_1, k_\omega),$$

$sim_c(c_i, c_x)$ is semantic proximity of concepts c_i, c_x in ontology O_i^*;

$sim_R(z_j, z_y)$ is semantic proximity of relations z_j, z_y in ontology O_i^*;

$sim_I(i_k, i_z)$ is semantic proximity of contextual meta-data in specimen of concepts i_k and i_z;

$f(k_1, k_\omega)$ are functions used to account for the importance factors of statements.

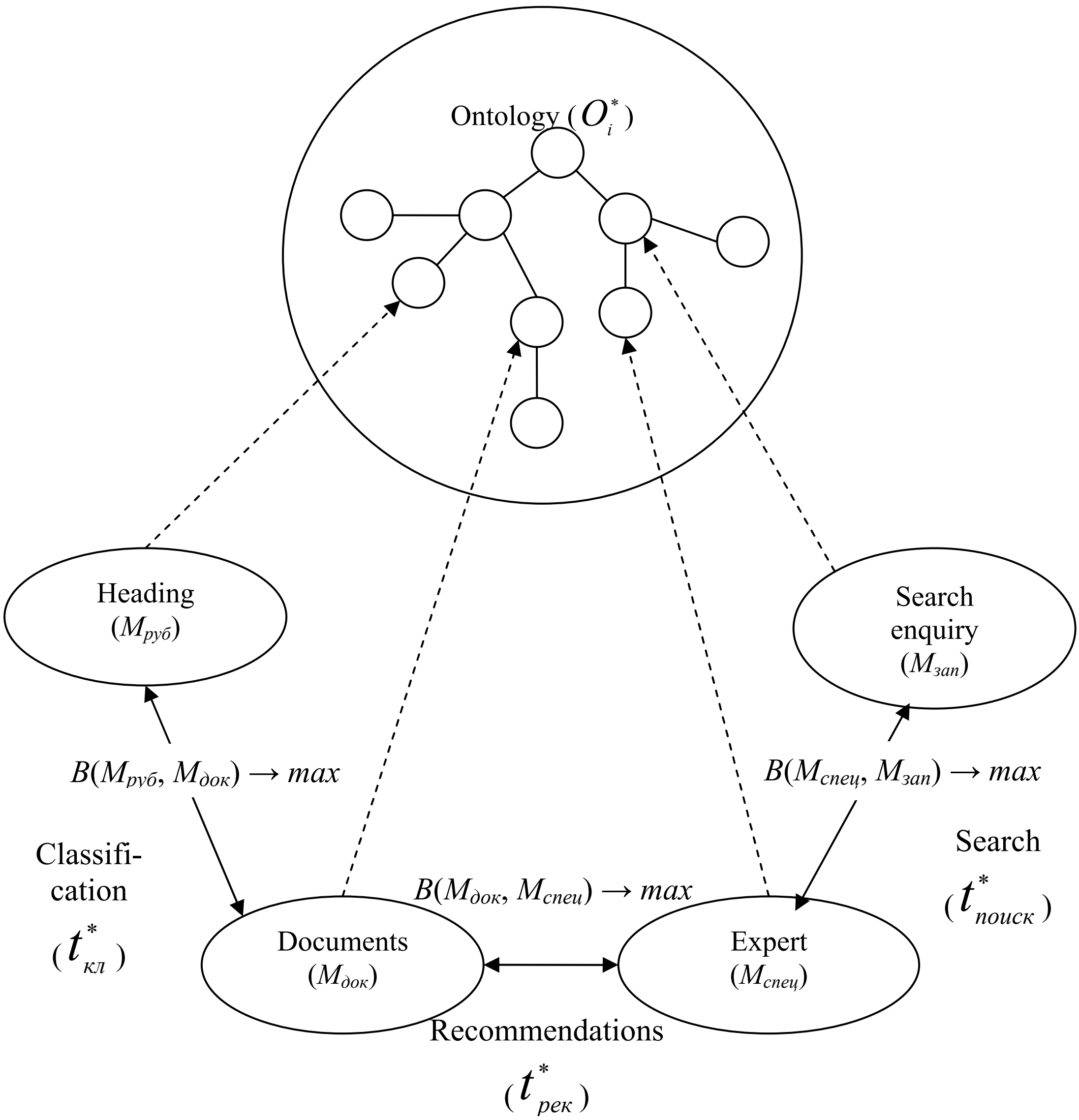


Figure 5. Second stage of selection with the use of optimization model and semantic proximity

In papers [4-12] we can see similar problems, connected with control of complex objects.

Therefore, the suggested model not only provides for semantic description of storing the virtual expert resource knowledge for its further automated processing, integration and repeated use, but offers a formalized procedure for user interaction with that resource to select well-founded managerial decisions with the use of optimization model.

Conclusion

As a result of the conducted research the option of creation of system of support of adoption of administrative decisions based on integration of formal and informal methods thar make an expert and virtual resource is offered. The semantic description of basic components of an expert and virtual resource of knowledge type with the use of ontologies is given. The model which provides the formalized procedure of interaction of users with a resource for a choice of reasonable administrative decisions is suggested.

References

[1] Developing SW Applications for Corporate Information Systems / Voronezh State Tech. Univ.; Sci. Ed. Ya.E. Lvovich. – Voronezh, 2006.

[2] Rogushina, Yu.V., Gladun, A.Ya. Ontological Approach to Multi-Linguistic Analysis of Information Resources in the Internet // Collected Works VI Int. Conf. “Intellectual Analysis of Information IAI-2006”. – Kiev: Prosvita, 2006.

[3] Lvovich, Ya.E. Multi-Alternative Optimization: Theory and Applications. – Voronezh: “Quarta” Publ. House, 2006.

[4] Preobrazhensky Y. P. Evaluation of the effectiveness of the system of intelligent decision support / Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 116-119.

[5] Zyablov. E. L. Creating object-semantic model of the control system / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 029-030.

[6] Panevin R. Y. optimal control of a multi-stage technological processes / R. Y. Panevin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 77-80.

[7] Zazulin A. V. Peculiarities of building a semantic domain models / A. V. Zazulin, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2008. No. 3. P. 026-028.

[8] Zyablov. E. L. Development of the linguistic means of intellectual support simulation-based-semantic modeling / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2009. No. 5. P. 024-026.

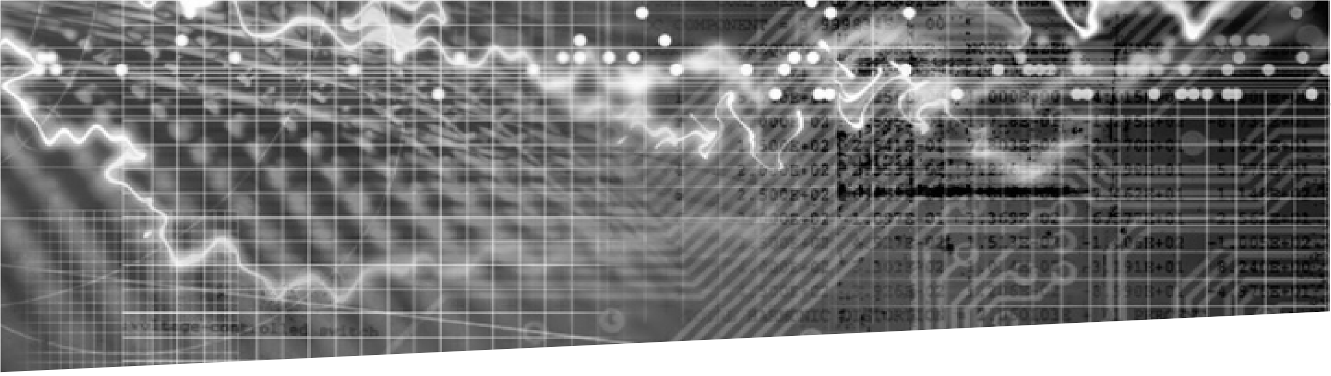
[9] Lvovich Ya. E. Adaptive control of Markov processes in a conflict situation / Ya. E. Lvovich, Y. P. Preobrazhensky, R. Y. Panevin // Herald of the Voronezh state technical University. 2008. Vol. 4. No. 11. P. 170-171.

[10] Zyablov. E. L. Markov decision processes of the first type with multiple absorbing States / E. L. Zyablov, Y. P. Preobrazhensky // Herald of the Voronezh Institute of high technologies. 2010. No. 6. P. 68-71.

[11] Lvovich Ya. E. Decision making in expert-virtual environment / Ya.E.Lvovich, I.Ya.Lvovich // Voronezh: Publishing house "Science book", 2010, p. 139.

[12] Glushanok T. M. Management of regional economic systems in the modernization of / by T. M. Glushanok, A. M. Luck, J. V. Komarov, I. V. Korolev, A. I. Korolev, I. Y. L., O. G. Pavlov, V. I. Panshin, A. P. Preobrazhensky, E. R. Safargaliev, N. G.Urvantsev, V. N. Filipova // monograph, Saratov, Publisher: limited liability Company "Center of professional management "Business Academy" (Saratov), 2013, P. 197.

Doc. RNDr. Eugen Ruzicky, CSc.
Paneuropean University, Bratislava, Slovakia
eugen. ruzicky@paneurouni.com



The use of quantile regression in data analysis tasks

Применение квантильной регрессии в задачах анализа данных

E. Ruzicky, V. N. Kostrova
Е. Ружицки, В. Н. Кострова

Abstract:

Theoretical and applied aspects of the use of quantile regression in data mining tasks are considered. The features of application of quantiles for modeling and data mining are introduced. It is shown that the most complete picture of the impact of the explanatory variable on the shape of the distribution provides the finding of conditional quantiles, that is, the use of quantile regression. The methods of construction of quantile regression are analyzed and the two-step procedure with the use of the method of k-nearest neighbors for the computation of the quantile functions and kernel smoothing for the final quantile regressions is carried out. The analysis of the efficiency of application of quantile regression models in different socio-economic areas is analyzed.

Аннотация:

Рассматриваются теоретические и прикладные аспекты использования квантильной регрессии в задачах интеллектуального анализа данных. Приводятся особенности применения квантилей в задачах моделирования и интеллектуального анализа данных. Показано, что наиболее полную картину влияния поясняющей переменной на форму распределения обеспечивает нахождение условных квантилей, то есть, использование квантильной регрессии. Проведен анализ методов построения квантильной регрессии, рассмотрена двухступенчатая процедура с использованием метода k-ближайших соседей для вычисления квантильных функций и ядерного сглаживания для финальных квантильных регрессий. Проведен анализ эффективности применения квантильно-регрессионных моделей в различных социально-экономических областях.

Key words:

Monitoring, data analysis, regression analysis, quantile regression.

Ключевые слова:

Мониторинг, анализ данных, регрессионный анализ, квантильная регрессия.

ACM Computing Classification System:

Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

Квантильная регрессия используется в связи с экстремальными событиями – этот тип предполагает намеренное введение смещения в результат, повышая точность модели. Метод квантильной регрессии при анализе трендов в различныхзадачах позволяет получить информацию о трендах по всему диапазону значений квантилей от 0 до 1 распределений зависимой переменной, что дает больше информации, чем использование традиционной, основанной на методе наименьших квадратов (МНК) регрессионной техники, дающей возможность получить оценки трендов лишь для средних значений зависимой переменной. В данной работе мы рассмотрим примеры использования квантильной регрессии для решения различных задач.

1. Применения квантилей в задачах интеллектуального анализа данных

Целью мониторинга как информационного процесса является производство нового знания, которое будет использоваться для принятия управленческих решений. Для нахождения нового знания на основе наблюдаемых мониторинговых индикаторов необходимо построение моделей и оценка их результатов. На сегодняшний день наиболее распространенным методом описания существующих закономерностей в данных является регрессионный анализ.

Содержание регрессионного анализа составляют методы нахождения зависимости средних значений от контролируемой переменной. Если регрессор X – независимая случайная величина, то можно рассматривать регрессионную модель как условную относительно действительно наблюдавшихся значений регрессора x_i, т. е.

$$\mathrm{M}\left[Y_i \middle| X_i = x_i\right] = f(x_i) + \varepsilon_i \tag{1}$$

где $f \in C^2[a,b]$ – некоторая сглаживающая функция, определенная на интервале x∈(a,b);

ε_i – отклонение (случайная ошибка) – нормально распределенная случайная величина с нулевым средним и единичной дисперсией ε∈ $N(0,\sigma^2)$.

Являясь зависимостью всего одного параметра распределения, регрессионная функция не отражает влияние объясняющих переменных. Так, если зависимая переменная цензурирована слева или справа, зависимость, отражаемая регрессионной функцией, искажается. Из характеристик положения медиана наиболее устойчива к выбросам и цензурированию, чем среднее значение.

В [2] указывается что, если регрессионная кривая обобщает средние значения распределений, соответствующих множеству X, то несколько регрессионных кривых для разных частей распределений позволяет получить полную картину этого множества. Значение, характеризующее отдельную часть выборки или долю (процент) наблюдений, составляет квантиль или процентиль.

Строгое математическое определение квантили согласно следующее: если Y – случайная переменная, имеющая функцию распределения F(y) или плотность распределения f(y), то квантилью q_τ порядка $\tau \in [0,1]$ одномерного распределения называется такое значение y_τ случайной величины Y, для которого функция распределения принимает значение τ или имеет место «скачок» со значением меньше τ до значения больше τ, т. е. $P\{Y \le y_\tau\} = F(y_\tau) = \tau$. Для непрерывных распределений, квантиль порядка τ, где число $\tau \in [0,1]$, определяется как решение уравнения:

$$F(q) = \int_{-\infty}^{q} f(y)dy = \tau . \qquad (2)$$

Поскольку τ может рассматриваться как переменная, определенная на интервале от 0 до 1, то $q(\tau)$ может быть функцией от вероятности τ или обратная к функции распределения:

$$q_y(\tau) = F_y(\tau)^{-1},$$

каждое значение которой есть значение случайной величины Y, вероятность наступления которого больше или равна $\tau \in [0,1]$.

Так, если Y – случайная переменная, имеющая функцию распределения F(y), то для любого значения τ, принадлежащего интервалу от 0 до 1, квантильная функция может быть определена как точная нижняя граница множества R точек y, при которых функция распределения принимает значения больше или равные τ:

$$Q_Y(\tau) = \inf_{y \in \Re}\{F(y) \ge \tau\}.$$

Квантильная функция, являясь обратной к функции распределения, может также полно описывать случайную величину Y как и функция распределения.

Самая известная квантиль – это медиана или квантиль порядка 0,5, которая делит выборку пополам, оставляя с обеих сторон от себя равное число наблюдений, т. е. $P\{Y \le \mu\} = P\{Y \ge \mu\} = 0{,}5$. Медиана является робастной оценкой центральной тенденции выборки и характеризует положение центра распределения. Далее, по аналогии, определяется нижняя и верхняя квартиль, так же разделяющие половины выборочных значений на две равные по объему выборки, F(y)=1/4 и F(y)=3/4 соответственно. Расстояние между квартилями – интерквантильный размах, который аналогично среднеквадратическому отклонению служит мерой рассеивания случайной величины. Затем также определяется нижняя и верхняя октили. В [3] указывается, что «даже значения квартилей и медианы дают хорошие сведения о характере распределения».

Часто встречается употребление вместо термина квантиль понятия процентиль, т. е. квантиль порядка τ. Процентили и центили делят выборку на 100 и 10 частей, попадания в которые равновероятны. Любая $\tau \cdot 100$ процентиль – это такое значение, меньше которого наблюдается $\tau \cdot 100$ % значений, а все остальные выборочные значения $(1-\tau) \cdot 100\,\%$ больше этого значения. Процентили определяют относительное положение данного значения в выборке, т.. прямо указывают на отклонения от принятых или допустимых пределов, а также могут трактоваться как рейтинговая оценка каждого наблюдения по 100 бальной шкале.

Рассмотрим пример – динамика заболеваемости нефропатиями по 33 районам Воронежской области за 2004-2008 гг. (рис. 2). Построенная линия регрессии имеет некоторую тенденцию к увеличению, что свидетельствует о росте заболеваемости в целом в регионе.

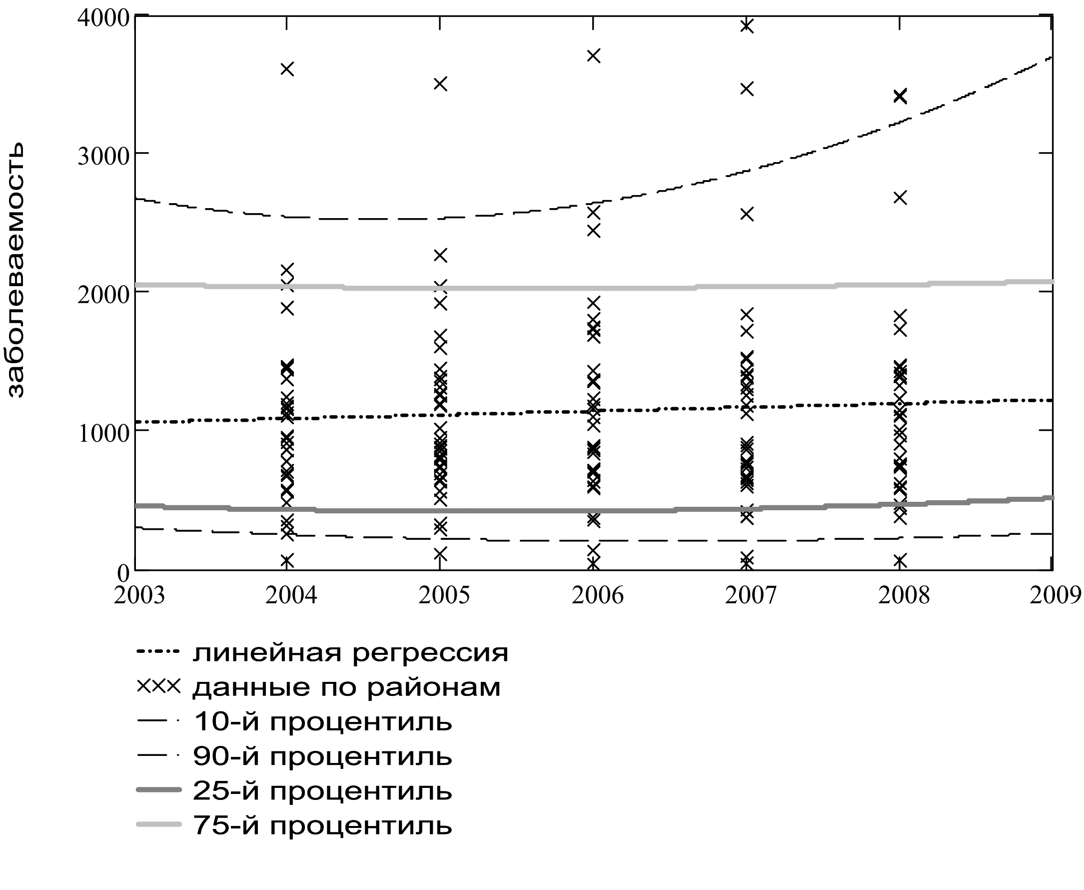


Рис. 2. Динамика заболеваемости по годам

Результат 10 % выборочных наблюдений с наименьшими значениями заболеваемости, свидетельствует о тенденции к стабилизации заболеваемости, а в 10 % выборки с наибольшими значениями наблюдается высокий рост заболеваемости. В выборке из 25 % наблюдений со значениями меньше среднего прослеживается рост заболеваемости, а выборке такого же объема, но с большими среднего значениями наоборот убывание. Следовательно, для районов с исходно высоким уровнем характерен рост заболеваемости, тогда как в районах, имеющих низкий уровень, значение заболеваемости практически не меняется. Также имеется ряд районов, в которых при исходно невысоком уровне наблюдается увеличение заболеваемости, и ряд районов с исходно высоким уровнем, где, вероятно, идет профилактическая работа, способствующая снижению уровня заболеваемости.

Таким образом, использование анализ процентильных значений создает наиболее полную картину, необходимую для мониторинга заболеваемости нефропатиями и позволяет оптимально корректировать обстановку в регионе.

Лечебный эффект может быть определен как «горизонтальное расстояние» $\Delta(x)$ между настоящим распределением F в контрольной группе и новым G в группе после вмешательства: $F(x)=G\{x+\Delta(x)\}$. Это может быть записано в терминах обратных функций, т. е. квантильных, как: $\Delta(x) = G^{-1}\{F(x)\} - x$.

Для $\tau=F(x)$ квантильный лечебный эффект определен как: $\delta(\tau) = \Delta\{F^{-1}(\tau)\} = G^{-1}(\tau) - F^{-1}(\tau)$ для каждой квантили настоящего распределения F.

Стандартная линейная регрессия $\mathrm{M}[Y_i|D_i] = \alpha + \beta \cdot D_i$ с оценкой параметров методов наименьших квадратов показала только наличие значимых различий между группами. Квантильная регрессия для $\tau\in(0,1)$ $Q_y[\tau|D_i] = \alpha(\tau) + \beta(\tau) \cdot D_i$ позволила выявить индивидуальные различия в каждой квантили.

Поскольку функция условной квантили дает более полную картину условного распределения зависимой переменной, чем функция условного среднего. По аналогии с регрессионным анализом статистический метод, позволяющий оценить параметры условных квантильных функций, известен как квантильная регрессия.

Так как каждая квантильная регрессия – это функция некоторого множества регрессоров для значения квантили заданного порядка:

$$\mathrm{Q_y}[\tau|\mathrm{x}] = \mathrm{g}(\beta\mathrm{x}) + \varepsilon_\mathrm{i}, \tag{3}$$

то множество функций, соответствующих определенному набору порядков составляют семейство условных квантильных функций. Самым распространенным является следующее упорядоченное множество значений порядков квантилей:

$$\{0{,}05;\ 0{,}10;\ 0{,}25;\ 0{,}50;\ 0{,}75;\ 0{,}90;\ 0{,}95\}.$$

2. Анализ методов построения квантильной регрессии

Квантильная регрессия довольно «старинный» статистический метод, упоминание этого термина в математической статистике датируется еще XIX веком.

Во многом его «забвение» связано с широкой распространенностью метода наименьших квадратов для вычисления линейной регрессии. Ставшее сейчас классическим определение квантильной регрессии было введено Коенкером и Бассетом в 1978 г., как расширение понятия порядковых квантилей или процентилей в локальных моделях к общему классу линейных моделей, в которых условные квантили имели линейную форму [4].

По аналогии с нахождением условного среднего, которое можно рассматривать как решение задачи минимизации остаточной суммы квадратов:

$$\arg\min_{\mu\in\Re}\sum_{i=1}^{n}(y_i-\mu)^2,$$

где y_i – значение наблюдения из выборки объема n,

μ – выборочное среднее, оцениваемое по этой выборке,

Поиск медианы может быть осуществлен как минимизация суммы абсолютных остатков.

Таким образом, нахождение квантили q заданного порядка τ можно рассматривать как поиск аргумента минимума специальной целевой функции:

$$\arg\min_{q\in\Re}\sum_{i=1}^{n}\rho_\tau(y_i-q),$$

где $\rho_\tau(u)$ – контрольная функция, обеспечивающая τ-баланс наблюдаемых значений и заданная в виде:

$$\rho_\tau(u)=\begin{cases}\tau\cdot u, & u\ge 0\\ (\tau-1)\cdot u, & u<0\end{cases}.$$

Если в регрессионном анализе условное среднее представляется в виде: $M\{Y/x\}=f(õ_i,\beta)+\varepsilon_i$, то оценка параметров β регрессионной функции есть решение оптимизационной задачи:

$$\arg\min_{\beta\in\Re}\sum_{i=1}^{n}(y_i-\mu(x_i,\beta))^2. \quad (4)$$

По аналогии с регрессионным анализом можно перейти к определению квантильно-регрессионных функций $\nu(x_i,\beta)$, каждая из которых представляет собой некоторую регрессию условной квантили $q(\tau)\{Y_\tau/x\}$. Тогда построение квантильно-регрессионных моделей можно рассматривать как задачу оценки параметров функций $\nu(x_i,\beta)$ и находить решение минимизацией:

$$\arg\min_{\beta\in\Re}\sum_{i=1}^{n}\rho_\tau(y_i-\nu(x_i,\beta)). \quad (5)$$

Решение представленной минимизационной проблемы, когда $\nu(x_i,\beta)$ – линейная функция с неизвестными параметрами, эффективно осуществляется методами линейного программирования.

В частности, линейная квантильно-регрессионная модель [5] заданная в виде:

$$y_i = \mathbf{x_i} \cdot \beta_\tau + \varepsilon_{\upsilon i} \text{ или } \tau = \int_{-\infty}^{\mathbf{x_i} \cdot \beta_\tau} f_y(s \mid x_i) ds ,$$

где β – неизвестный вектор регрессионных параметров, оценивается как решение минимизационной задачи:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i: y_i > x_i \beta} \tau \left| y_i - x_i \cdot \beta \right| + \sum_{i: y_i < x_i \beta} (1 - \tau) \left| y_i - x_i \cdot \beta \right| \right),$$

где $\varepsilon_{\upsilon i}$ – неизвестный вектор ошибок, так что условная квантиль порядка τ его значений равна нулю.

Такого рода оценки получили название L2-оценки, так как они основываются на минимизации взвешенных остатков. При решении задач оценивания параметров регрессионных моделей наиболее распространен критерий наименьших квадратов по (4), т. е. М-оценки.

Альтернативой является использование оценок максимального правдоподобия или L-оценок, которые в отличие от МНК не требуют нормально распределенных ошибок. Оценка максимального правдоподобия находится минимизацией функции правдоподобия [6, 7]:

$$\arg\min L_y(\theta_i), \tag{6}$$

где L – функция правдоподобия выборки размера n для заданных выборочных значений $\{y_1, y_2 ... y_n\}$ от неизвестного параметра Θ функции распределения вероятностей случайной величины Y:

$$L = \prod_{i=1}^{n} f_i(y_i, \theta) .$$

Наиболее общий характер носит процедура байесовского оценивания. Если задана функция потерь $W(\hat{\theta}, \theta)$, определяющая потери вследствие отклонения полученных оценок параметров θ от истинных значений, то байесовские оценки параметров находятся из условия минимума функции риска [7, 8].

Оценки максимального правдоподобия и оценки наименьших квадратов могут быть получены как частные случаи байесовских оценок при определенном выборе функции потерь и при соответствующих вероятностных мерах $dG(\theta)$ и $dF_\theta(Y_N)$ [7, 9]. Поэтому возможно множество подходов к оценке параметров квантильной регрессии для различных вероятностных мер и заданной функции потерь.

В исходной работе [10] сглаживание всех трех функций, описывающих модель, производилось отдельно с помощью сплайнов, и выбор модели сводился к подбору оптимального числа узлов каждой из трех функций. В дальнейшем использовались различные модификации, так Yee предложил оценивать все три функции совместно вектором сплайнов [11].

В качестве альтернативы нормальному распределению различными учеными предлагалось использовать t-распределение, кривые Джонсона, экспоненциально-модальное распределение, гамма-распределение.

В работе [12] предлагается использовать степенное экспоненциальное распределение Box-Tiao или общих ошибок, которое является общим вариантом задания различных одномодальных распределений от нормального до равномерного для данных, имеющих слишком большой эксцесс после использования трансформации Бокса-Кокса.

При отсутствии априорной информации о форме распределения предложены непараметрические квантильно-регрессионные модели. В частности, обсуждается вопрос об использовании ядерного оценивания функции условного распределения и получение условной квантили обращением этой функции. Решая вычислительную проблему обращения оцененной функции условного распределения Yu and Jones (1998) использовали двойную ядерную аппроксимацию как минимизацию [13]:

$$\arg\min_{\nu} \sum_{i=1}^{n} \rho_{\tau}(y_i - \nu) \cdot K_h(x_i - x),$$

где $\nu=\nu(x)$ – оценка квантильной регрессии,

К – ядро с заданной шириной окна h.

Соответствующая функция реализована Yu для пакета S-PLUS, разработанный алгоритм гарантирует сходимость.

Наибольшее многообразие решений квантильной регрессии представлено в специальном статистическом ПО – SAS, в котором на сегодняшний день реализованы – симплекс-алгоритм, алгоритм с внутренней точкой, сглаживающий алгоритм, которые основаны на преобразовании задачи (5) в задачу линейного программирования.

Реализация LMS-метода выполнена в специальном прикладном ПО lmsChartMaker, разработанном T. Cole and H. Pan, имеется публикации как исходного текста FORTRAN-программы Коула, так и код Кери (Carey) для ее реализации S-PLUS на http://biosun1.harvard.edu/`carey/.

Множество работ Коенкера (Koenker) по оценке линейной квантильной регрессии представлено в качества отрытого кода на языке R (http://cran.r-project.org) и в виде функции на языке S для пакета S-PLUS (http://econ.uiuc.edu/roger).

Специализированный статистический пакет STATA имеет команду «qreg» для оценки квантильной регрессии, библиотека подпрограмм команд STATA постоянно расширяется и все множество пользовательских версий вычисления квантильной регрессии можно найти на http://www.jstor.org.

В специальном математическом ПО XploRe имеется возможность оценки параметров квантильной функции и ряд сервисных процедур по проверке гипотез и построения графиков [14].

Также известны специальные прикладные реализации некоторых алгоритмов квантильной регрессии – пакет VGAM, разработанный T. W. Yee (http://www.stat.auckland.ac.nz), реализующий LMS-метод, оценку квантилей при исходном гамма распределении значений и использующий свою модификацию трансформации к нормальности для положительных и отрицательных значений.

Проект GAMLSS (Generalized Additive Models for location, scale and shape – обобщенные аддитивные модели для положения, смещения и формы) разработан для полупараметрических регрессионных моделей.

Предполагается, что параметрическими должны быть функции распределения зависимых переменных (отклика) Y, а моделирование параметров распределения как функций независимых регрессоров X может осуществляться с использованием непараметрических сглаживающих функций. В основе моделей GAMLSS лежит экспоненциальное семейство распределений, которое позволяет моделировать непрерывные и дискретные распределения зависимых переменных Y с высоким коэффициентом асимметрии и эксцесса. Систематическая часть моделей разложена, чтобы позволить моделировать не только среднее (положение), но все другие параметры распределения Y как линейными, так и нелинейными параметрическими и аддитивными непараметрическими функциями независимых переменных со случайными эффектами (http://www.gamlss.com).

Применение подхода на основе сглаживания предполагает, что для каждого фиксированного значения переменной X осуществляется выборка соответствующих значений зависимой переменной Y, по которой вычисляется выборочная квантиль заданного порядка. Упорядоченные по фиксированным значениям X квантили одного и того же порядка интерполируются гладкой непрерывной функцией. Поскольку подобная задача решается в два этапа:

1) расчет эмпрического квантиля заданного порядка по выборке;

2) сглаживание множества эмпирических квантилей по независимой переменной, зафиксированной для каждого рассчитанного квантиля, то соответствующий подход принято считать двухступенчатым.

На сегодняшний день наиболее известна реализация двухступечатого подхода при обработке исследований Американского центра по контролю за питанием (CDC) для построения справочных диаграмм развития детей. Для измеренных значений были получены первоначальные сглаженные кривые выбранных главных процентилей и на втором этапе получены параметры, которые были использованы для построения финальных сглаженных кривых и дополнительных процентилей. В качестве сглаживающих функций были использованы полиномиальная 5-й степени, локально взвешенная регрессия (locally weighted regression). Подгонка модели основывалась на минимизации остаточного среднего квадрата ошибок (RMSE), коэффициента детерминации (R2). Подробное описание вычислительных процедур можно найти на http://www.cdc.gov/growthcharts.

3. Анализ эффективности применения квантильно-регрессионных моделей в различных социально-экономических областях

Квантильная регрессия как статистический метод используется в различных социально-экономических областях.

В медико-социальном мониторинге широко используются справочные или стандартные показатели физического развития детей. Справочные показатели используются как на уровне популяции, так и на индивидуальном уровне. На уровне популяций справочные показатели могут помочь в оценке распространенности, определении причин заболевания и идентификации групп риска, при выборе направлений и оценке эффектов лечебного вмешательства [16].

На индивидуальном уровне справочные показатели роста являются эффективным инструментом для скрининга-диагностики и прогнозирования заболеваний у детей, связанных с развитием [17].

Справочные показатели физического развития детей представляют процентильными кривыми, соответствующие значениям 3-го, 10-го, 25-го 50-го, 75-го, 90-го, 97-го процентиля длины и массы тела, окружности головы здорового ребенка определенного пола для каждого возраста [18, 19]. Для детского организма характерна сильная возрастная зависимость показателей физического развития, поэтому применение квантильно-регрессионных моделей – оптимальный способ получения диагностических критериев. Методы квантильной регрессии были применены при разработке Британских национальных стандартов физического развития (LMS-метод) [0], национальным центром контроля за заболеваемостью и питания США (двухступенчатый метод) [18], стандартов Всемирной организации здравоохранения [17].

На сегодняшний день ведутся разработки процентильных кривых для других медико-биологических показателей, в частности, работы [20, 21], посвящены анализу артериального давления.

В финансовой сфере распространен VaR-анализ [22], позволяющий рассчитать величину риска в случаях, когда набор финансовых инструментов, составляющих портфель, не ограничивается 1-2 видами ценных бумаг. Цель VaR определить, что с заданной вероятностью за N дней возможна потеря не более p % от начального капитала. Выбор вероятности варьируется в пределах 95-99 %. Под убытком понимается отрицательный логарифм отношения капитала портфеля в некий момент времени t из интервала [0, N] к стоимости портфеля при t = 0, который рассматривается как случайная величина Y, зависящая, например, от валютного курса и т. п. Следовательно, для вычисления VaR необходимо знать характер поведения логарифмов изменений стоимости портфеля.

Если приращения логарифмов приемлемо описываются нормальным законом используется, параметрическая модель VaR. В реальности «хвосты» распределений логарифмов приращений оказываются более тяжелыми, чем у нормального распределения. Альтернатива использованию семейства гиперболических распределений – построение эмпирической функции распределения по архивным данным. Построение квантильно-регрессионных моделей актуально для решения задач портфельной оптимизации по критерию доходность – риск [23, 24].

В экономике труда и управлении персоналом широко применяются анализ и моделирование социально-трудовых показателей. Они позволяют выявить наиболее существенные черты изучаемого явления путем замещения объекта исследования его моделью, как правило, экономико-математической.

Размер заработной платы зависит от уровня квалификации работника, интенсивности труда, условий труда, а также отрасли, в которой занят работник, территориального размещения предприятий и организаций и других факторов. Поэтому квантильно-регрессионные модели очень эффективны. Наиболее часто цитируемая работа по изучению рынка – [5], где проведен анализ уровня доходов от различных факторов с использованием квантильной регрессии. В работе [25] исследовано распределение доходов в Соединенном Королевстве, в [26] сравнивается изменчивость в уровне заработных плат в США и Германии.

Информация о структуре доходов населения необходима для изучения спроса населения, поскольку объем и структура спроса существенно зависят от размера дохода той или иной группы населения. Поэтому при изучении потребительского рынка применение квантильно-регрессионных моделей позволяет учитывать социальную дифференциацию потребителей по полу, возрасту, уровню доходов, состоянию здоровья и образования и другим параметрам. Так в [27] получены зависимости спроса на электричество в зависимости от времени суток.

Квантильно-регрессионные модели дифференциции доходов могут быть использованы при расчетах средств, необходимых для повышения минимальной заработной платы и пенсий, тарифных ставок, размеров и методов налогообложения групп населения с различными уровнями доходов. Для осуществления целевой социальной политики, регулирования потребительского рынка необходим мониторинг и прогноз структуры потребления в зависимости от факторов социальной дифференциации, эффективным инструментом которого являются квантильно-регрессионные модели.

При проектировании сложных человеко-машинных систем остро встает вопрос о качестве такой системы. Как правило, для подобных систем проводится анализ риска, при котором получаются количественные показатели – вероятность наступления неблагоприятных событий и размер ущерба. На основе анализа статистических данных по неблагоприятным событиям, имевшим место в прошлом, определяются средние и предельные значения, которые являются квантилями эмпирических распределений. Так, для оценки максимального ущерба используют 95-й, 99-й и 99,9-й процентиль, что соответствует вероятности превышения максимально приемлемого уровня ущерба с частотой один раз в 20, 100 и 1000 лет [28]. В связи с этим применение квантильно-регрессионных моделей будет эффективно и при прогнозировании чрезвычайных ситуаций.

Использование квантильной регрессии наиболее часто встречается в зарубежных социальных исследованиях. Поскольку для построения квантильно-регрессионных моделей необходим большой объем данных, то в качестве исходных данных используются результаты национальных исследований здоровья (National Population Health Survey). Так в работе [29] для оценки взаимосвязи между состоянием здоровья и уровнем доходов использованы данные исследование здоровья Канадского населения. Чтобы учесть нелинейность и неоднородность данных была выбрана частично линейная полупараметрическая квантильно-регрессионная модель.

Другая область применения квантильной регрессии – задачи климатического мониторинга [30]. Меры по адаптации к изменениям климата должны строиться с учетом всего диапазона изменения метеовеличин, поэтому актуально использование не средних, а значений квантилей заданного порядка. На сегодняшний день для определения режимом работы систем отопления, вентиляции и кондиционирования используются 0,5-й, 6-й, 98-й процентиль суточных температур. Накопленные базы данных метеорологических переменных, представляют многомерные временные ряды, которые могут быть представлены квантильно-регрессионными моделями исследуемого района. Полученные результаты могут быть использованы не только для прогнозирования, но для формирования управленческих решений.

Методы регрессионного анализа могут быть использованы при решении различных практических задач [32-49].

Заключение

В работе рассмотрены особенности применения квантилей в задачах интеллектуального анализа данных. Проведен анализ методов построения квантильной регрессии. Дана оценка эффективности применения квантильно-регрессионных моделей в различных социально-экономических областях.

Литература

[1] Мостеллер Ф, Тьюки Дж. Анализ данных и регрессия. Вып. 1. – М.: Финансы и статистика. – 1982. – 319 с.

[2] Айвазян С.А., Енюков И.С., Мешалкин А.Д. Прикладная статистика. Исследование зависимостей. – 1983.

[3] Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация. Том. 1: - М.: Издательский дом «Вильяме», 2001. – 250 с.

[4] Koenker R. Quantile Regression / R. Koenker, K. F. Hallock // J. of Economic Perspectives – 2001. – Vol. 15. – P. 143–156.

[5] Buchinsky M. Quantile regression, Box-Cox transformation model, and U.S. wage structure, 1963-1987. – J. Econometr. – 1995. – V. 65 – P. 109-154.

[6] Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, 2003. – 686 с.

[7] Круг Г.К., Кабанов В.А., Фомин Г.А., Фомина Е.С. Планирование эксперимента в задачах нелинейного оценивания и распознавания образов. – М.: Наука, 1981. – 172 с.

[8] Математическая теория планирования эксперимента / Под. Ред. С.М. Ермакова. – М.: Наука, 1983. – 392 с.

[9] Васильев Ф.П. Численные методы решения экстремальных задач. – М.: Наука, 1980. – 518 с.

[10] Cole TJ Smoothing reference centile curves: the LMS method and penalized likelihood / TJ Cole, PJ. Green // Statistics in Medicine. – 1992. – Vol. 11. – P. 1305-1319.

[11] Yu. K. Smoothing regression quantile by combining k-NN estimation with local linear kernel fitting // Statist. Sinica. – 1999. – V. 9. – P. 759-774.

[12] Rigby RA Generalized additive models for location, scale and shape / RA Rigby, DM Stasinopoulos // Journal of the Royal Statistical Society, Series C – Applied Statistics. – Vol. 54. – P. 507–544.

[13] Yu. K. Local linear regression quintile estimation// Journal American Statist. Ass. – 1998. – V. 93. – P.228-238

[14] Cizek P. Quantile Regression/ XploRe Application Guide, ed. by W. Härdle, Z. Hlavka, and S. Klinke. – Springer, Berlin. – 2003. – P. 19-48.

[15] Buchinsky M. Recent Advances in Quantile Regression Models: A Practical Guidline for Empirical Research // Journal of Human Resources, 1998. – Vol. 33(1). – P. 88-126.

[16] Кучма В. Р. Оценка физического развития детей и подростков в гигиенической диагностике системы «Здоровье населения – среда обитания» / Р.В. Кучма. – М.: Издательство ГУНЦЗД РАМН, 2003. – 316 с.

[17] WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards: Length/height-for-age, Weight-for-age, Weight-for-length, Weight-for-height and Body mass index-for-age: Methods and Development. – Geneva: World Health Organization, 2006.

[18] Borghi E. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves / E Borghi, M. de Onis, C Garza, Van den Broeck J, EA Frongillo // Statistics in Medicine. – 2006. – Vol. 25. – P. 247–265.

[19] De Onis M. WHO Multicentre Growth Reference Study (MGRS): Rationale, Planning and Implementation / M de Onis, C Garza, CG Victora, MK Bhan, KR Norum // Food and Nutrition Bulletin. (Suppl. 1) – 2004.– P. 1-89.

[20] Jackson Lisa V. Blood pressure centiles for Great Britain / L.V. Jackson, Nandu KS Thalange, TJ Cole // Arch. Dis. Child. – 2007. – V.92. – P. 298-303.

[21] Минакова О.В. Оптимизация и рациональный выбор тактики определения отношения артериального давления с показателями физического развития / О. В. Минакова, А. С. Настаушева, В. П. Ситникова, Л. И. Стахурлова // Системный анализ и управление в биомедицинских системах: журнал практической и теоретической биологии и медицины. – М., 2008. – Т. 7. – №2. – С. 310-313.

[22] Волошин И. Оценка банковских рисков: Новые подходы. – Эльга, 2004 – 216 с.

[23] Basset G. W. Portfolio style: return-based attribution using quintile regression / G. W. Basset, H. Chen // Emp. Econ – 2001. – V. 26. – P. 293-305.

[24] Taylor J. A quantile regression approach to estimating the distribution of multiperiod returns. – J. Deriv. – 1999 – V. 24. – P.64-78.

[25] Gosling A. The Changing Distribution of Male Wages in the U.K. / A. Gosling, S. Machin, C. Meghir – Review of Economic Studies. – 2000. – Vol. 67, issue 4. – P. 635-66.

[26] Conley T. Nativity and Wealth in Mid-Ninete-enth-Centure Cities. / T. Conley, D. Galenson // J. of Economic History. – V. 58. – P. 468-493.

[27] Hendricks W. hierarchical splane models for conditional quantiles and the demand for electricity // W. Hendricks, R. Koenker – J. of Am. Stat. Assoc. – 1991 – V.87. – P. 58-68.

[28] Хохлов Н.В. Управление риском. – М., 1999.

[29] Sun Y. The absolute health income hypothesis revisited: a semiparametric quantile regression approach / Yiguo Sun // Empirical Economics. – 2008. – V. 35. – P. 395-412.

[30] Стерлин А.М. К учету тенденций изменения вариабельности и экстремальности климата в выработке стратегий адаптации / А.М. Стерлин, А.А. Тимофеев // Материалы международной конференции по глобальному изменению климата. – М., 2009.

[31] Koenker R. Conditional Quantile Estimation and Inference for ARCH Models // Koenker R., Zhao Q. – Econometric Theory, – 2002, V. 12. – P. 793-814.

[32] Klimenko G.Ya. Optimization of medical aid for pregnant women with iron deficiency anemia based on predictive modeling of their health state with due consideration of medical and social risk factors / G.Ya. Klimenko, S.A. Pyataeva, I.Ya. Lvovich, O.N. Choporov, N.V. Naumov. – Lorman, MS, USA. Science Book Publishing House, 2012. – 144 p.

[33] Амвросов Д.Э. Прогнозирование качества жизни больных после перенесенной травмы нижних конечностей и результатов их лечения / Д.Э. Амвросов, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 44. – № 1.3. – С. 383-388.

[34] Болгов С.В. Прогнозирование стоматологической заболеваемости по медико-биологическим и социально-гигиеническим факторам риска / С.В. Болгов, К.А. Разинкин, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 49. – № 6.2. – С. 294-301.

[35] Интегральное оценивание и прогностическое моделирование состояния здоровья беременных, рожениц и родильниц с учетом их медико-социальных характеристик / О.Н. Чопоров, В.П. Косолапов, Н.В. Наумов, Х.А. Гацайниева // Вестник Воронежского института высоких технологий. – 2012. – № 9. – С. 91-95.

[36] Клименко Г.Я. Индивидуальное прогнозирование заболеваемости туберкулезом органов дыхания по медико-социальным факторам риска / Г.Я. Клименко, В.А. Николаев, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2010. – Т. 9. – № 4. – С. 892-896.

[37] Моделирование и прогнозирование заболеваемости миомой матки в сочетании с аденомиозом по медико-социальным факторам риска / О.Н. Чопоров, Н.Н. Кудинова, М.В. Фролов, Г.Я. Клименко // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Kudinova_soavtori_3_13_1.pdf.

[38] Моделирование и прогнозирование качества жизни беременных женщин и пути его улучшения / В.И. Стародубов, Г.Я. Клименко, С.В. Говоров, Н.Б. Костюкова, О.Н. Чопоров. – Воронеж: Изд-во «Истоки», 2009. – 188 с.

[39] Прогнозирование развития онкологической заболеваемости по индивидуальным медико-социальным факторам риска / О.Н. Чопоров, А.И. Агарков, Г.Я. Клименко, Ю.Ю. Шуршуков // Моделирование, оптимизация и информационные технологии. - 2013. – № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Agarkov_soavtori_3_13_1.pdf.

[40] Прогнозирование удовлетворенности стоматологической помощью по медико-социальным факторам риска / С.В. Болгов, О.Н. Чопоров, Г.Я. Клименко, К.А. Разинкин // Системный анализ и управление в биомедицинских системах. – 2013. – Т. 12. – № 4. – С. 1171-1176.

[41] Разработка и использование моделей для прогнозирования качества жизни беременных по их медико-социальным характеристикам / Х.А. Махер, Н.В. Наумов, Г.Я. Клименко, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2011. – Т. 10. – № 4. – С. 789-793.

[42] Чембарцева Н.Я. Моделирование и прогнозирование состояние здоровья новорожденных по медико-социальным факторам риска / Н.Я. Чембарцева, Г.Я. Клименко. – Воронеж, 2006. – 133 с.

[43] Чопоров О.Н. Оптимизация функционирования медицинских систем на основе интегральных оценок и классификационно-прогностического моделирования: дисс.: д-ра техн. наук / О.Н. Чопоров. – Воронеж, 2001. – 325 С.

[44] Бережная Е.В. Оценка риска для здоровья населения г. Воронежа при воздействии химических веществ, загрязняющих атмосферный воздух / Е.В.Бережная // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – С. 2.

[45] Преображенский Ю.П. Оценка эффективности применения системы интеллектуальной поддержки принятия решений / Ю.П. Преображенский // Вестник Воронежского института высоких технологий. – 2009. – № 5. – С. 116-119.

[46] Львович Я.Е. Принятие решений в экспертно-виртуальной среде / Я.Е. Львович, И.Я. Львович. – Воронеж: Изд-во «Научная книга», 2010. – 139 с.

[47] Гафанович Е.Я. Прогнозирование исходов и выбор рационального лечения артериальной гипертензии с применением математических методов / Е.Я. Гафанович, И.Я. Львович // Вестник Воронежского государственного технического университета. – 2013. – Т. 9. – № 4. – С. 84-86.

[48] Калаев В.Н. Регрессионный анализ в биологических исследованиях / В.Н. Калаев, Е.А. Калаева, А.П. Преображенский, О.В. Хорсева // Системный анализ и управление в биомедицинских системах. – 2007. – Т. 6. – № 3. – С. 755-759.

[49] Преображенский Ю.П. Применение имитационно-семантического моделирования и полумарковских процессов принятия решений в клинической практике / Ю.П. Преображенский, Н.С. Преображенская // Вестник Воронежского института высоких технологий. – 2010. – № 6. – С. 83-89.

....

Doc. RNDr. Eugen Ruzicky, CSc.
Paneuropean University, Bratislava, Slovakia
eugen. ruzicky@paneurouni.com

Prof. Vera Kostrova, D. Sc.
Voronezh Institute of High Technologies

Methods of calculating of quantiles of various orders

Методы вычисления квантилей различного порядка

Y. E. Lvovich, A. G. Yurochkin
Я. Е. Львович, А. Г. Юрочкин

Abstract:

The methods of calculation of quantiles of various orders are considered in the work. The procedure for the definition of quantiles on selected theoretical distribution is introduced. It presents an algorithm for determining the distribution function, including the procedure of histograming, the choice of the formula for the empirical distribution function, the determination of the parameter estimates by the method of quantiles with the examples for various practical cases. The technique for the determination of quantiles approximated by the empirical distribution function based on nuclear estimation and interpolation by splines and orthogonal polynomials is presented.

Аннотация:

В работе рассмотрены методы вычисления квантилей различных порядков. Приведена процедура определения квантилей по подобранному теоретическому распределению, в рамках которой представлен алгоритм определения функции распределения, включающий процедуру построения гистограмм, выбора вида эмпирической функции распределения, определения оценок параметров методом квантилей с примерами для различных практических случаев. Представлена методика определения квантилей по аппроксимированной функции эмпирического распределения на основе ядерного оценивания, интерполяции сплайнами и ортогональными полиномами.

Keywords:

Distribution function, quantiles, histogram, approximation, interpolation, nuclear assessment, orthogonal polynomials.

Ключевые слова:
Функция распределения, квантили, гистограмма, аппроксимация, интерполяция, ядерное оценивание, ортогональные полиномы.

ACM Computing Classification System:
Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

Квантильная регрессия широко используемый статистический метод в эконометрике, в финансовых и биомедицинских исследованиях, при изучении окружающей среды и других прикладных областях. В связи с появлением новых вычислительных процедур, высокопроизводительных алгоритмов актуальность ее применения будет только расти. Во многих задачах необходимо проводить вычисление квантилей различного порядка, соответствующие методы рассмотрены в данной работе.

1. Определение квантилей по подобранному теоретическому распределению

Аппроксимация известным теоретическим распределением и вычисление квантилей теоретического распределения – достаточно распространенная задача. Если для данных можно правильно подобрать теоретическое распределение, то все другие характеристики могут быть вычислены с необходимой точностью, поскольку функциональное описание распределения дает самое полное представление о любого множестве выборочных данных.

Так как выбор вида функции распределения не поддается какой-либо формализации, вид закона выбирается по результатам визуального анализа эмпирической функции распределения или плотности распределения (например, по гистограмме) и оценке параметров распределения.

Алгоритм выбора теоретической функции распределения по экспериментальным данным включает следующие шаги:

1. Ввод опытных данных.
2. Составление вариационного ряда.
3. Построение гистограммы.
4. Выбор вида эмпирической функции распределения.
5. Оценка параметров выбранной функции распределения.
6. Проверка гипотезы о виде функции распределения.
7. Оценка точечных и интервальных значений параметров функции распределения.
8. Вывод функции распределения.

Важным вспомогательным средством при принятии гипотезы о виде функции распределения является гистограмма. Для ее построения к вариационному ряду наблюдаемых значений применяется следующий алгоритм.

Шаг 1. Размах вариационного ряда делится на k интервалов. Наиболее распространено вычисление числа интервалов по формуле Старджесса (Sturges): $k = 1 + 3,3\lg(N_0)$.

В задачах контроля качества для определения «оптимального» числа интервалов рекомендуют формулу Брукса и Каррузера [1]:

$$k = 5\lg(N_0).$$

При больших объемах выборок разброс значений k, задаваемых различными формулами, достаточно велик. Поэтому на практике при выборе числа интервалов руководствуются тем, чтобы в интервалы попадало число наблюдений не менее 5-10. В рекомендациях ВНИИМетрологии в зависимости от числа наблюдений – N_0 предлагают следующие значения k:

N_0	40–100	100–500	500–1000	1000–10000
k	7–9	8-12	10–16	12–22

Шаг 2. Для каждого из интервалов вычисляется значение частоты попаданий:

$$\overline{f}_i = \frac{n(\Delta x_i)}{\Delta x_i \cdot N_0},$$

где $n(\Delta x_i)$ – число членов вариационного ряда, попавших в i-й интервал,

Δx_i – ширина интервала, N_0 – общее число членов ряда (наблюдений).

Шаг 3. Графически гистограмма изображается рядом прямоугольников шириной Δt_i и высотой $\overline{f_i}$. Считается, что в гистограмме не должно быть пропусков интервалов и многочисленных инверсий высоты $\overline{f_i}$.

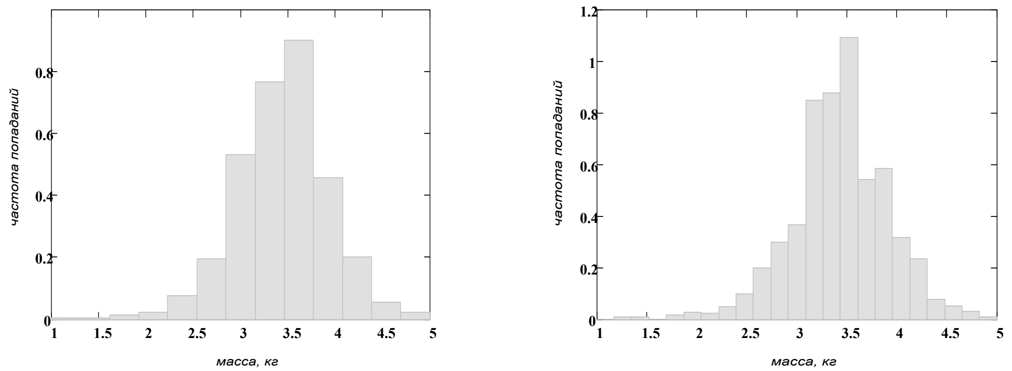
При построении гистограммы следует использовать несколько различных значений длины интервала и выбирать наименьшее, при котором гистограмма имеет более плавную форму.

Задача 1. Построение гистограммы массы тела при рождении мальчиков 1997-2000 года рождения на основании данных исследования физического развития [2]. Размер выборки 1558 наблюдений.

Результаты задачи 1. На рисунке 1 представлены гистограммы, построенные для различного числа интервалов.

При значительном увеличении числа интервалов наблюдалось появление множества изломов – инверсий, при выборе меньшего значения числа интервалов истинная форма оказывалась скрытой слишком широкими столбцами и пик распределения не выявлялся. Следует отметить, что в значительных пределах от 10 до 25 интервалов форма гистограммы в рассматриваемой задаче значительно не изменялась.

На рисунке 2 представлено сопоставление гистограммы с теоретическим распределением – нормальным и логнормальным, параметры которых оценены по выборке. После выбора подходящего вида распределения производится оценка его параметров.



а) при k=1+3.3lg(1558)=13 б) при k=23

Рис. 1. Пример построения гистограмм при задании различного числа интервалов (задача 1)

Для заданного семейства непрерывных распределений, от нормальных до гамма-распределений, существует несколько альтернативных способов определения плотности распределения вероятностей. Однозначность представления обеспечивает параметризация, т. е. задание его параметров – положения, масштаба и формы.

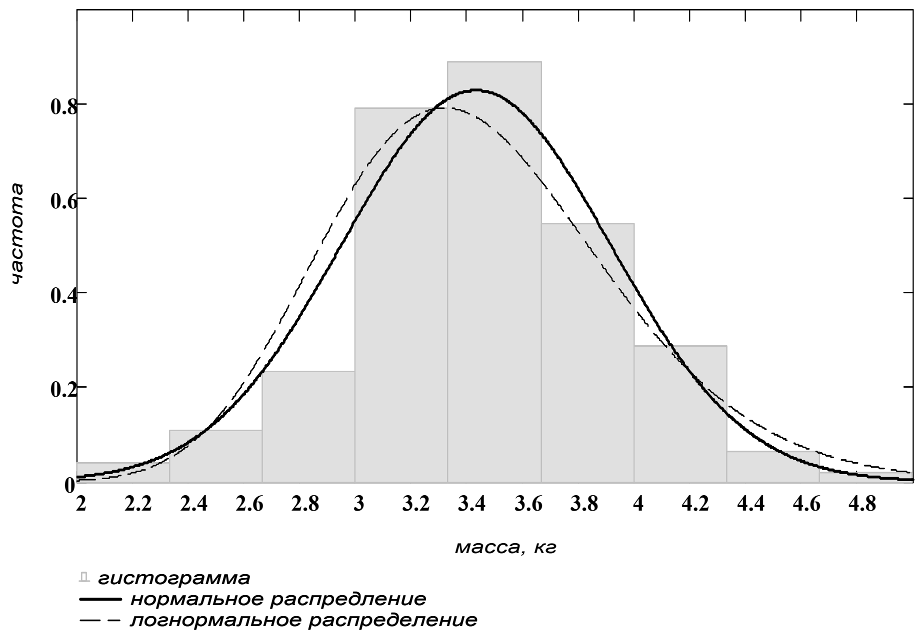
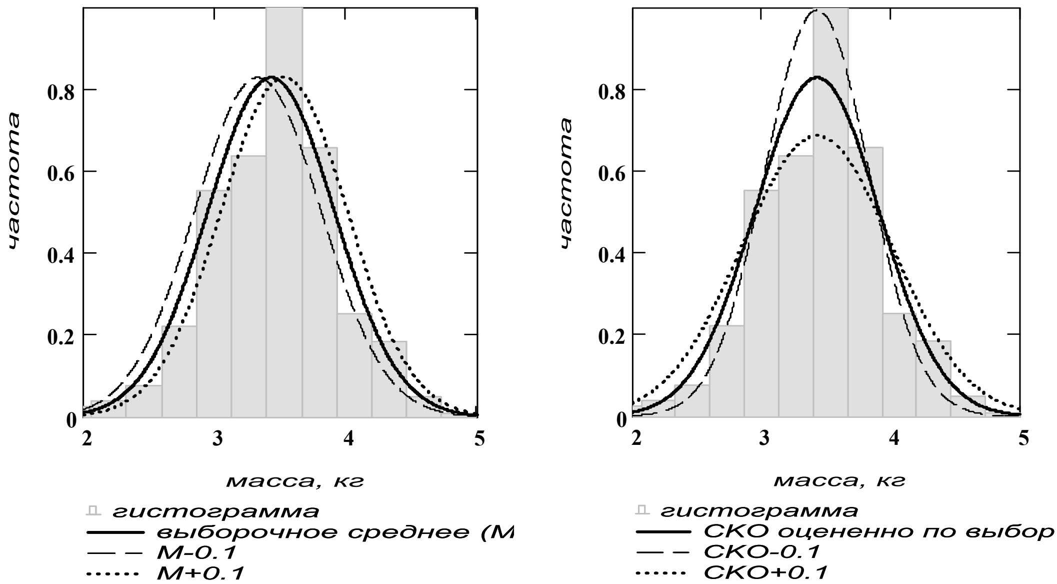


Рис. 2. Сравнение гистограммы с теоретическим распределением (задача 1)

Параметр положения определяет положение области распределения значений по оси абсцисс, обычно – это средняя (М для нормального распределения) или нижняя конечная точка области распределения (параметр сдвига). При изменении параметра положения (сдвига) соответствующее распределение просто сдвигается влево и вправо без каких-либо изменений.

Масштабный параметр определяет масштаб изменения значений в диапазоне распределений (σ – для нормального распределения). Изменяя параметр масштаба, соответствующее распределение можно сократить или увеличить без изменения его основной формы. На рисунке 3 представлен подбор нормального распределения с различными значениями положения (М) и масштаба (σ) для выборочных данных задачи 1.

Параметр формы определяет форму распределения в общем семействе распределений и его изменение приводит, как правило, к другому виду распределения данного семейства. С одной стороны оценить параметр формы по гистограмме нельзя, но с другой стороны, некоторые распределения имеют характерные значения этого параметра, которые однозначно указывают вид распределения. Пример представлен решением задачи 2.



а) изменение положения　　　б) изменение масштаба

Рис. 3. Подбор параметров распределения (задача 1)

Задача 2. Подбор теоретического распределения заболеваемости нефропатиями среди подростков в 32 районах Воронежской области и городе Воронеж на основании данных работы нефроцентра ВОДКБ за десятилетний период. Заболеваемость определялась как число новых случаев на 100 тыс. подросткового населения района.

Результаты задачи 2. На рисунке 4 представлена гистограмма распределения заболеваемости нефропатиями и функции плотности экспоненциального и логнормального распределения с параметрами, оцененными по выборочным данным.

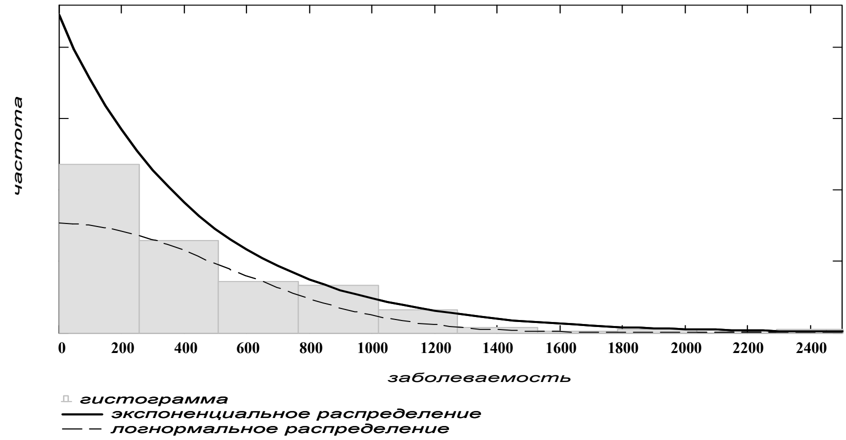


Рис. 4. Сопоставление гистограммы распределения заболеваемости с теоретическим распределением (задача 2)

Значение коэффициента асимметрии по экспериментальным данным составило 2,60 и коэффициента эксцесса – 10,4, что свидетельствует о показательном распределении данных, так как у теоретического экспоненциального распределения коэффициент асимметрии равен 2, а эксцесса 9.

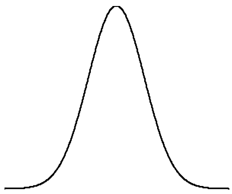
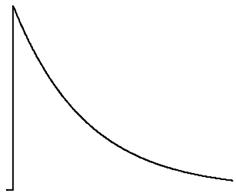
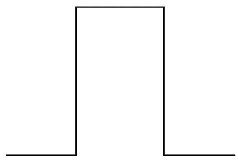
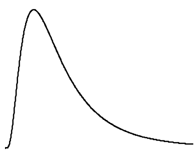
Для оценки параметров распределения используют методы максимального правдоподобия, моментов или квантилей.

Пример. Получение оценок максимального правдоподобия экспоненциально распределенной величины – показателя заболеваемости в регионе.

Для каждого j-го района запишем апостериорную вероятность некоторого значения заболеваемости: $f(x_j;\lambda)dx = \lambda \cdot e^{-\lambda \cdot x}dx$. Для наблюдаемых значений показателя заболеваемости в N районах $x_1, x_2, \ldots x_N$ функция правдоподобия $L = \prod_{j=1}^{N} f(x_j;\lambda)dx$ равна $L = \lambda^N \exp\left\{-\lambda \cdot \sum_{j=1}^{N} x_j\right\}$, так как $T = \dfrac{\sum_{j=1}^{N} x_j}{N}$ – среднее значение выборочных наблюдений, то $L = \lambda^N \exp\left\{-\lambda \cdot T \cdot N\right\}$. Логарифм функции правдоподобия $\ell = \ln(L) = N\ln(\lambda) - \lambda \cdot T \cdot N$. Производная от логарифма равна: $\dfrac{d\ell}{d\lambda} = N \cdot \left(\dfrac{1}{\lambda} - T\right)$. Уравнение правдоподобия принимает вид: $N \cdot \left(\dfrac{1}{\lambda} - T\right) = 0$. Оно имеет решение $\hat{\lambda} = \dfrac{\sum_{j=1}^{N} x_j}{N}$. Следовательно, оценка максимума правдоподобия равна обратной величине от среднего значения результатов заболеваемости по районам.

В отечественной практике оценки максимального правдоподобия для параметров наиболее распространенных распределений регламентированы государственными стандартами. В таблице 1 приведены формулы для расчета оценок широко распространенных распределений по выборочным данным.

Таблица 1. Оценки параметров распространенных распределений

Функция плотности распределения $f(x)$	График функции $f(x)$	Оценки максимального правдоподобия
нормальное $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$		$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i$ $\hat{\sigma}=\sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2}$
экспоненциальное $\lambda\cdot\begin{cases}\exp(-\lambda\cdot x), & x\geq 0\\ 0, & x<0\end{cases}$		$\hat{\lambda}=\frac{N}{\sum_{i=1}^{N}x_i}$
равномерное $\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & x<a, x>b\end{cases}$		$\hat{a}=\min_{1\leq i\leq N}x_i$ $\hat{b}=\max_{1\leq i\leq N}x_i$
логнормальное $\frac{x^{-1}}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$		$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}\ln x_i$ $\hat{\sigma}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln x_i-\hat{\mu})^2}$

Важным этапом при выборе вида распределения является проверка согласия выбранного теоретического распределения с экспериментальными данными.

Критерии согласия применяются, чтобы формально оценить, являются ли наблюдения независимой выборкой из предполагаемого распределения. Наиболее распространен критерий согласия – χ^2 Пирсона. При использовании этого критерия область значений делится на k прилежащих интервалов. Мера расхождения теоретического и эмпирического распределения или статистика критерия определяется как [1]:

$$H=\sum_{i=1}^{k}\frac{(n_i-N_0p_i)^2}{N_0p_i}, \qquad (1)$$

где N_0 – общее число наблюдений;

n_i – число наблюдений в i-м интервале,

p_i – теоретическая частота попаданий в i-й интервал.

Вычисленная по (1) мера расхождения H есть случайная величина, имеющая χ^2-распределение с числом степеней свободы r=k-1-S. Значение S – число параметров проверяемого закона распределения, для нормального и логнормального закона S=2, экспоненциального S=1.

Сравнение вычисленной меры расхождения H с квантилью χ^2-распределения по уровню 1-α с r степенями свободы, где α – допустимая вероятность ошибки, дает основание отвергнуть (H>χ^2(1- α,r)) или принять проверяемую гипотезу о соответствии анализируемых данных нормальному распределению. Следует отметить, что χ^2-критерий действителен, т. е. имеет уровень α только асимптотически при $n \to \infty$ [3].

Критерий Колмогорова-Смирнова лишен указанных недостатков, так как обеспечивает сравнение функции эмпирического распределения с функцией теоретического и не требует никакого группирования данных и как следствие – потери информации. Преимущество его еще и в том, что он точно достоверен для любого объема выборки, но только если известны все параметры теоретического распределения [4, 5].

Статистика Колмогорова-Смирнова формально определяется [6]:

$$D_n = \sup\left\{\left|F_n(y) - \hat{F}_n(y)\right|\right\},$$

где $F_n(y)$ – эмпирическая функция распределения,

$\hat{F}_n(y)$ – оцениваемая функция распределения.

Вычисление супремума осуществляется с применением алгоритма сортировки данных по формуле [6]:

$$D_n = \max\left\{\max_{1\le i\le n}\left[\frac{i}{n} - F(z_i)\right], \max_{1\le i\le n}\left[F(z_i) - \frac{i-1}{n}\right]\right\}, \qquad (2)$$

где $F(z_i)$ – значение функции подобранного распределения в точке z_i.

В зависимости от вида распределения имеются модифицированные формы критерия Колмогорова-Смирнова, позволяющие осуществлять проверку согласия без априорного знания параметров распределения.

Результаты задачи 2. В результате анализа формы распределения по гистограмме, а также ее параметров формы было сделано предположение об экспоненциальном характере распределения заболеваемости нефропатиями в районах.

С помощью преобразования $z_i = \dfrac{y_i - \mu}{\upsilon}$, где $\upsilon = \dfrac{n\cdot(\bar{x} - x_1)}{n-1}$ и $\mu = x_1 - \dfrac{\upsilon}{n}$ для упорядоченной выборки $x_1, x_2, \ldots x_N$ переходим к нормированному экспоненциальному преобразованию, т. е. с параметром, равным 1.

Поскольку параметр априорно известен, то может быть использован критерий Колмогорова-Смирнова в виде (2.2), где $F(z_i) = 1 - e^{-z_i}$ – функция экспоненциального распределения с параметром, равным 1.

В результате расчетов значение D_n=0,186, после применения поправки Стефенса, устраняющей зависимость процентных точек распределений Колмогорова-Смирнова от объема выборки n D_n=0,618. Из таблиц процентильных точек [7] для уровня значимости α=0,05 критическое значение 1,224 и $D_n < 1,224$, гипотеза о экспоненциальности распределения заболеваемости в районах не отклоняется.

Следует отметить, что в случае годовой выборки заболеваемости по районам, когда ее размер был меньше 50 (n=33), значение D_n изменялось в пределах от 0,318 до 0,816 за каждый наблюдаемый год и для уровня значимости α=0,05 гипотеза также не отклонялась, что позволяет считать и ежегодное распределение заболеваемости нефропатиями в регионе экспоненциальным.

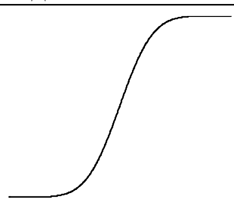
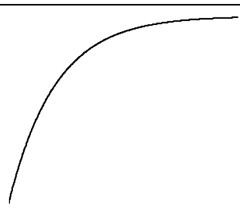
<u>*Пример.*</u> Получение обратной функции экспоненциально распределенных случайных значений показателя t – заболеваемость в регионе.

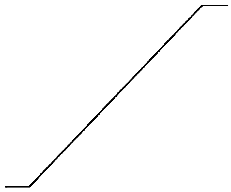
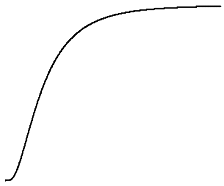
Плотность распределения имеет вид: $f(t) = \begin{cases} \exp(-\lambda \cdot t), & t \geq 0 \\ 0, & t < 0 \end{cases}$, функция распределения имеет вид: $F(t) = \dfrac{1}{\lambda}\int_0^t f(t)dt = 1 - e^{-\lambda \cdot t}$. Если $x = F(t)$, то для получения обратной функции $t = F^{-1}(x)$ необходимо выразить t, следовательно, из соотношения $x = 1 - e^{-\lambda \cdot t}$ определить t, т. е. $t = -\lambda \cdot \ln(1 - x)$.

Использование этого подхода лежит в основе работы большинства генераторов случайных чисел с заданным законом распределения, реализующих преобразования равномерно распределенных чисел в заданное распределение с использованием формулы для вычисления квантилей теоретического распределения.

В таблице 2 приведены формулы для вычисления квантилей наиболее распространенных распределений.

Таблица 2. Квантили распространенных распределений

Функция распределения	Вид	Квантили
нормальное $F(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$		$\mu + \sigma \cdot z^{-1}(\tau)$
экспоненциальное $F(x) = \begin{cases} 1 - \exp(-\lambda \cdot x), & x \geq 0 \\ 0, & x < 0 \end{cases}$		$\dfrac{1}{\lambda}\ln\left(\dfrac{1}{1-\tau}\right)$

равномерное $F(x)=\begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & a<x<b \\ 1 & x \ge b \end{cases}$		$a+(b-a)\cdot\tau$
логнормальное $F(x)=$ $\dfrac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$		$\mu+\sigma\cdot z^{-1}(\tau)$

2. Определение квантилей по аппроксимированной функции эмпирического распределения

Альтернативой, позволяющий учесть все характерные особенности формы эмпирического распределения, является аппроксимация эмпирического распределения по гистограмме. В этих целях используется ядерное оценивание, интерполяция сплайнами и ортогональными полиномами.

В технических задачах распространена непараметрическая (ядерная) оценка плотности распределения в заданной точке, известная как оценка Розенблатта-Парзена. Гистограмма может быть представлена как:

$$\overline{f}(x)=\frac{1}{\Delta x\cdot N}\sum_{i=1}^{k} I(x,x_0+i\cdot\Delta x),$$

где $I(x,x_0+i\cdot\Delta x)=\begin{cases} 1, & (x_0+i\cdot\Delta x)\le x<(x_0+(i+1)\cdot\Delta x) \\ 0, & x<(x_0+i\cdot\Delta x), x\ge(x_0+(i+1)\cdot\Delta x) \end{cases}$ – индикаторная функция.

Замена индикаторной функции на ядро, представляющая симметричную взвешивающую функцию, позволяет оценить плотность $w_1(y)$ выборки n независимых наблюдений $\{y_1, y_2 ... y_n\}$ следующей функцией:

$$w_1(y)=\frac{1}{n\cdot h(n)}\cdot\sum_{i=1}^{n} K\left[\frac{y-y_i}{h(n)}\right], \tag{3}$$

где K(y) – произвольное ядро аппроксимации,

h(n) – масштаб ядра или ширина окна аппроксимации.

Ядерная функция K(u) – неотрицательная и удовлетворяет условиям:

$$0<K(u)<\infty,\ \lim_{y\to\pm\infty} uK(u)=0,\ \int_{-\infty}^{\infty} K(u)du=1.$$

Цель ядра – обеспечить гладкость и дифференцируемость результирующей оценки. В качестве критерия оценки обосновано использование минимизации интегральной среднеквадратической ошибки.

Зачастую выбор ядра определяется вычислительной сложностью. На практике наиболее распространено нормальное (гауссово) ядро:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2 / 2) .$$

Хотя ядерная функция важна, проведение качественного непараметрического оценивания в основном зависит от выбора подходящего для поставленной задачи масштаба ядра h(n), который зависит от числа наблюдений, и должно удовлетворять условию: $\lim_{n\to\infty} h(n) = 0$.

На практике в качестве стартовой используется универсальная формула выбора оптимальной ширины окна, известная как правило Сильвермана:

$$h(n) = 1,364 \left(\frac{R_K}{\sigma_K^4} \right)^{1/5} \sigma \cdot n^{-1/5} ,$$

где σ – выборочное стандартное отклонение,

$\sigma^2_K = \int u^2 K(u) du$,

$R_K = \int K(u)^2 du$ – константы, зависящие от выбора ядра.

Для гауссового ядра: $h(n) = 1.059\sigma \cdot n^{-1/5}$.

Как показывают результаты эксперимента, оптимальная ширина окна обеспечивает гладкость оцениваемой функции и близость к гистограмме, полученной для числа интервалов по формуле Старджесса. На рисунке 5 представлены результаты ядерного сглаживания с гауссовым ядром данных задачи 1 при различной ширине окна.

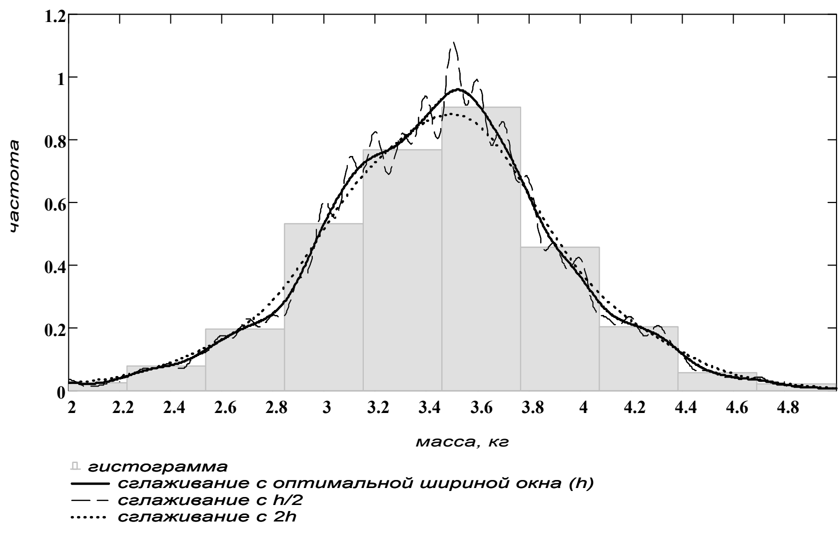


Рис. 5. Результаты ядерного сглаживания с различной шириной окна (задача 1)

Результаты сглаживания ядрами второго порядка практически совпадают с гауссовым ядром при использовании оптимальной ширины окна сглаживания, использования треугольного ядра значительно нарушает гладкость оцениваемой функции. На рисунке 6 представлены результаты ядерного сглаживания с различными ядерными функциями данных задачи 1 при оптимальной для гауссового ядра ширине окна.

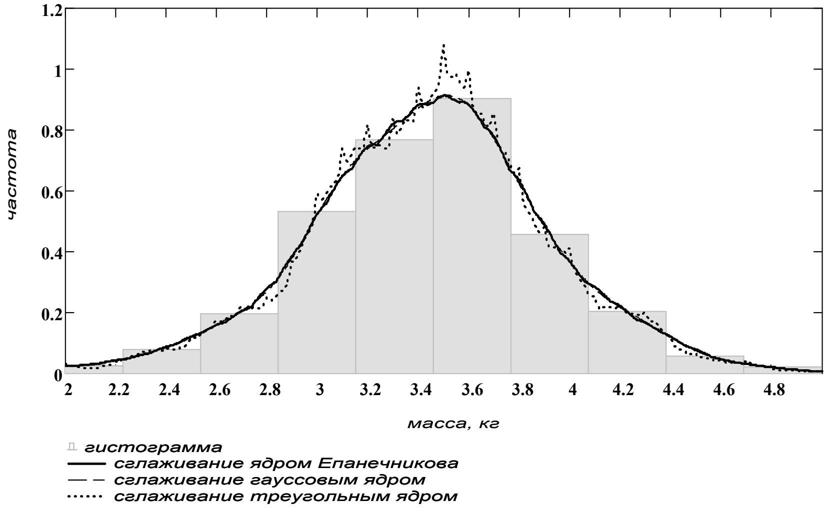


Рис. 6. Результаты ядерного сглаживания с оптимальной шириной окна и различными ядрами (задача 1)

Таким образом, свойства ядерных оценок зависят как от выбора вида ядра, так и от параметра сглаживания h(n). И как следствие оценка квантилей также будет зависеть от этих параметров.

Другим локальным непараметрическим методом оценки функции и плотности распределения является оценка ближайших соседей, ее существенным преимуществом перед ядерным оцениванием считается наличие оценки в любом случае. Так если необходимо получить оценку функции распределения в точке $\hat{y}$ по выборке n независимых наблюдений $\{y_1, y_2 ... y_n\}$, то можно увеличить интервал вблизи точки $\hat{y}$, так чтобы в него попадало k наблюдений. Тогда оценка плотности определяется как:

$$\hat{w}(y) = \frac{k}{n \cdot 2d_k(\hat{y})},$$

где k – число наблюдений в интервале $\left[\hat{y} - d_k(\hat{y}), \hat{y} + d_k(\hat{y})\right]$, т. е. ближайших соседей.

Целое число k контролирует степень сглаживания и должно быть много меньше размера выборки, обычно выбирают $k \approx \sqrt{n}$.

Поэтому используется комбинирование этих методов и оценка определяется как:

$$\hat{w}(y) = \frac{1}{n \cdot d_k(y)} \cdot \sum_{i=1}^{n} K\left[\frac{y - y_i}{d_k(y)}\right] \quad (4)$$

и рассматриваться как ядерное оценивание с шириной окна $d_k(y)$. Это позволяет управлять сглаживанием посредством целого k, используя при оценке любую точку наблюдения вблизи оцениваемой.

Некоторой разновидностью этого подхода можно считать ядерное оценивание с переменным ядром:

$$\hat{w}(y) = \frac{1}{n} \cdot \sum_{j=1}^{n} \frac{1}{h \cdot d_{j,k}} K\left[\frac{y - y_j}{h \cdot d_{j,k}}\right], \quad (5)$$

где K – ядерная функция,

k – положительное целое число,

$d_{j,k}$ – расстояние от точки y_j до k ближайших соседних точек на множестве включенных других n-1 точек данных,

h – сглаживающий параметр.

Оценка неизвестного закона распределения может также быть получена в виде рядов ортонормированных полиномов:

$$w(x) = \sum_{i=0}^{K} \beta_i \cdot \phi_i(x),$$

где $\phi_k(x)$ – система ортогональных функций, для которых:

$$\frac{1}{l} \sum_{i=1}^{n} \varphi_p(x_i)\varphi_q(x_i) = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases}.$$

Коэффициенты β определяются с помощью стандартных приемов линейной алгебры минимизацией остаточной суммы квадратов:

$$\beta = (\Phi_p^T \Phi_p)^{-1} \Phi_p^T Y,$$

где Y – вектор значений $y_1, ..., y_n$;

Φ_p – матрица, определяемая $\Phi_p = \begin{vmatrix} \phi_1(x_1) & ... & \phi_p(x_1) \\ ... & ... & ... \\ \phi_1(x_n) & ... & \phi_p(x_n) \end{vmatrix}$.

До k=6 решение легко может быть получено численными методами, для определения коэффициентов полиномов старших степеней требуется поиск специальных решений ввиду сильной обусловленности матрицы Φ_p.

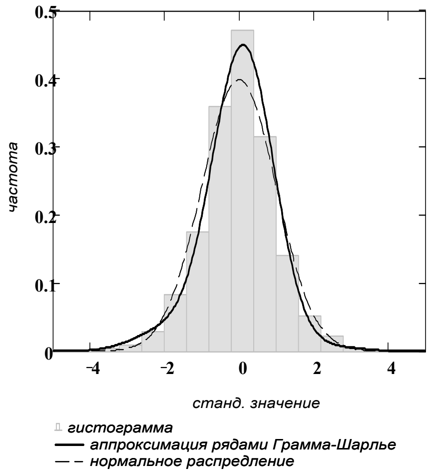
Для распределений, незначительно отличающихся от нормального, может использоваться разложение с помощью рядов Грамма-Шарлье [8]:

$$f(t) = \varphi(x) + \frac{b1}{6}\varphi'''(x) + \frac{b2 - 3}{24}\varphi''''(x), \quad (6)$$

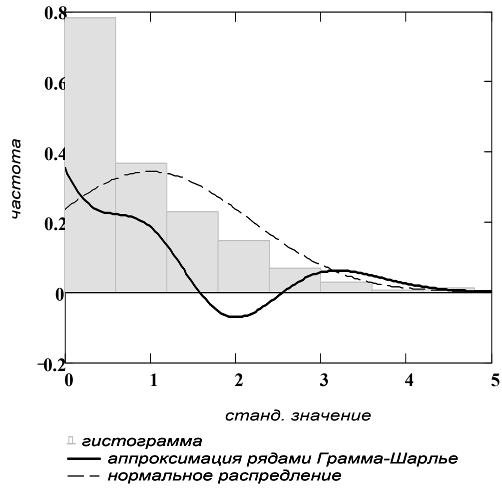
где $\varphi(x)$ – плотность стандартного нормального распределения;

b1 и b2 – коэффициенты асимметрии и эксцесса.

На рисунке 7 представлена аппроксимация распределения данных задачи 1 рядами Грамма-Шарлье, значение коэффициента асимметрии было отрицательное - 0,33, а коэффициент эксцесса равен 1,13. Для сравнения представлена аппроксимация распределения данных заболеваемости в районах задачи 2, для которых асимметрия составляла 2,60 и эксцесс – 10,4.



а) аппроксимации распределения массы тела при рождении (задача 1)



а) аппроксимации заболеваемости нефропатиями (задача 2)

Рис. 7. Аппроксимация распределений данных задач 1 и 2 рядами Грамма-Шарлье

Следует отметить важность оценки параметров формы – коэффициентов асимметрии (β) и эксцесса (γ) при выборе метода оценки функции распределения.

Так, равномерное распределение имеет очень высокое значение эксцесса – 1,8 при нулевой асимметрии. При экспоненциальном распределении коэффициент асимметрии постоянен и равен 2, а коэффициент эксцесса 9. Полученные по выборке задачи 2 оценки третьего и четвертого моментов очень близки, и форма гистограммы сильно похожа на кривую экспоненциального распределения.

Следует отметить, что для распределений, отличающихся от нормального, ряд Грамма-Шарлье может вести себя нерегулярно, т. е. при увеличение количества членов ряда наблюдается снижение точности аппроксимации. При коэффициенте асимметрии превышающем 0,7 сумма конечного числа членов ряда приводит к отрицательным значениям аппроксимирующих функций, особенно на краях распределений. Появление выбросов недопустимо, так как вычисление квантили по аппроксимации в ряд осуществляется численными методами.

Вычисление квантили q(τ) для функции распределения F(x) при заданной вероятности τ эквивалентно нахождению нуля функции:

$$\delta(x) = \tau - F(x). \qquad (7)$$

Эта задача решается в два этапа. На первом определяется интервал (x_0, x_1), на котором эта функция обращается в нуль. На втором этапе этот интервал уменьшается до тех пор, пока его длина не станет меньше заданной погрешности вычисления ε.

Функция F(x) монотонная по определению, то и δ(x) тоже монотонна, следовательно, можно воспользоваться специальной функцией MathCad root(δ (x), x). Точность вычисления определяется специальной константой TOL=ε. В функции root могут быть использованы необязательный параметры a и b, которые задают интервал поиска корней уравнения, q=root(δ (x), x, [a, b]).

Решение уравнения $\delta(q) = \tau - \int_{-\infty}^{q} f(y)dy$ по аппроксимированной плотности распределения численными методами не обеспечивает хорошей сходимости итерационного процесса.

Предложенные методы статистического анализа данных могут быть использованы в различных практических приложениях [9-34].

Заключение

В работе проведено определение квантилей по подобранному теоретическому распределению. Даны оценки параметров распространенных распределений. Показано, каким образом определяются квантили по аппроксимированной функции эмпирического распределения.

Литература

[1] Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир. – 1970. – 368 с.

[2] Физическое развитие (рост, масса) детей Воронежской области / Под ред. В.Н. Пенкина. – Воронеж, 2000. – 41 с.

[3] Кендалл М., Стьюарт А. Теория распределений. – М.: Наука. – 1966. – 587 с.

[4] Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ. – 2006. – 816 с.

[5] Орлов А.И. Практическая статистика. – М.: Экзамен. – 2006. – 312 с.

[6] Шурыгин А.М. Прикладная статистика: робастность, оценивание, прогноз. – М.: Финансы и статистика. – 2000. – 360 с.

[7] ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. – М.: Изд-во стандартов. 2002. – 30 с.

[8] Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука. – 1978. – 832 с.

[9] Klimenko G.Ya. Optimization of medical aid for pregnant women with iron deficiency anemia based on predictive modeling of their health state with due consideration of medical and social risk factors / G.Ya. Klimenko, S.A. Pyataeva, I.Ya. Lvovich, O.N. Choporov, N.V. Naumov. – Lorman, MS, USA. Science Book Publishing House, 2012. – 144 p.

[10] Амвросов Д.Э. Прогнозирование качества жизни больных после перенесенной травмы нижних конечностей и результатов их лечения / Д.Э. Амвросов, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 44. – № 1.3. – С. 383-388.

[11] Анализ динамики и прогнозирование распространенности дерматозов среди населения воронежской области / Е.Н. Бугакова, Г.Я. Клименко, О.Н. Чопоров, Г.В. Сыч // Врач-аспирант. – 2010. – Т. 43. – № 6.2. – С. 259-267

[12] Болгов С.В. Прогнозирование стоматологической заболеваемости по медико-биологическим и социально-гигиеническим факторам риска / С.В. Болгов, К.А. Разинкин, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 49. – № 6.2. – С. 294-301.
[13] Интегральное оценивание и прогностическое моделирование состояния здоровья беременных, рожениц и родильниц с учетом их медико-социальных характеристик / О.Н. Чопоров, В.П. Косолапов, Н.В. Наумов, Х.А. Гацайниева // Вестник Воронежского института высоких технологий. – 2012. – №9. – С. 91-95.
[14] Классификация районов белгородской области по распространенности злокачественных новообразований и результаты краткосрочного прогнозирования / А.И. Агарков, Г.Я. Клименко, О.Н. Чопоров, Ю.Ю. Шуршуков // Системный анализ и управление в биомедицинских системах. – 2013. – Т. 12. – № 4. – С. 1134-1138.
[15] Клименко Г.Я. Индивидуальное прогнозирование заболеваемости туберкулезом органов дыхания по медико-социальным факторам риска / Г.Я. Клименко, В.А Николаев, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2010. – Т. 9. – № 4. – С. 892-896.
[16] Методы предварительной обработки информации при системном анализе и моделировании медицинских систем / О.Н. Чопоров, Н.В. Наумов, Л.А. Куташова, А.И. Агарков // Врач-аспирант. – 2012. – Т. 55. – № 6.2. – С. 382-390.
[17] Моделирование и прогнозирование заболеваемости миомой матки в сочетании с аденомиозом по медико-социальным факторам риска / О.Н. Чопоров, Н.Н. Кудинова, М.В. Фролов, Г.Я. Клименко // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Kudinova_soavtori_3_13_1.pdf.
[18] Моделирование и прогнозирование качества жизни беременных женщин и пути его улучшения / В.И. Стародубов, Г.Я. Клименко, С.В. Говоров, Н.Б. Костюкова, О.Н. Чопоров. – Воронеж: Изд-во «Истоки», 2009. – 188 с.
[19] Оптимизация управления функционированием медицинских систем различного уровня / О.Н. Чопоров, И.Я. Львович, К.А. Разинкин, А.А. Рындин // Системы управления и информационные технологии. – 2013. – Т. 53. – №3. – С. 100-104.
[20] Прогнозирование изменения течения беременности по медико-социальным факторам риска / В.П. Косолапов, П.Е. Чесноков, Г.Я. Клименко, О.Н. Чопоров [и др.] // Врач-аспирант. – 2011. – Т. 44. – № 1.4. – С. 572-578.
[21] Прогнозирование развития онкологической заболеваемости по индивидуальным медико-социальным факторам риска / О.Н. Чопоров, А.И. Агарков, Г.Я. Клименко, Ю.Ю. Шуршуков // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Agarkov_soavtori_3_13_1.pdf.
[22] Прогнозирование удовлетворенности стоматологической помощью по медико-социальным факторам риска / С.В. Болгов, О.Н. Чопоров, Г.Я. Клименко, К.А. Разинкин // Системный анализ и управление в биомедицинских системах. – 2013. – Т. 12. – № 4. – С. 1171-1176.
[23] Разработка и использование моделей для прогнозирования качества жизни беременных по их медико-социальным характеристикам / Х.А. Махер, Н.В. Наумов, Г.Я. Клименко, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2011. – Т. 10. – № 4. – С. 789-793.

[24] Разработка классификационно-прогностических моделей развития гнойно-септических осложнений у родильниц / А.В. Чернов, В.Ю. Бригадирова, О.Н. Чопоров, В.И. Чернов // Системный анализ и управление в биомедицинских системах. – 2012. – Т. 11. – № 1. – С. 261-266.

[25] Химина И.Н. Рационализация управления медицинской помощью больным с заболеваниями желудка и двенадцатиперстной кишки на основе комплексного мониторинга и классификационно-прогностического моделирования / И.Н. Химина, В.Н. Эктов, О.Н. Чопоров. – Воронеж: Изд-во «Научная книга», 2014. – 181 с.

[26] Хопина О.А. Прогнозирование заболеваемости глаукомой по индивидуальным медико-социальным факторам риска / О.А. Хопина, Г.Я. Клименко, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 49. – № 6.1. – С. 203-208.

[27] Чембарцева Н.Я. Моделирование и прогнозирование состояние здоровья новорожденных по медико-социальным факторам риска / Н.Я. Чембарцева, Г.Я. Клименко. – Воронеж, 2006. – 133 с.

[28] Чопоров О.Н. Оптимизация функционирования медицинских систем на основе интегральных оценок и классификационно-прогностического моделирования: дисс.: д-ра техн. наук / О.Н. Чопоров. – Воронеж, 2001. – 325 С.

[29] Бережная Е.В. Оценка риска для здоровья населения г. Воронежа при воздействии химических веществ, загрязняющих атмосферный воздух / Е.В. Бережная // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – С. 2.

[30] Преображенский Ю.П. Оценка эффективности применения системы интеллектуальной поддержки принятия решений / Ю.П. Преображенский // Вестник Воронежского института высоких технологий. – 2009. – № 5. – С. 116-119.

[31] Львович Я.Е. Принятие решений в экспертно-виртуальной среде / Я.Е. Львович, И.Я. Львович. – Воронеж, Изд-во «Научная книга», 2010. – 139 с.

[32] Гафанович Е.Я. Прогнозирование исходов и выбор рационального лечения артериальной гипертензии с применением математических методов / Е.Я. Гафанович, И.Я. Львович // Вестник Воронежского государственного технического университета. – 2013. – Т. 9. – № 4. – С. 84-86.

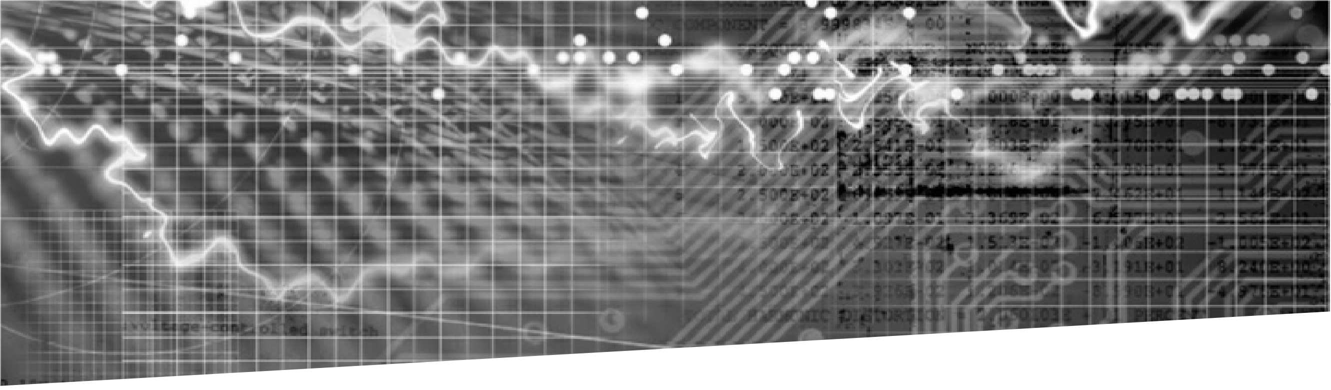
[33] Калаев В.Н. Регрессионный анализ в биологических исследованиях / В.Н. Калаев, Е.А. Калаева, А.П. Преображенский, О.В. Хорсева // Системный анализ и управление в биомедицинских системах. – 2007. – Т. 6. – № 3. – С. 755-759.

[34] Преображенский Ю.П. Применение имитационно-семантического моделирования и полумарковских процессов принятия решений в клинической практике / Ю.П. Преображенский, Н.С. Преображенская // Вестник Воронежского института высоких технологий.– 2010. –№ 6. – С. 83-89.

…………………………………………………………………………………………...

Prof. Yakov Lvovich, D. Sc.
the honored scientist of the Russian Academy of Natural Sciences
President of Voronezh Institute of High Technologies
office@vivt.ru

Prof. Anatoly Yurochkin, D. Sc.
The Voronezh branch of the Russian Academy of state service when
the President of the Russian Federation
kafec@vrn.ranepa.ru



Parametric quantile-regression models and those similar to them

Параметрические и близкие к ним квантильно-регрессионные модели

I. Y. Lvovich, Y. E. Lvovich
И. Я. Львович, Я. Е. Львович

Abstract:

The parametric approach to the construction of quantile regression models based on the selection and evaluation of parameters of the functional dependence of the moments of the first and second order distribution of the dependent variable from the values of the independent variable is considered. The mathematical and algorithmic maintenance of the procedures for the construction of parametric models is described. Comparative analysis of the statistical validation criteria distribution for normality is carried out. The information criteria that ensure selection of the optimal number of model parameters are given. The possibility of using the parametric quantile regression models for charting the age dynamics of the growth of children is considered.

Аннотация:

Рассматривается параметрический подход к построению квантильно-регрессионных моделей, основанный на выборе и оценке параметров функциональной зависимости моментов первого и второго порядка распределения зависимой переменной от значения независимой переменной. Описано математическое и алгоритмическое обеспечение процедур построения параметрических моделей. Проведен сравнительный анализ статистических критериев проверки распределения на нормальность. Приведены информационные критерии, позволяющие обеспечивать выбор оптимального числа параметров модели. Рассмотрена возможность использования параметрических квантильно-регрессионных моделей для построения диаграмм возрастной динамики роста детей.

Key words:

Parametric models, regression analysis, error, criterion, growth curves.

Ключевые слова:

Параметрические модели, регрессионный анализ, ошибка, критерий, кривые роста.

ACM Computing Classification System:

Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

В настоящее время регрессионные модели имеют большое многообразие форм, степеней сложности и возможностей применения для решения теоретических и прикладных задач. В основе квантильных статистических регрессионных моделей различных структур и процессов лежит широкое использование математической техники условных медиан и квантилей многомерных вероятностных распределений. Требуется определять информационные критерии, которые позволяют давать выбор оптимального числа параметров модели.

1. Математическое и алгоритмическое обеспечения параметрических моделей

Наиболее часто при анализе данных допускается нормальное распределение показателей. Теоретической предпосылкой для этого является центральная предельная теорема Ляпунова, утверждающая, что распределение суммы независимых случайных величин с любым исходным распределением будет нормальным, если число слагаемых достаточно велико, а вклад каждого в сумму мал [1].

При предположении о нормальности распределения квантиль заданного порядка $\tau \in [0,1]$ может быть вычислена как [2]:

$$Q_y(\tau) = M(x) + \sigma(x) \cdot \Phi^{-1}(\tau)\,, \qquad (1)$$

где $\hat{M}(x)$ – условное среднее значение;

$\hat{\sigma}(x)$ – условное СКО;

$\Phi^{-1}(\tau)$ – обратная функция N(0, 1).

Таким образом, в основе параметрического подхода к построению квантильно-регрессионных моделей лежит выбор и оценка параметров функциональной зависимости моментов первого и второго порядка распределения зависимой переменной y от значения независимой переменной x. Эти зависимости в статистике известны как регрессионная $\mathrm{M(x)}$ и скедастическая $\sigma^2(\mathrm{x})$.

На практике наиболее распространенная задача – определение зависимости средних значений. Методы нахождения таких зависимостей и оценка их статистических свойств составляют содержание регрессионного анализа.

Схема регрессионного анализа включает в себя последовательное решение следующих задач: выбор вида регрессионной функции, оценка ее параметров, проверка статистической значимости выборочной регрессии в сравнении с безусловным разбросом значений y, имеющих дисперсию, определение доверительных границ, с заданной вероятностью включающих в себя истинную регрессию.

Вид регрессионной функции $\eta(x)$ выбирается исходя из особенностей исследуемого процесса. На основе анализа графической зависимости между *y* и *x,* подбирается форма функциональной зависимости. Чаще всего стараются использовать линейную модель, а в случае явно выраженной нелинейной зависимости – использовать различные линеризующие преобразования переменных *x* и *y*.

Простейшей моделью регрессии является модель одномерной линейной регрессии:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$

в которой ошибки ε_i предполагаются независимыми случайными величинами, каждая из которых имеет распределение $N(0,\sigma^2)$. Если и регрессор X – случайная величина, то можно рассматривать модель как условную относительно действительно наблюдавшихся значений регрессора x_i:

$$\mathrm{M}\left[Y_i \middle| X_i = x_i\right] = \beta_0 + \beta_1 x_i .$$

На практике делаются различные допущения, чтобы подобрать регрессионную зависимость, но необходимо, чтобы соответствующая связь была устойчивой и воспроизводимой. Под устойчивостью понимается, что при повторении эксперимента в различных условиях:

1) уравнение регрессии остается без изменений, даже когда изменяются другие аспекты данных;

2) получаемые в различных условиях линии регрессии параллельны;

3) линии регрессии всегда получаются удовлетворительными, но их наклоны и расположение различны.

Непараметрическая регрессионная модель может быть записана в виде:

$$y_i = \eta(\tilde{o}_i, \beta) + \varepsilon_i ,$$

где β – неизвестный вектор параметров регрессии;

$\eta \in C^2\left[a,b\right]$ – некоторая сглаживающая функция;

ε_i – отклонение (случайная ошибка) – нормально распределенная случайная величина со средним, равным нулю, и некоторой дисперсией.

При построении модели возможно использование различных классов функций $\eta(x,\hat{\beta})$ при условии, что их первая и вторая производные непрерывны.

Учитывая, что при программной реализации используются численные методы, т. е. большинство функций реализуется разложением в ряд, без обоснованного предположения о форме нелинейной зависимости нет необходимости использовать специальные математические функции. Поэтому в большинстве практических задач оптимален выбор полиномиальной регрессии:

$$\eta(x,\beta) = \sum_{i=1}^{p} \beta_i \cdot \phi_i(x),$$

где $\phi_k(x) = \sum_{s=1}^{k} \alpha_s x^{s-1}$ полином степени k-1.

Хотя подбор по n наблюдениям полинома степени до n-1 включительно принципиально возможен всегда, при больших значениях k возникают трудности с практической реализацией. Так, при значениях k порядка шести и более регрессионная матрица становится плохо обусловленной [3]. Поэтому одним из решений является использование системы ортогональных функций $\phi_k(x)$.

В такой ситуации возникает вопрос об интерпретации вида подобранного аппроксимирующего полинома, тогда может оказаться уместным использовать линейную комбинацию одночленов. Следует отметить, что на сегодняшних день применяемые в специальных математических пакетах, в частности MathCad, алгоритмы обеспечивают практически одинаковое качество оценки регрессионных моделей как одночленами, так и ортогональными полиномами.

Отрицательно на объясняющих свойствах модели сказывается как отсутствие значимой переменной, так и избыточное присутствие незначимой.

В случае, когда в модель не включена существенная переменная (существенной называют переменную, которая должна быть в модели согласно правильной теории), наблюдаются следующие последствия:

– исчезает возможность правильной оценки и интерпретации уравнений;

– коэффициенты при оставшихся переменных становятся смещенными;

– стандартные ошибки коэффициентов и t-статистики некорректны и не могут быть использованы для суждения о качестве подгонки предлагаемой модели.

Включение несущественной переменной в модель не приводит к смещению оценок коэффициентов, но появляется другой недостаток – растут стандартные ошибки коэффициентов. Оценки становятся статистически незначимыми.

Если точная спецификация модели неизвестна, то используют информационные критерии, позволяющие извлекать максимум информации из исходных данных и тем самым обеспечивать выбор оптимального числа параметров.

Наиболее распространенными критериями является критерий Шварца (Schwarz) и критерий Акайке (Akaike). Оба критерия позволяют выбирать наилучшую модель из множества различных спецификаций. Критерии численно построены так, чтобы учесть влияние на качество подгонки модели двух противоположных тенденций.

При построении моделей предпочтительнее робастная процедура, позволяющая выбрать модель в предположении нестрогой нормальности ошибок и позволяющая использовать различные робастные оценки, одной из которых является формулировка критерия Акайке в виде: [0]:

$$AIC(df) = \ln(\hat{\sigma}^2) + \frac{2 \cdot df}{N},$$

где $\hat{\sigma}$ – робастная оценка параметра рассеивания,

N – размер базисной выборки.

Критерий Шварца:

$$SIC(df) = \ln(\hat{\sigma}^2) + \frac{\ln(N) \cdot df}{N}.$$

В этих формулах первое слагаемое представляет собой штраф за большую дисперсию, второе – штраф за использование дополнительных переменных. Критерии рассчитываются для каждой рассматриваемой спецификации. При сравнении различных моделей предпочтение отдается той, которая имеет наименьшие значения критериев.

В регрессионном анализе выбор оптимальной модели всегда сопровождается оценкой параметров каждого из рассматриваемых вариантов регрессии. При решении задач оценивания параметров регрессионных моделей наиболее распространен критерий наименьших квадратов, который записывается в следующем виде [5]:

$$\hat{a} = \arg \min_{a} \sum_{i=1}^{N} \left[y_i - \eta(a, x_i) \right]^2,$$

где $\hat{a}$ – вектор оценок параметров модели,

$\eta(a, x_i)$ –- регрессионная модель,

y_i, x_i – результаты наблюдений зависимого и независимого измерения.

Во многом именно эффективность вычислительных алгоритмов метода наименьших квадратов (МНК) обусловила широкое применение линейного регрессионного анализа как метода восстановления различных зависимостей. Зачастую на практике пытаются использовать МНК, не рассматривая допустимые условия его применимости. Использование метода наименьших квадратов предполагает, что оценки $\hat{a}$ распределены нормально. При негауссовых случайных ошибках наблюдения оптимальные оценки находят методом максимального правдоподобия.

Применение метода максимального правдоподобия позволяет получать оптимальные оценки независимо от формы распределения случайных ошибок выбранной модели [6], поэтому она эффективна для оценки параметров как линейной так и нелинейной модели. Функция правдоподобия для оценивания среднего и СКО нормального распределения имеет вид [1]:

$$L = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - M\right)^2}{2\sigma^2} \right)$$

и ее логарифм равен:

$$\ell = -\frac{1}{2} \sum_{i=1}^{N} \frac{\left(y_i - M\right)^2}{2\sigma^2} - N\ln(\sigma) - \frac{N}{2}\ln(2\pi).$$

Оценка параметров функций $\hat{M}(t)$ и $\hat{\sigma}(t)$, которые являются параметрами нормального распределения – характеристиками положения и рассеивания соответственно, осуществляется максимизацией по выборке объема N функции:

$$\ell(M, \sigma) = -\frac{1}{2} \sum_{i=1}^{N} \left[\frac{\left(y_i - M(t_i)\right)^2}{\sigma(t_i)^2} + \ln(\sigma(t_i)^2) \right]. \qquad (2)$$

Таким образом, при нормальном распределении измерений, построение модели сводится к выбору вида и оценке параметров аппроксимирующих функций $\hat{M}(t)$ и $\hat{\sigma}(t)$.

Поскольку вид распределения априорно известен, то множество различных статистических критериев проверки на нормальность может быть успешно использовано.

На сегодняшний день действует стандарт ГОСТ Р ИСО 5479-2002 «Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения» [7]. Согласно стандарту применение наиболее распространенного критерия Шапиро-Уилкса допускается только на небольших выборках (N<100). Для построения квантильно-регрессионных моделей необходим большой размер выборки, в таком случае согласно стандарту осуществляется проверка на асимметричность и эксцесс Д'Агостино [8].

В действующем стандарте отсутствует описание вычислительной процедуры, а приведены лишь пороговые точки. На основании оригинальной статьи Д'Агостино составлен алгоритм вычислений, позволяющий производить оценку выборочных значений на симметричность и эксцесс распределения. Исходные данные для алгоритма выборка значений yi, i=1..N. Последовательность шагов оценки на симметричность распределения выборки значений следующая [0].

1. Определение коэффициента асимметрии β по выборочным данным.
2. Вычисление временных переменных:

$$t1 = |b1| \cdot \sqrt{(N+1)(N+3)/6/(N-2)}\,,$$

$$t2 = -1 + \sqrt{2}\sqrt{\frac{3(N^2+27N-70)(N+1)(N+3)}{(N-2)(N+5)(N+7)(N+9)} - 1}\,,$$

$$t3 = \sqrt{2/(t2-1)}\,.$$

3. Расчет выборочной Z-статистики:

$$z_\beta = \ln(\sqrt{(t1/t3)^2 + t1/t3 + 1}) / \sqrt{\ln(\sqrt{t2})}\,.$$

4. В соответствие с заданным уровнем значимости α нахождение 1-α-квантили стандартного нормального распределения N(0,1).

Гипотеза о согласии эмпирического и нормального распределения отвергается, если условие $Z_\beta \geq Z_{1-\alpha}$ выполняется. Значение выборочной z-статистики может быть использовано в качестве количественной оценки симметричности не только при определении типа модели, но и в процессе подгонки модели к данным, как оценка качества соответствия модели данным.

Последовательность шагов оценки на эксцесс распределения выборки значений следующая:

1. Определение коэффициент эксцесса γ по выборочным данным.
2. Вычисление временных переменных:

$$t4 = \frac{|b2| - 2(N-1)/(N+1)}{\sqrt{24N(N-2)(N-3)/(N+1)^2/(N+3)/(N+5)}}\,,$$

$$t5 = \frac{(N+7)(N+9)}{3(N^2-5N+2)}\sqrt{\frac{N(N-2)(N-3)}{6(N+3)(N+5)}}\,,$$

$$t6 = 6 + 4 \cdot t5 \cdot (t5 + \sqrt{t5^2+1})\,.$$

3. Расчет выборочной Z-статистики по формуле:

$$Z_\gamma = \frac{1-2/9/t6-\sqrt[3]{(1-2/t6)/(1+t4\sqrt{2/(t6-4)}}}{\sqrt{2/9/t6}}.$$

4. В соответствии с заданным уровнем значимости α нахождение 1-α-квантили стандартного нормального распределения N(0,1).

Гипотеза о согласии эмпирического и нормального распределения отвергается, если условие $Z_\gamma \geq Z_{1-\alpha}$ выполняется. Значение выборочной статистики Z_γ также может быть использовано в качестве количественной оценки крутизны эмпирических распределений на различных этапах построения и верификации моделей.

Совместный тест (omnibus test), оценивающий нормальность выборки по асимметрии и эксцессу, состоит в вычислении статистики $K = Z_\beta^2 + Z_\chi^2$, проверке ее соответствия χ^2-распределению с 2 степенями свободы [8].

Таким образом, оценки Z_{b1} и Z_{b2} могут быть использованы как количественная мера несоответствия эмпирического распределения нормальному. Эти оценки могут быть использованы как для определения шагов по снижению асимметрии или эксцесса на этапе подготовки данных, так и для оценки качества получаемых моделей.

Поскольку результатом разработки модели является определение функциональной зависимости $M(x)$ – условного среднего и $\sigma(x)$ – СКО, то исходные выборочные значения y_i, могут быть стандартизованы (нормированы) к величине z_i:

$$z_i = \frac{y_i - M(x_i)}{\sigma(x_i)}, \tag{3}$$

имеющей стандартное нормальное распределение с параметрами 0 и 1. Вследствие чего возможна проверка гипотез равенства выборочного среднего нулю, а СКО – 1 с помощью различных как параметрических, так и непараметрических критериев.

Из широкого множества альтернатив в мировой практике наиболее распространено использование χ^2-критерия. Этот критерий также применяется для проверки гипотезы о совпадении законов распределений для серии G выборок.

Для проверки гипотезы H_0: $M\{z\}=0$ для множества G подвыборок из базисной выборки может быть использована так называемая Q-статистика, вычисляемая по формуле:

$$Q_M = \sum_{g=1}^{G} n_g \overline{z}_g^{\,2} \tag{4}$$

и основанная на том, что статистические характеристики выборки при объеме каждой N→∞ имеют нормальное распределение, а сумма квадратов независимых нормально распределенных величин с χ^2-распредление с G-dfm степенями свободы.

Для проверки гипотезы H_0: $D\{z\}=1$ для множества G подвыборок из базисной выборки для вычисления Q-статистики на основе трансформации дисперсии к нормальности Вилсона-Хилферти используется формула:

$$Q_s = \sum_{g=1}^{G} \frac{\left[d_g^{2/3} - \left(1 - \frac{2}{9(n_g - 1)} \right) \right]^2}{\left(\frac{2}{9(n_g - 1)} \right)}, \tag{5}$$

которая имеет χ^2 – распредление с G-dfs степенями свободы.

Последовательность шагов для проверки нормальности следующая:

1. Разброс значений независимого переменного X разбивается на G групп исходя из следующего соотношения:

$$\max\left\{n^{0.25}, \max(dfo) + 3\right\} < G < \min\left\{n^{0,4}, n/50\right\},$$

где dfo – общее число степеней свободы, функций всех параметров, описывающих модель.

При этом необходимо, чтобы G-dfo>3 и размер каждой группы не меньше 50.

2. Для каждой группы вычисляется выборочное среднее и трансформированная Вилсона-Хилферти дисперсия.

3. По формулам (4) и (5) вычисляется Q-статистика.

4. В соответствии с заданным уровнем значимости α нахождение 1-α-квантили χ^2-распредление с G-dfo степенями свободы.

Гипотеза о согласии эмпирического и нормального распределения отвергается, если условие $Q\chi<Q$м и $Q\chi<Qs$ выполняется.

Наибольшую трудность составляет проверка сложной гипотезы нормальности распределений с неизвестным средним и дисперсией. Для проверки такой гипотезы с точки зрения удобства программной реализации эффективен критерий Лина-Мудхолкара в представлении Нельсона. Применение критерия не требует ни упорядочивания, ни преобразования переменных, ни выборочной оценки параметров распределения [9].

Критерий основан на анализе n средних и дисперсий, рассчитанных по n подвыборкам из базовой выборки, получаемым путем исключения одного наблюдения. Полученные пары значений среднего и дисперсии используются для проверки их независимости посредством линейного коэффициента корреляции Пирсона. Использование этого коэффициента корреляции возможно только в нормальной двухмерной совокупности.

В отличие от распределения средних значений, распределение дисперсий не является нормальным, поэтому к нему применяется нормализующее преобразование Вилсона-Хилферти:

$$d_i = \frac{1}{n}\left[\sum_{j\neq i} sd_j^2 - \frac{1}{n-1}\left(\sum_{j\neq i} sd_j \right)^2 \right]^{1/3},$$

где $sd_i = \frac{1}{n}\sum_{j\neq i}\left(x_i - M^2\right)$ – выборочная дисперсия наблюдений за исключением i-го.

Для построения статистики критерия используется нормализующая трансформация Фишера для линейного коэффициента корреляции r:

$$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

и если $|Z|$ очень велико, гипотеза о нормальности отвергается, т. е. в числовом выражении:

$$|Z| < z(0,5+\alpha/2),$$

где z – критическое значение статистики критерия, определяемое через соответствующие уровню значимости α квантили стандартного нормального распределения z^{-1} по формуле:

$$z(\tau) = \left[\frac{3}{n} - \frac{7,324}{n^2} + \frac{53,005}{n^3}\right] \cdot \left\{ z^{-1}_{\tau} + \frac{\left(z^{-1}_{\tau}\right)^3 - 3 \cdot z^{-1}_{\tau}}{24} \cdot \left[-\frac{11,7}{n} + \frac{55,06}{n^2}\right]\right\}.$$

Следует отметить, что критерии, основанные на оценках максимального правдоподобия – отношении правдоподобия, М-оценках, не рассмотрены преднамеренно, так как этот метод использован для оценки параметров модели.

2. Практическая задача – разработка региональных справочных кривых роста детей

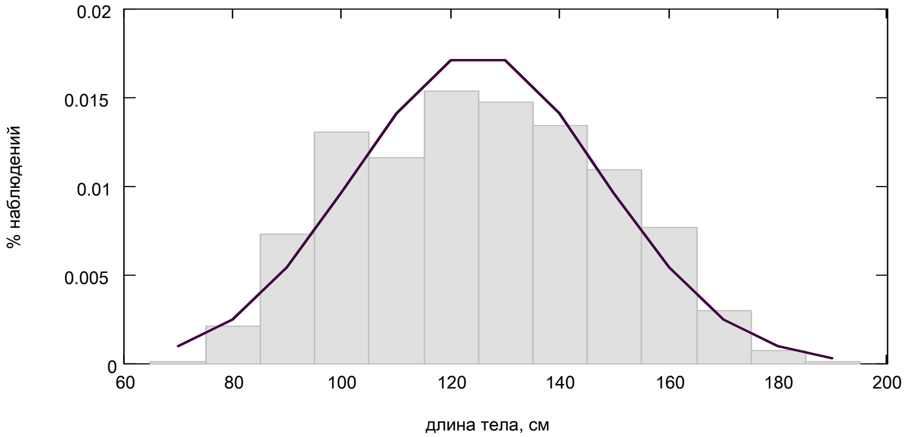
Экспериментальный материал. Для анализа в качестве базисной группы были использованы данные перекрестного обследования 8422 детей в возрасте от 2 до 14 лет. Все измерения были проведены в условиях поликлиники или стационара и выполнены в соответствии со стандартной методикой специально подготовленным медицинским персоналом. Были исключены дети, имеющие выраженную эндокринную и другую патологию развития, а также родившиеся недоношенными. Каждое измерение состояло из четырех параметров: значение массы и длины тела, пол и хронологический возраст ребенка, ИМТ был вычислен с точностью до трех значащих цифр после запятой.

Предварительный анализ данных. В таблице 1 представлены основные описательные статистики выборки.

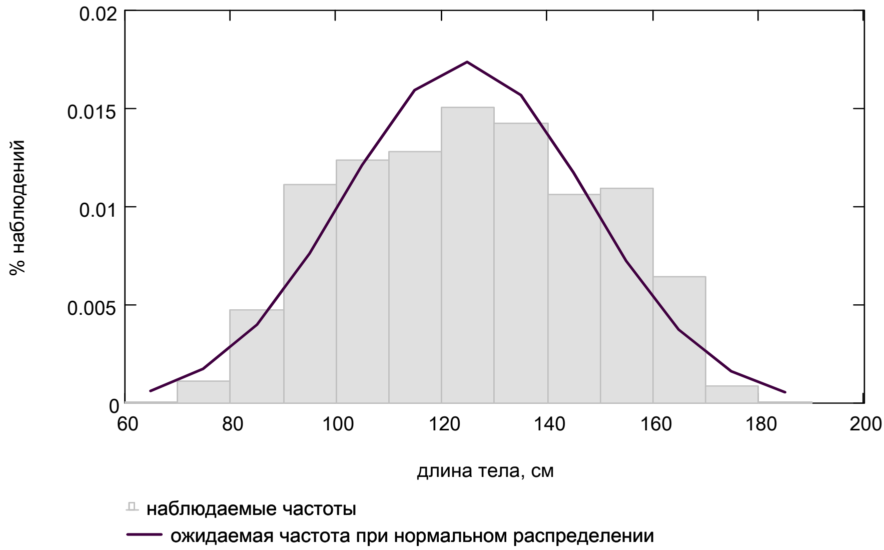
Таблица 1. Описательные статистики исходной выборки

	среднее (M)	*ошибка M*	*нижняя квар-тиль*	*меди ана*	*верхняя квар-тиль*	*СКО*	*асим-метрия*	*экс-цесс*
девочки (N=4199)								
Масса	27,5	0,2	18,0	24,0	35,0	11,9	0,9	0,2
Длина	124,4	0,4	105,0	125,0	142,0	22,9	0,0	-0,9
ИМТ	16,9	0,1	15,1	17,0	18,4	2,7	1,2	4,2
Возраст	8,0	0,1	4,8	7,8	11,0	3,6	0,1	-1,2
мальчики (N=4223)								
Масса	27,9	0,2	18,0	25,0	35,0	12,1	0,9	0,3
Длина	125,2	0,3	106,0	124,5	142,0	23,7	0,1	-0,8
ИМТ	17,1	0,0	15,4	17,2	18,3	2,7	1,7	8,2
Возраст	8,0	0,1	4,8	7,7	11,0	3,6	0,1	-1,1

Распределение длины тела девочек и мальчиков представлены гистограммой на рисунке 1 совместно с кривой нормального распределения, среднее и СКО которой оценены по соответствующим выборочным данным.



а) мальчиков



б) девочек

Рис. 1. Гистограмма эмпирического распределения длины тела мальчиков и девочек в возрасте от 2 до 14 лет

Проведенный анализ распределений исходных данных, преобразованных функциями натурального (ln(x)) и десятичного логарифма (log10(x)), отношением к 1 (1/x), извлечением квадрата ($\sqrt{x}$), экспонентой (e^{-x}) и логистой ($1/(1-e^{-x})$) длины тела, выявил значимые отличия от нормального. Коэффициенты асимметрии трансформированных выборок значимо отличались от нуля, за исключением преобразования $\sqrt{x}$ (-0,097 у мальчиков и -0,147 у девочек).

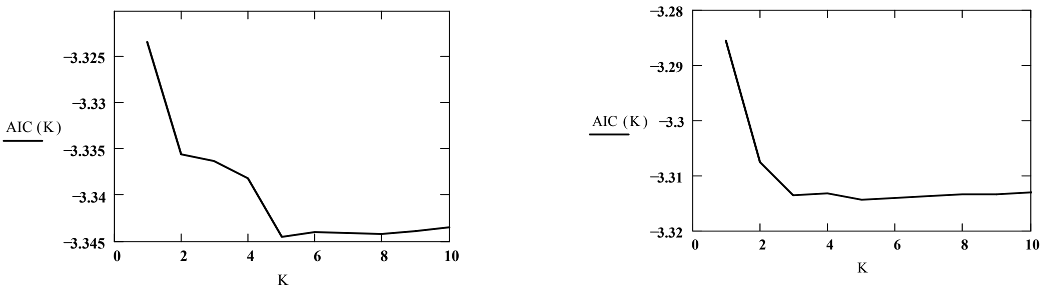
Коэффициенты эксцесса выборок всех указанных преобразований ненамного превышали значение в исходной выборке. Исключение составляло преобразование – 1/x, когда коэффициенты эксцесса были равны -0,15 у девочек и -0,16 у мальчиков, но при этом коэффициенты асимметрии были высоки:0,67 и 0,64.

Проверка по критерию Колмогорова-Смирнова с поправкой Лиллифорс не выявила согласия распределения ни одной перечисленной трансформации длины тела у мальчиков и девочек с нормальным.

В ходе проведенного анализа было выявлено, что простое преобразование – возведение в квадрат, позволяет получить нормальное распределение в большинстве половозрастных групп. Каждое выборочное значение длины тела было трансформировано по формуле: $y_i^* = \left[\frac{y_i}{100}\right]^2$ для построения параметрической квантильно-регрессионной модели длины тела.

<u>Процесс построения модели.</u> Основной задачей процесса построения модели являлась аппроксимация связывающих функций M(x) и σ(x), являющихся регрессионной и скедастической зависимостями. Поскольку регрессионный анализ является довольно распространенной задачей, большинство его методов довольно широко реализовано в большинстве специальных математических и статистических пакетов, поэтому выбор функции M(x) не представляет большой сложности.

На первом шаге был осуществлен выбор вида регрессионной функции, для первоначального анализа была выбрана полиномиальная функция. Выбор степени полинома определялся с использованием критерия Акайке (AIC). Результаты расчета в виде зависимости значения AIC от степени полинома K в пределах от 1 до 10 представлены на рисунке 2.



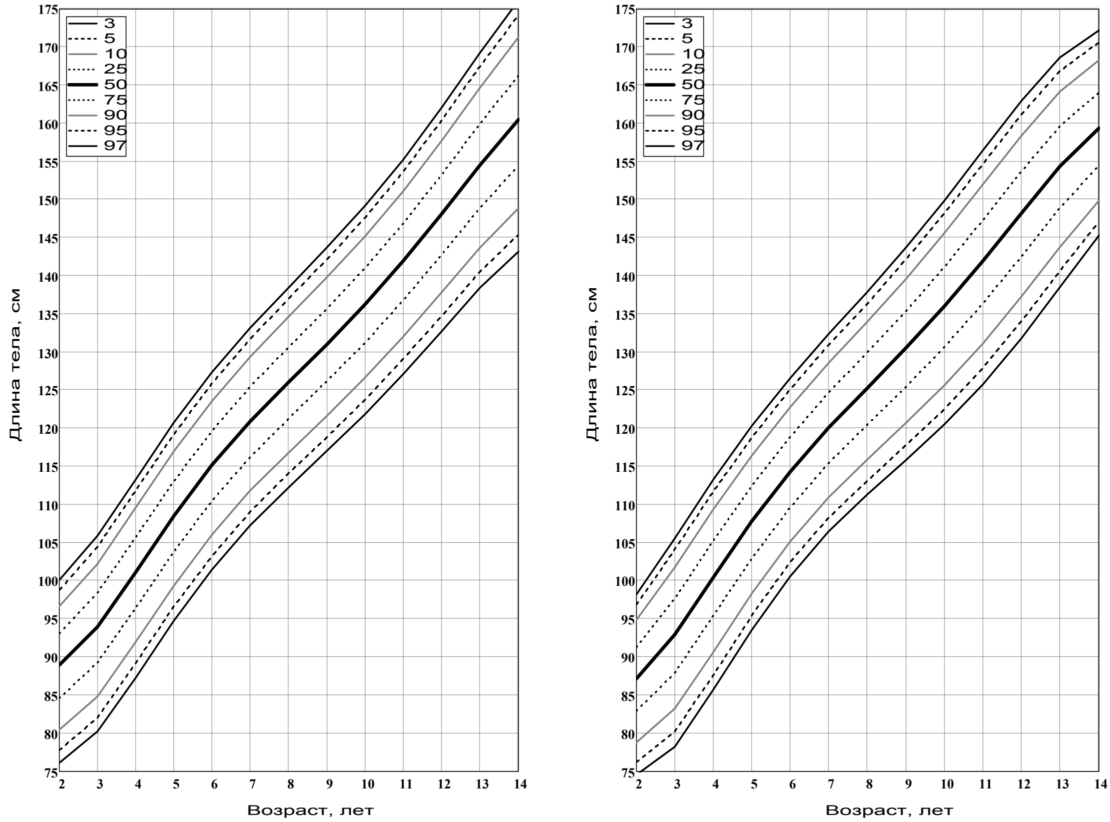
а) девочек *б) мальчиков*

Рис. 2. Зависимость значения критерия AIC от степени полинома, выбранного для аппроксимации среднего

Для модели длины тела девочек степень полинома, минимизирующая значения критерия Акайке, была равна 5 и для мальчиков – 3.

На втором шаге был осуществлен выбор вида аппроксимирующей функции для СКО. Для удобства вычислений была использована полиномиальная регрессия для квадрата отклонений каждого наблюдения от среднего значения. Определение оптимальной степени аппроксимирующего полинома также осуществлено с помощью критерия AIC. Реализация расчета включает совместный выбор степени полиномов для регрессионной и скедастической зависимостей.

Для аппроксимации функции σ(x) модели длины тела девочек выбран полином 4 степени и мальчиков – 3. Следует отметить, что для модели длины тела мальчиков совместная оценка AIC двух функций M(x) и σ(x) указывает на необходимость увеличения степени аппроксимирующего полинома M(x) до 5 (рис. 3).



а) мальчиков *б) девочек*

Рис. 3. Разработанные диаграммы роста детей

Таким образом, модель квадрата длины тела девочек имеет вид:

$$M(x) = 0,785 - 0,134x + 0,086x^2 - 0,012x^3 + 7,7 \cdot 10^{-4} x^4 - 1,8 \cdot 10^{-5} x^5;$$

$$S(x) = 0,061 + 3,1 \cdot 10^{-2} x - 3,4 \cdot 10^{-3} x^2 + 1,7 \cdot 10^{-4} x^3,$$

где x – хронологический возраст.

Модель квадрата длины тела мальчиков имеет вид:

$$M(x) = 0,928 - 0,301x + 0,146 \cdot x^2 - 0,022x^3 + 1,5 \cdot 10^{-3} x^4 - 3,7 \cdot 10^{-5} x^5;$$

$$S(x) = -0,049 + 0,045x - 0,01x^2 + 9,8 \cdot 10^{-4} x^3 - 3,2 \cdot 10^{-5} x^4,$$

где x – хронологический возраст.

Для получения диаграмм возрастной динамики длины тела мальчиков и девочек (рис. 3, вычисленные по формуле (1), значения квантилей были трансформированы функцией квадратного корня.

Вычисленные по модели длины тела мальчиков z-оценки измерений имели – нулевое среднее (расхождение в 8-м знаке), единичную дисперсию (расхождение в 3-м знаке после запятой), коэффициент асимметрии -0,019 и эксцесса 0,526. Проверка на симметричность по Д'Агостино позволила принять нулевую гипотезу, значение эксцесса было статистически значимо больше нуля, что привело к отклонению нулевой гипотезы при использовании совместного теста на симметричность и эксцесс.

Вычисленные по модели длины тела девочки z-оценки измерений имели – нулевое среднее (расхождение в 5-м знаке), единичную дисперсию (в 3-м знаке после запятой), коэффициент асимметрии -0,048 и эксцесса 0,555. Проверка на симметричность по Д'Агостино позволила принять нулевую гипотезу, значение эксцесса статистически значимо больше нуля.

Для верификации полученных моделей весь возрастной разброс значения длины тела мальчиков был разбит на 17 групп, с одинаковым возрастным интервалом и числом наблюдений не менее 200. Для каждой группы были рассчитаны z-оценки, по которым были вычислены выборочные средние и СКО. Вычисленное по z-оценкам значение Q-статистики среднего длины тела мальчиков составило 4,3 и девочек – 9,1, при значении 5-й процентильной точки $\chi 2$-распредления с 12 степенями свободы – $\chi 2(0{,}95,\ 12)=19{,}7$. Значение Q-статистики дисперсий составило 21,3 по выборке z-оценок длины тела мальчиков и 24,5 – девочек, что меньше значения $\chi 2(0{,}95,\ 15)=25{,}2$. Варьирование числа групп в допустимых для имеющегося объема выборки и числа степеней свободы модели в пределах – от 8 до 28 не привело к значимому увеличению Q-статистики.

Оцененное по z-оценкам длины тела девочек значение Q-статистики среднего составило 31,6, что меньше значения 5-й процентильной точки $\chi 2$-распредления с 12 степенями свободы – $\chi 2(0{,}95,\ 12)=19{,}7$. Значение Q-статистики дисперсии составило 31,3, что меньше значения $\chi 2(0{,}95,\ 15)=25{,}2$. Варьирование числа групп в допустимых для имеющегося объема выборки и числа степеней свободы модели в пределах – от 8 до 28 не привело к увеличению Q-статистики. Отклонения коэффициентов асимметрии также не значительно отличались от нуля, тогда как отклонение эксцесса было значительным и в некоторых выборках превышало 1.

Причиной слишком высокого эксцесса является неоптимальный выбор трансформации первичных измерений, поэтому эффективно осуществлять не подбор простого преобразования, а построение LMS-модели, которая включает в себя выбор оптимальной трансформации к нормальности.

В работе [10] представлена LMS-модель длины тела, построенная по такому же экспериментальному материалу, а результаты ее верификации показали статистически незначимое отклонение эксцесса.

Следует отметить, что, несмотря на высокую скорость вычислений, оценки полиномиальных коэффициентов зависимостей среднего и СКО, полученные методом наименьших квадратов, менее устойчивые, чем оценки, полученные методом максимального правдоподобия. Поэтому для построения параметрической квантильно-регрессионной модели предлагается использовать алгоритм с оцениванием параметров методом наименьших квадратов.

Поскольку, для вычислений требуется решение оптимизационной задачи, то при реализации вычислений на практике эффективна схема с предварительной оценкой параметров методом наименьших квадратов, которые будут использоваться для инициализации оптимизационной процедуры.

Выбор числа степеней свободы для каждой из функций, входящих в модель, осуществляется минимизацией обобщенного Акайке критерия с заданным штрафом. Для верификации полученной модели использованы статистические характеристик z-оценок исходных данных.

Методы регрессионного анализа могут быть использованы при решении различных практических задач [11-35].

Заключение

В работе описаны математическое и алгоритмическое обеспечения параметрических моделей. Рассмотрена практическая задача, связанная с разработкой региональных справочных кривых роста детей. В ходе проведенного анализа было выявлено, что простое преобразование – возведение в квадрат, позволяет получить нормальное распределение в большинстве половозрастных групп. Разработаны диаграммы роста детей.

Литература

[1] Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, 2003. – 686 с.

[2] Кокс Д., Хинкли Д. Теоретическая статистика. – М.: Мир. – 1974. – 560 с.

[3] Себер Дж. Линейный регрессионный анализ. – М.: Мир. – 1980. – 456 с.

[4] Робастность в статистике. Подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль – М.: Мир, 1989. – 512 с.

[5] Круг Г.К., Кабанов В.А., Фомин Г.А., Фомина Е.С. Планирование эксперимента в задачах нелинейного оценивания и распознавания образов. – М.: Наука. – 1981. – 172 с.

[6] Математическая теория планирования эксперимента / Под. ред. С.М. Ермакова. – М.: Наука. – 1983. – 392 с.

[7] Лемешко Б.Ю. Сравнительный анализ критериев проверки отклонения распределения от нормального закона / Б.Ю. Лемешко, С.Ю. Лемешко // Метрология. – 2005. – №.2. – С. 3. – 23.

[8] D'Agostino R.B. Test for the normal distribution/ Goodness-of-fit techniques // R.B. D'Agostino, M.A. Stephens. – 1986. – New York. Marcel Dekker. – P. 367 – 419.

[9] Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ. – 2006. – 816 с.

[10] Львович И.Я. Определение справочных показателей физического развития детей с применением LMS-метода/ И.Я. Львович, О.В. Минакова, В.П. Ситникова // Вестник ВГТУ. – 2007. – № 10. – С. 96-101.

[11] Klimenko G.Ya. Optimization of medical aid for pregnant women with iron deficiency anemia based on predictive modeling of their health state with due consideration of medical and social risk factors / G.Ya. Klimenko, S.A. Pyataeva, I.Ya. Lvovich, O.N. Choporov, N.V. Naumov. – Lorman, MS, USA. Science Book Publishing House, 2012. – 144 p.

[12] Анализ динамики и прогнозирование распространенности дерматозов среди населения воронежской области / Е.Н. Бугакова, Г.Я. Клименко, О.Н. Чопоров, Г.В. Сыч // Врач-аспирант. – 2010. – Т. 43. – № 6.2. – С. 259-267

[13] Болгов С.В. Прогнозирование стоматологической заболеваемости по медико-биологическим и социально-гигиеническим факторам риска / С.В. Болгов, К.А. Разинкин, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 49. – № 6.2. – С. 294-301

[14] Бугакова Е.Н. Прогнозирование заболеваемости населения аллергическими дерматозами по медико-социальным факторам риска / Е.Н. Бугакова, Г.Я. Клименко, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2010. – Т. 9. – № 4. – С. 801-804.

[15] Интегральное оценивание и прогностическое моделирование состояния здоровья беременных, рожениц и родильниц с учетом их медико-социальных характеристик / О.Н. Чопоров, В.П. Косолапов, Н.В. Наумов, Х.А. Гацайниева // Вестник Воронежского института высоких технологий. – 2012. – №9. – С. 91-95.

[16] Классификация районов белгородской области по распространенности злокачественных новообразований и результаты краткосрочного прогнозирования / А.И. Агарков, Г.Я. Клименко, О.Н. Чопоров, Ю.Ю. Шуршуков // Системный анализ и управление в биомедицинских системах. – 2013. – Т. 12. – № 4. – С. 1134-1138.

[17] Клименко Г.Я. Индивидуальное прогнозирование заболеваемости туберкулезом органов дыхания по медико-социальным факторам риска / Г.Я. Клименко, В.А. Николаев, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2010. – Т. 9. – № 4. – С. 892-896.

[18] Методы предварительной обработки информации при системном анализе и моделировании медицинских систем / О.Н. Чопоров, Н.В. Наумов, Л.А. Куташова, А.И. Агарков // Врач-аспирант. – 2012. – Т. 55. – № 6.2. – С. 382-390.

[19] Моделирование и прогнозирование заболеваемости миомой матки в сочетании с аденомиозом по медико-социальным факторам риска / О.Н. Чопоров, Н.Н. Кудинова, М.В. Фролов, Г.Я. Клименко // Моделирование, оптимизация и информационные технологии. – 2013. - № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Kudinova_soavtori_3_13_1.pdf.

[20] Моделирование и прогнозирование качества жизни беременных женщин и пути его улучшения / В.И. Стародубов, Г.Я. Клименко, С.В. Говоров, Н.Б. Костюкова, О.Н. Чопоров. – Воронеж: Изд-во «Истоки», 2009. – 188 с.

[21] Оптимизация управления функционированием медицинских систем различного уровня / О.Н. Чопоров, И.Я. Львович, К.А. Разинкин, А.А. Рындин // Системы управления и информационные технологии. – 2013. – Т. 53. – №3. – С. 100-104

[22] Прогнозирование развития онкологической заболеваемости по индивидуальным медико-социальным факторам риска / О.Н. Чопоров, А.И. Агарков, Г.Я. Клименко, Ю.Ю. Шуршуков // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Agarkov_soavtori_3_13_1.pdf.

[23] Прогнозирование удовлетворенности стоматологической помощью по медико-социальным факторам риска / С.В. Болгов, О.Н. Чопоров, Г.Я. Клименко, К.А. Разинкин // Системный анализ и управление в биомедицинских системах. – 2013. – Т. 12. – № 4. – С. 1171-1176.

[24] Разработка и использование моделей для прогнозирования качества жизни беременных по их медико-социальным характеристикам / Х.А. Махер, Н.В. Наумов, Г.Я. Клименко, О.Н. Чопоров // Системный анализ и управление в биомедицинских системах. – 2011. – Т. 10. – № 4. – С. 789-793.

[25] Разработка классификационно-прогностических моделей развития гнойно-септических осложнений у родильниц / А.В. Чернов, В.Ю. Бригадирова, О.Н. Чопоров, В.И. Чернов // Системный анализ и управление в биомедицинских системах. – 2012. – Т. 11. – № 1. – С. 261-266.

[26] Химина И.Н. Рационализация управления медицинской помощью больным с заболеваниями желудка и двенадцатиперстной кишки на основе комплексного мониторинга и классификационно-прогностического моделирования / И.Н. Химина, В.Н. Эктов, О.Н. Чопоров. – Воронеж: Изд-во «Научная книга», 2014. – 181 с.
[27] Хопина О.А. Прогнозирование заболеваемости глаукомой по индивидуальным медико-социальным факторам риска / О.А. Хопина, Г.Я. Клименко, О.Н. Чопоров // Врач-аспирант. – 2011. – Т. 49. – № 6.1. – С. 203-208.
[28] Чембарцева Н.Я. Моделирование и прогнозирование состояние здоровья новорожденных по медико-социальным факторам риска / Н.Я. Чембарцева, Г.Я. Клименко. – Воронеж, 2006. – 133 с.
[29] Чопоров О.Н. Оптимизация функционирования медицинских систем на основе интегральных оценок и классификационно-прогностического моделирования: дисс.: д-ра техн. наук / О.Н. Чопоров. – Воронеж, 2001. – 325 С.
[30] Бережная Е.В. Оценка риска для здоровья населения г. Воронежа при воздействии химических веществ, загрязняющих атмосферный воздух / Е.В.Бережная // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – С. 2.
[31] Преображенский Ю.П. Оценка эффективности применения системы интеллектуальной поддержки принятия решений / Ю.П. Преображенский // Вестник Воронежского института высоких технологий. – 2009. – № 5. – С. 116-119.
[32] Львович Я.Е. Принятие решений в экспертно-виртуальной среде / Я.Е. Львович, И.Я. Львович. – Воронеж: Изд-во «Научная книга», 2010. – 139 с.
[33] Гафанович Е.Я. Прогнозирование исходов и выбор рационального лечения артериальной гипертензии с применением математических методов / Е.Я. Гафанович, И.Я. Львович // Вестник Воронежского государственного технического университета. – 2013. – Т. 9. – № 4. – С. 84-86.
[34] Калаев В.Н. Регрессионный анализ в биологических исследованиях / В.Н. Калаев, Е.А. Калаева, А.П. Преображенский, О.В. Хорсева // Системный анализ и управление в биомедицинских системах. – 2007. – Т. 6. – № 3. – С. 755-759.
[35] Преображенский Ю.П. Применение имитационно-семантического моделирования и полумарковских процессов принятия решений в клинической практике / Ю.П. Преображенский, Н.С. Преображенская // Вестник Воронежского института высоких технологий. – 2010. – № 6. – С. 83-89.

Prof. Igor Lvovich, D. Sc.
Paneuropean University, Bratislava, Slovakia
office@vivt.ru

Prof. Yakov Lvovich, D. Sc.
the honored scientist of the Russian Academy of Natural Sciences
President of Voronezh Institute of High Technologies
office@vivt.ru

Building of semi-quantile regression models

Построение полупараметрических квантильно-регрессионных моделей

I. Y. Lvovich, Y. E. Lvovich, O. N. Choporov
И. Я. Львович, Я. Е. Львович, О. Н. Чопоров

Abstract:

The article describes the mathematical background of building of semi-parametric quantile regression models (LMS-models). The basic characteristics of statistical distributions are considered and the algorithms of their transformation to normality (box-cox transformation, Manly exponential transformation, modular transformation) are given and their modification is proposed. The criteria for selecting optimal models are defined. The developed integrated algorithm of LMS - model construction based on the transformation of the primary measurements to normality, including stages of the model initialization, the model selection and the model configuration is presented. The scheme of a process of the model configuration with the use of a modified chain graph is introduced. The reference chart of body mass, which can be used to monitor obesity in the pediatric population are built with the application of the proposed LMS-method in different quantile modes.

Аннотация:

В статье описаны математические предпосылки построения полупараметрических квантильно-регрессионных моделей (LMS-моделей). Рассмотрены основные характеристики статистических распределений и приведены алгоритмы их трансформации к нормальности (трансформация бокса-кокса, экспоненциальная трансформация Манли, модульная трансформация), предложена их модификация. Определены критерии выбора оптимальной модели. Представлен разработанный интегрированный алгоритм построения LMS-модели на основе трансформации первичных измерений к нормальности, включающий этапы инициализации модели, выбора модели и настройки модели. Представлена схема процесса настройки *модели с использованием модифицированного цепочечного графика. С применением предложенного LMS-метода в различных квантильных режимах построены справочные диаграммы массы тела, которые могут быть использованы для мониторинга ожирения среди детской популяции.*

Key words:

Semi-parametric models, quantile regression models, LMS method, transformation to normality, chain graph, body weight diagram.

Ключевые слова:
Полупараметрические модели, квантильно-регрессионные модели, LMS-метод, трансформация к нормальности, цепочечный график, диаграмма массы тела.

ACM Computing Classification System:

Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

Во многих случаях задача полупараметрического анализа связана с оценкой регрессионных коэффициентов, определяющих положения распределения переменных. Важно знать меру точности получаемых оценок при использовании полупараметрического подхода в квантильно-регрессионных моделях. В работе необходимо рассмотреть основные характеристики статистических распределений и дать алгоритм построения требуемой модели для различных квантильных режимов.

1. Математические предпосылки построения полупараметрических моделей

На практике редко встречаются распределения, точно соответствующие нормальному [1], поэтому при анализе данных стремятся подобрать некоторую трансформацию исходного значения, в результате которой обеспечивалось бы соответствие нормальному закону распределения. Такой подход, основанный на трансформации к нормальности, принято считать полупараметрическим.

Первые два момента – матожидание и дисперсия – характеризуют положение и рассеивание распределения, третий и четвертый – являются характеристиками формы и позволяют количественно представлять форму распределения. Так как коэффициенты асимметрии и эксцесса характеризуют форму распределения, то следовательно, они могут быть выбраны в качестве меры, определяющей, с какой точностью эмпирическое распределение может быть аппроксимировано нормальным [0-4]. Считается, что изменение коэффициента асимметрии приводит к более фундаментальному изменению свойств распределения, чем изменение параметра положения и рассеивания [2, 5, 6].

В прикладной статистике наиболее распространено применение преобразования Бокса-Кокса с целью трансформации первичных данных к нормальному распределению.

В оригинальной работе показано, что степенная трансформация Бокса-Кокса [7] $y^*(\lambda)=\begin{cases}(y^\lambda-1)/\lambda, & \lambda\neq 0\\ \log(y), & \lambda=0\end{cases}$, позволяет не только устранить асимметричность формы распределения, но и стабилизировать дисперсию ошибок, а также избавиться от нелинейности. Указанная трансформация включает все известные простые преобразования: степенное, логарифмическое, обратное, квадратного и кубического корня, а также все возможные промежуточные точки – и позволяет избежать разрыва при $\lambda=0$. Представленная форма трансформации Бокса-Кокса не применима к отрицательным значениям, поэтому в таких случаях используется модификация:

$$y^*(\lambda)=\begin{cases}((y+\lambda_2)^{\lambda_1}-1)/\lambda, \lambda\neq 0\\ \log(y+\lambda_2),\ \lambda=0\end{cases},$$

где λ=(λ1, λ2).

Очевидно, что не все данные могут быть трансформированы к нормальности. Исследования этой проблемы показали, что даже когда степенная трансформация $y^*(\lambda)$ не может приводить распределение к нормальному, полученное в ходе оценивания значение λ, обеспечивает нулевое значение коэффициента асимметрии.

Существует и другие преобразования для трансформации к нормальности. Манли предложил экспоненциальную трансформацию вида: $y^*(\lambda)=\begin{cases}(e^{\lambda\cdot y}-1)/\lambda, & \lambda\neq 0\\ y, & \lambda=0\end{cases}$, допускающую использование отрицательных значений исходных данных. Эта трансформация эффективна для одномодальных асимметричных распределений, но не применима к бимодальным и U-образной формы.

Для случаев, когда форма распределения близка к симметричному виду, предложена «модульная трансформация» вида:

$$y^*(\lambda)=\begin{cases}sign(y)\cdot((|y|+1)^\lambda-1)/\lambda, & \lambda\neq 0\\ sign(y)\cdot\log(|y|+1), & \lambda=0\end{cases},$$

где $sign(y)=\begin{cases}1, & y\geq 0\\ -1, & y<0\end{cases}$.

Так для некоторого λ преобразованные наблюдения $y^*(\lambda)$ удовлетворяют предположениям нормальной теории, т. е. $y^*(\lambda)\in N(X\beta,\sigma^2 I)$ и ее плотность распределения равна:

$$f^*(\lambda)=(2\pi\sigma^2)^{-(n/2)}\exp\left\{-\frac{1}{2\sigma^2}(y^2-X\beta)(y^\lambda-X\beta)\right\}$$

При этом функция правдоподобия для исходных наблюдений имеет вид:

$$L=(2\pi\sigma^2)^{-(1/2)n}exp\left\{-\frac{1}{2\sigma^2}(y^\lambda-X\beta)'(y^\lambda-X\beta)\right\},$$

где $J = \prod_{i=1}^{n} y_i^{\lambda-1}$ – абсолютная величина якобиана преобразования от y к y*.

С точностью до константны отношение максимального правдоподобия равно [7]:

$$L_{\max}(\lambda) = -\frac{n}{2}\log\left\{RSS(\lambda, \mathbf{y})\right\} + \log J \cdot$$

Следовательно, максимизируя отношения правдоподобия, можно определить значение λ. При программной реализации вычисление степени трансформации непосредственно по графику зависимости логарифмической функции правдоподобия $L_{\max}(\lambda)$ от λ, которое вычисляется как [8]:

$$L_{\max}(\lambda) = -0.5n \cdot \log(RSS(z^{(\lambda)})), \tag{1}$$

где RSS – остаточная сумма квадратов регрессии значения z на независимую переменную x,

z – трансформация зависимого переменного y вида:

$$z_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}}{\dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y}\log y_i, & \lambda = 0 \end{cases},$$

где $\dot{y}$ – среднее геометрическое выборочных значений y_i,

может быть заменено поиском по сетке.

Для проверки пригодности преобразования, соответствующего некоторому значению $\lambda = \lambda_0$, в [8] предлагается использовать статистику критерия отношения правдоподобия $2\left(L_{\max}(\lambda_0) - L_{\max}(\hat{\lambda})\right)$, асимптотическим распределением которой при проверке гипотезы H: $\lambda = \lambda_0$ является χ2-распредление.

Модифицированная трансформация каждого выборочного значения y_i позволяет получить стандартизованную величину z, распределенную по нормальному закону N(0,1) по формуле [9]:

$$z = \begin{cases} \dfrac{\left(\dfrac{y}{\mu}\right)^{\lambda} - 1}{\lambda \cdot \sigma}, & \lambda \neq 0, \\ \dfrac{\log(y/\mu)}{\sigma}, & \lambda = 0 \end{cases} \tag{2}$$

где λ – степень трансформации Бокса-Кокса выборочных значений,

μ – медиана выборочных значений y,

σ – СКО выборочных значений y.

Замена параметра трансформации функцией независимого переменного приводит к следующей формуле преобразования измеренного значения y к величине z, имеющей стандартное нормального распределение:

$$z = \begin{cases} \dfrac{\left(\dfrac{y}{M(x)}\right)^{L(x)} - 1}{L(x) \cdot S(x)}, & L(x) \neq 0 \\ \dfrac{\log(y / M(x))}{S(x)}, & L(x) = 0 \end{cases} \tag{3}$$

где L(x) – функция, представляющая зависимость степени трансформации от независимого переменного x, устраняющая асимметричность распределения зависимой переменной y;

M(x) – функция, представляющая зависимость медианы переменной y от независимого переменного x;

S(x) – функция, представляющая зависимость коэффициента вариации переменной y от независимого переменного x.

Апостериорная вероятность некоторого результата трансформированного измерения $z \in N(\lambda 1, \lambda 2)$ равна [10]:

$$f(z; \lambda 1, \lambda 2)dz = \frac{1}{\lambda 2\sqrt{2\pi}} \exp\left(-\frac{(z - \lambda 1)^2}{2\lambda 2^2}\right)dz,$$

Так как трансформированная величина Z должна удовлетворять стандартному нормальному распределению, значение параметров должны быть λ1=0 и λ2=1. Подставляя вместо Z формулу трансформации исходного измерения, получаем вероятность измерения y:

$$f(y, L, M, S)dy = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \left\{\frac{\left(\dfrac{y}{M}\right)^L - 1}{L \cdot S}\right\}^2\right) \cdot \frac{\left(\dfrac{y}{M}\right)^L}{y \cdot S} dy \cdot$$

Учитывая все N измерений, получаем функцию правдоподобия:

$$\Lambda = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \left\{\frac{\left(\dfrac{y_i}{M(x_i)}\right)^{L(x_i)} - 1}{L(x_i) \cdot S(x_i)}\right\}^2\right) \cdot \frac{\left(\dfrac{y_i}{M(x_i)}\right)^{L(x_i)}}{y_i \cdot S(x_i)},$$

и ее логарифм равен:

$$l(L, M, S) = \sum_{i=1}^{n} \left(L(x_i)\ln\frac{y_i}{M(x_i)} - \ln\{y_i S(x)\} - \frac{1}{2}\left\{\frac{\left[\dfrac{y_i}{M(x_i)}\right]^{L(x_i)} - 1}{L(x_i)S(x_i)}\right\}^2\right) + const$$

Максимизация логарифма функции правдоподобия позволяет оценить параметры функций L(x), M(t) и S(x) по выборке.

Наилучшая функция должна проходить через все N точек наблюдения. Очевидно, что такое качество подгонки избыточно, поэтому при сплайновой аппроксимации функций модели вводится штраф за усложнение вида функций:

$$R(L,M,S)=\frac{\alpha_L}{2}\int\left\{L''(x)\right\}^2dx+\frac{\alpha_M}{2}\int\left\{M''(x)\right\}^2dx+\frac{\alpha_S}{2}\int\left\{S''(x)\right\}^2dx,$$

где α_L, α_M, α_S, – сглаживающие параметры, необходимые для контроля качества подгонки.

Для подгонки множества данных наиболее мощным средством выбора среди различных моделей является общий информационный критерий Акайке [11]:

$$GAIC(p)=-2\cdot\hat{L}(df)+p\cdot df\text{ ,}\tag{4}$$

где df – количество параметров модели,

$\hat{L}(df)$ – оценка логарифмической функции правдоподобия модели с количеством параметров, равных df,

p – фиксированное значение штрафа.

Информационный критерий Акайке, основанный на теории информации, приближенно максимизирует энтропию модели. Применительно к регрессионным моделям с нормально распределенными ошибками критерий основывается на минимизации RSS и вычисляется [11]:

$$AIC(df)=K(n,\hat{\sigma})+RSS/\hat{\sigma}^2+2\cdot df\ ,$$

где $K(n,\hat{\sigma})$ – константа, зависящая от частных распределений,

RSS – остаточная сумма квадратов относительно среднеквадратичной оценки.

Таким образом, для выбора оптимальной LMS-модели может быть применен информационный критерий Акайке по (4) с числом степеней свободы $df=dfl+dfm+dfs+1$, где dfm, dfs и dfl – порядок полинома или количество узлов сплайновой аппроксимации связывающих функции M(x), S(x) и L(x) соответственно, и 1 добавляется при использовании сглаживающего параметра α.

Оптимальная LMS-модель должна обеспечивать соответствие трансформированных по (3) значений N(0,1). Частным случаем LMS модели является параметрическая модель без трансформации, когда L(x)=1 и dfl=1.

Особенностью LMS моделей является трансформация первичных измерений к нормальности, т. е. полученное с помощью функций моделей множество трансформированных значений исходных данных $i=\overline{1,N}$ должно соответствовать стандартному нормальному распределению N(0,1).

Следовательно, можно сформулировать нулевую гипотезу следующим образом – трансформированные с помощью LMS-модели измерения (z_i) являются независимыми и нормально распределенными случайными величинами, с нулевым средним и единичной дисперсией, т. е. $z_i\in N(0,1)$.

В рассматриваемой задаче в качестве эмпирической функции распределения рассматривается наблюдаемая функция распределения трансформированных измерений (z-оценок), а в качестве теоретической функции распределения – нормальное с известными параметрами, т. е. N(0,1).

Вычисление супремума выполняется с применением алгоритма сортировки данных по формуле [12]:

$$D_n = \max\left\{\max_{1\le i\le n}\left[\frac{i}{n} - \Phi(z_i)\right], \max_{1\le i\le n}\left[\Phi(z_i) - \frac{i-1}{n}\right]\right\}, \quad (5)$$

где $\Phi(z_i)$ – значение функции стандартного нормального распределения в точке z_i.

Статистика D_n может служить количественной мерой точности подгонки модели в целом, но при этом статистики могут проверять точность подгонки четырех основных характеристик – положения, рассеивания, асимметрии и эксцесса.

Нулевая гипотеза с уровнем значимости 1-α=0,95 на основании известного точного приближения Стефанса отклоняется, если [**Ошибка! Источник ссылки не найден.**3] выполняется условие:

$$(\sqrt{n} + 0{,}12 + 0{,}11/\sqrt{n}) \cdot D_n > 1{,}358. \quad (6)$$

Это позволяет устанавливать пороговое значение принятия модели с заданной доверительной вероятностью – 1-α.

Квантильные функции из полученной LMS-модели вычисляются по формуле:

$$Q_\tau(x) = M(x)\cdot(1 + L(x)\cdot S(x)\cdot z(\tau))^{1/L(x)}, \quad L(x) \ne 0, \quad (7)$$

$$Q_\tau(x) = M(x)\cdot \exp(S(x)\cdot z(\tau)), \quad L(x) = 0$$

где L(x), M(x) и S(x) – функции, составляющие квантильно-регрессионную модель показателя физического развития для данного пола и возраста x;

Z(τ) – квантиль порядка τ стандартного нормального распределения N(0,1) или известное значение z-оценки.

2. Алгоритмическое обеспечение построения LMS-моделей

Основу построения LMS-моделей составляет решение оптимизационной задачи – минимизации логарифма максимального правдоподобия. Сходимость оптимизационных алгоритмов зависит от выбора стартовой точки.

На рис. 1 представлен разработанный интегрированный алгоритм построения LMS-модели. Инициализация LMS-модели включает выбор вида связывающих функций L(x), M(x) и S(x) и задание начальных значений их параметров, а также задание простого штрафа для каждой из степеней свободы, используемой в модели.

Так как связывающие функции M(x), S(x) и L(x) интерпретируются как медиана, коэффициент вариации и степень трансформации, которые могут быть оценены по выборке, то диапазон изменения независимого переменного базисной выборки следует разбивать на множество интервалов, для каждого из которых определяются указанные статистические характеристики.

Для уменьшения сложности представления может быть выбрана полиномиальная функция $M(x) = \sum_{i=1}^{dfm} a_i x^{i-1}$, где dfm – степень свободы этой функции в модели и полиномиальная функция $S(x) = \sum_{i=1}^{dfs} a_i x^{i-1}$, где dfs – степень свободы этой функции в модели.

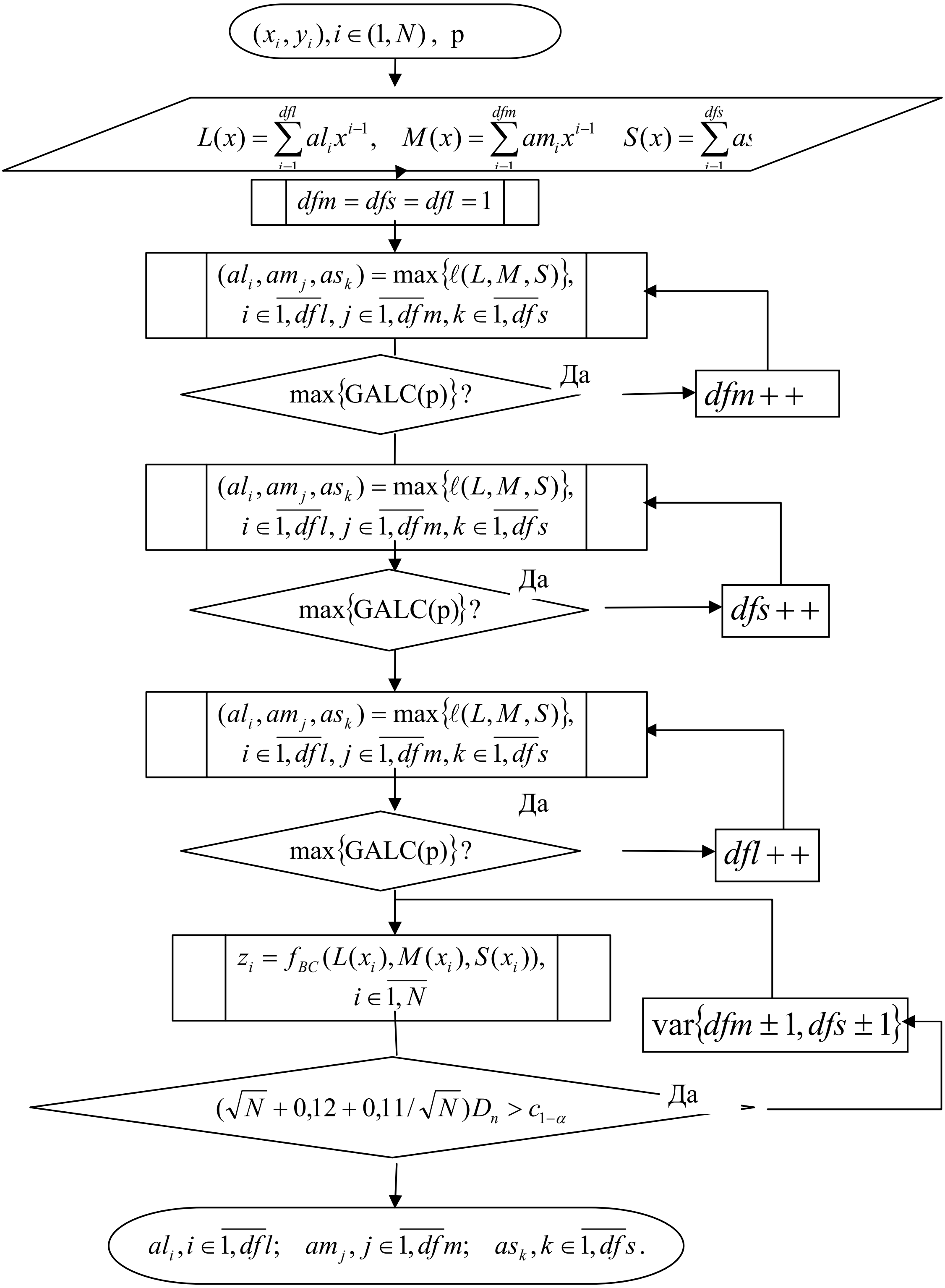


Рис. 1. Алгоритм построения модели на основе трансформации первичных измерений к нормальности

Значение степени трансформации определяется непосредственно по графику зависимости логарифмической функции правдоподобия $L_{\max}(\lambda)$ от λ, которое с точностью до константы равно [7]:

$$L_{\max}(\lambda) = -0.5n \cdot \log(RSS(z^{(\lambda)})),$$

где RSS – остаточная сумма квадратов регрессии значения z на независимую переменную x, z – трансформация зависимого переменного y вида:

$$z_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}}{\dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y}\log y_i, & \lambda = 0 \end{cases},$$

где $\dot{y}$ – среднее геометрическое выборочных значений y_i.

Для инициализации начального значения L(x) ввиду отсутствия однозначного математического определения была разработана дополнительная функция Lmax(X,Y) в среде MathCad, обеспечивающая переборный поиск максимального значения отношения правдоподобия:

Функция формирует массив средних остаточных квадратов ошибок регрессии трансформированных значений диагностического показателя различной степени в пределах от -4 до 6 с шагом 0,25 и возвращает значение коэффициента трансформации, соответствующее максимуму.

Для удобства вычислений в представленном алгоритме использована полиномиальная аппроксимация $L(x) = \sum_{i=1}^{dfl} a_i x^{i-1}$ с dfl числом степеней свободы в модели.

Единичному значению степени свободы соответствовала постоянная величина, равная значению медианы для функции M(x), коэффициента вариации для S(x) и степени трансформации для L(x), вычисленной по (6) в базисной выборке. При числе степеней свободы, равном 2, каждая из связывающих функций представляла простую линейную регрессию и при df>2 – полиномиальную регрессию степени df. Оценка параметров этих функций при df>1 на этапе инициализации осуществлялась МНК по множеству соответствующему статистик – медианы, коэффициента трансформации выборок близлежащих значений x.

Выбор значения простого штрафа p может быть осуществлен в диапазоне от 2 до 8, так как простой критерий Акайке соответствует значению p=2, а критерий Шварца – p=log(n), где n – размер выборки [14].

Второй этап процедуры построения модели состоит в переборе числа степеней свободы dfm, dfs и dfl с шагом 1, и поиске сочетания, обеспечивающего наименьшее значение критерия GAIC(p). Процесс выбора модели начинается с установления значений функции L(x) и S(x), соответствующих dfs=dfl=1 и определения порядка полинома функции M(x), при котором значение GAIC(p) минимально. На втором шаге при выбранной функции M(x) и фиксированном значении L(x) осуществляется поиск порядка полинома функции S(x), минимизирующего GAIC(p).

Завершает этот этап определение значения степени полинома L(x) при выбранных ранее функциях модели M(x) и S(x). На третьем этапе для каждой пары выборочных значений независимой переменной X и зависимого измерения Y по полученным в процессе подгонки модели функциям L(x), M(x), S(x) параметров μ, σ и ν вычисляется значение Z. Как было показано, основным свойством LMS-модели является то, что трансформированные функциями модели значения Z (z-оценки) имеют N(0,1), поэтому настройка модели должна состоять в проверке соответствия полученного распределения трансформированных измерений стандартному нормальному распределению.

$$\mathrm{Lmax}(X,Y) := \left| \begin{array}{l} i \leftarrow 0 \\ \text{for } \ m \in -4, -3.75 \,..\, 6 \\ \quad \left| \begin{array}{l} Z_{i,0} \leftarrow \left| \begin{array}{l} r1 \leftarrow \mathrm{gmean}\left(Y^{m-1}\right) \\ Z \leftarrow \left| \begin{array}{ll} \dfrac{Y^m}{r1} & \text{if } \ m \neq 0 \\ \ln(Y)\cdot \mathrm{gmean}(Y) & \text{otherwise} \end{array}\right. \\ \ln(\mathrm{stderr}(Z,X)) \end{array}\right. \\ Z_{i,1} \leftarrow m \\ i \leftarrow i+1 \end{array}\right. \\ a \leftarrow \mathrm{match}\left(\min\left(Z^{\langle 0\rangle}\right), Z^{\langle 0\rangle}\right) \\ Z_{a_0,1} \end{array}\right.$$

Для сравнения наблюдаемого распределения с теоретическим наиболее эффективен критерий Колмогорова-Смирнова, а именно, проверка условия (6). Этот критерий учитывает расхождение в форме, т. е. модели в целом, и не позволяет определить, какая из функций модели является наиболее вероятной причиной несоответствия модели данным.

Цепочечный (worm) график представляет собой трансформированный квантиль-квантиль график [15]. По вертикальной оси для каждого наблюдения отложено отклонение между его положением в теоретическом и эмпирическом распределениях. Точки данных каждого графика образуют ряд в виде цепочки. Ровный график показывает, что выборочные данные соответствуют предполагаемому распределению. Отклонения графика от прямой, проходящей вдоль оси X показывают, насколько данные отличаются от предполагаемого распределения.

Цель процесса подгонки модели к данным сводится к выравниванию точек цепочечного графика и устранению их флуктуаций. Анализ результатов проведенного моделирования [16] позволил предложить интерпретацию различных форм цепочечных графиков для изменения статистических характеристик – положения, рассеивания и формы, представленную в таблице 1.

Процесс подгонки модели по цепочечному графику можно совместить с вычислением точного значения статистики Колмогорова-Смирнова для проверки гипотезы о нормальности распределения. Для этого по вертикальной оси следует откладывать разницу эмпирической и теоретической функции распределения, рассчитанную по (5) для каждого значения z-оценки, а по оси X отражать значения самой z- оценки, полученной трансформацией (3) первичного измерения с применением анализируемой LMS-модели. Предложенная модификация цепочечных графиков позволяет обеспечить визуализацию процесса настройки модели с применением мощного статистического критерия Колмогорова-Смирнова.

Таблица 1. Интерпретация видов цепочечных графиков при подгонке модели к данным

Вид цепочечного графика		***Схема подгонки модели***	
Форма	***Положение***	***Характеристика***	***Изменение***
горизонтальная прямая	под осью абсцисс	среднее	уменьшение
	над осью абсцисс		увеличение
наклонная прямая	наклон положительный	СКО	увеличение
	наклон отрицательный		уменьшение
парабола	прямая	асимметрия	увеличение
	инвертированная		уменьшение
синусоида	прямая	эксцесс	уменьшение
	инвертированная		увеличение

На этапе настройки модели выбранные на предыдущем этапе степени свободы функций L(x), M(x) и S(x) варьируются в пределах ±1 и для каждой их комбинации совместно оцениваются все параметры LMS-модели для расчета z-оценок. Завершение этого этапа осуществляется при достижении нулевых значений среднего и коэффициентов асимметрии и эксцесса трансформированных значений измерений, а также минимизации значения статистики Dn, лежащей в основе критерия Колмогорова-Смирнова.

Проведенные экспериментальные исследования показали, что если велико значение эксцесса, то следует изменить вид аппроксимации связывающей функции M(x), если велико значение статистики Dn, то необходимо изменение функции S(x), и если распределение z-оценок асимметрично, то следует увеличить количество степеней свободы функции L(x).

3. Практическая задача – построение LMS-модели мониторинга избыточного веса у детей

Для апробации предложенного алгоритмического обеспечения было осуществлено построение модели массы тела девочек и мальчиков Воронежской области на основе экспериментальных данных базисной выборки, полученной в ходе перекрестного исследования. Базисная группа состояла из 8422 детей в возрасте от 2 до 14 лет, каждое наблюдение было представлено зависимой переменной y_i – измеренное значение показателя физического развития (длины или массы тела) и независимой x_i – хронологический возраст на момент проведения измерений.

Предварительно проведенный анализ выявил линейно-экспоненциальную зависимость медианы от возраста, в связи с этим было выполнено сопоставление полиномиальной и экспоненциальной регрессионной функции с выборочными значениями среднего за каждые 2 месяца и сплайновой интерполяции выборочных средних за каждый год (рис. 2.). Экспоненциальная регрессия практически совпадает с полиномиальной 2 степени, наилучшее приближение к выборочным данным обеспечивается с увеличением степени полинома. Полиномиальная функция $M(x)=\sum_{i=1}^{dfm} a_i x^{i-1}$, где dfm – степень свободы этой функции в модели, была выбрана на этапе инициализации.

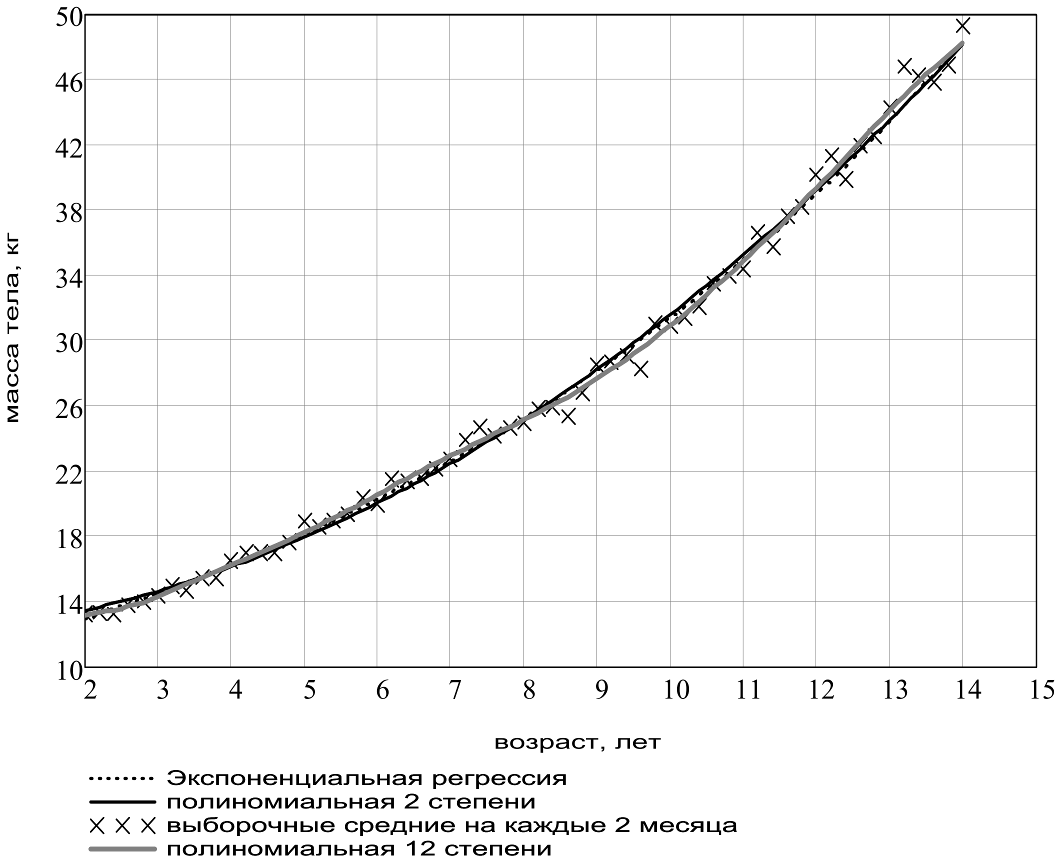
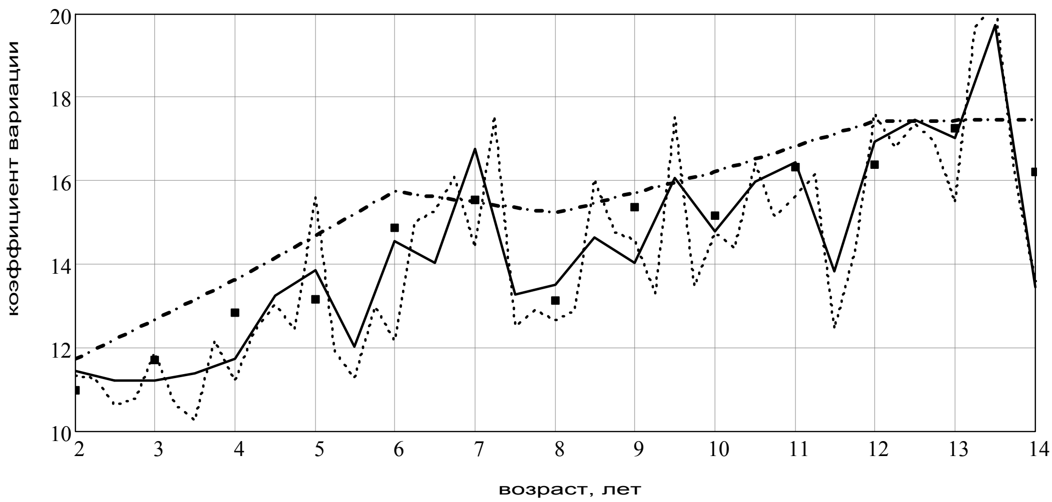


Рис. 2. Выбор вида регрессии для модели массы тела девочек

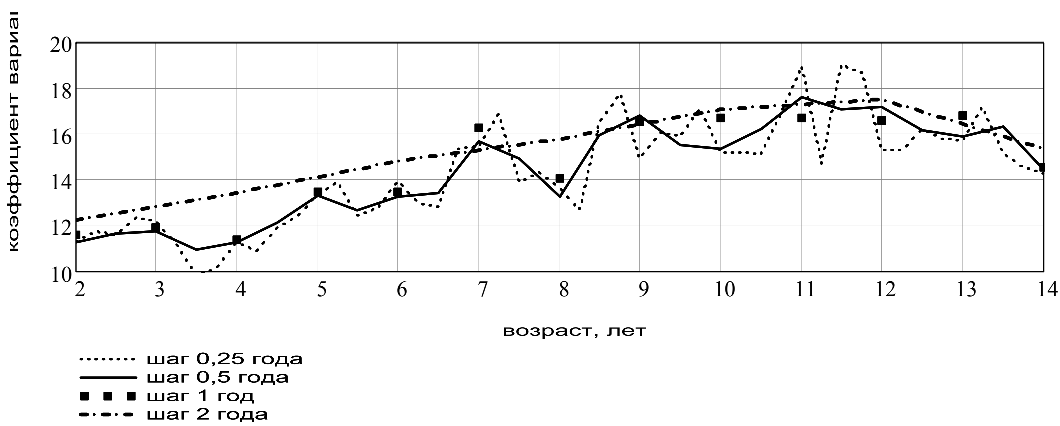
На рисунке 3 представлены зависимости значения коэффициента вариации мальчиков и девочек от возраста. Каждое значение вычислено по выборке, составленной в интервале (t±Δt/2), где t – значение возраста, Δt – устанавливаемый шаг разбиения.

Как видно, выборочный коэффициент вариации неустойчив, сильно изменяется в зависимости от возрастного диапазона формирования выборки и не может быть представлен нелинейной зависимостью одинаковой как для девочек, так и мальчиков. В таком случае могут быть рассмотрены два пути – полиномиальная и сплайновая аппроксимации.

Учитывая, что при высокой степени полинома достигается сходимость к соответствующей сплайновой интерполяции, а представление полиномом просто реализуется вычислительными методами, для связывающей функции S(x) также была выбрана полиномиальная аппроксимация $S(x)=\sum_{i=1}^{dfs} a_i x^{i-1}$, где dfs – порядок полинома и степень свободы этой функции в модели. Используя хронологический возраст $x \cdot 10$, осуществлено разнесение значений регрессора x для получения зависимости степени трансформации Бокса-Кокса в небольшом возрастном диапазоне. На рис. 5 представлены возрастные зависимости значений степени трансформации Бокса-Кокса мальчиков и девочек. Каждое значение определено по выборке, составленной в диапазоне 1, 2 и 3 года от соответствующего возраста.



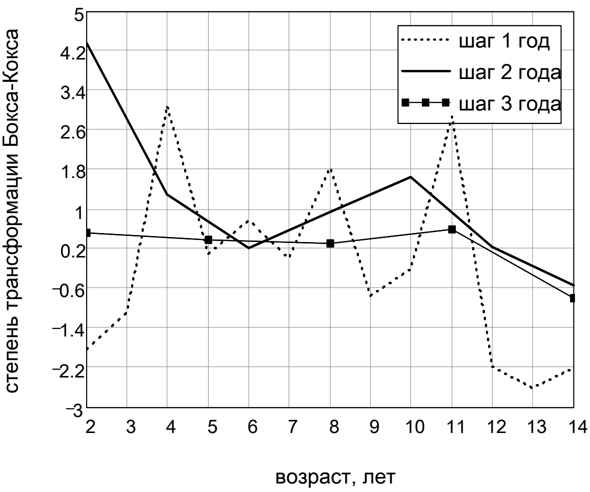
б) мальчиков



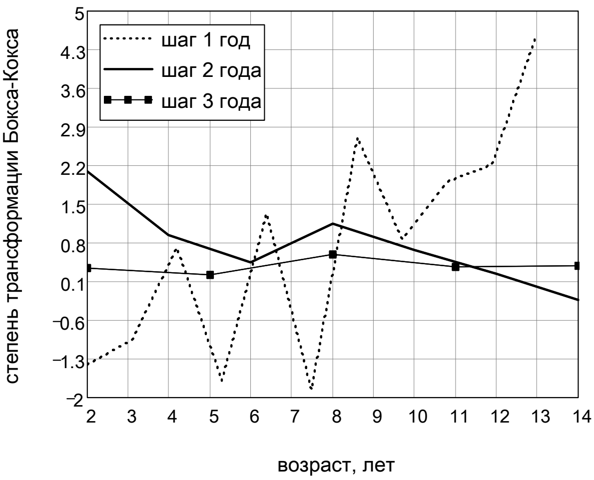
а) девочек

Рис. 3. Возрастные зависимости выборочного коэффициента вариации массы тела детей

На первом этапе построения модели были выбраны полиномиальные функции для M(x), S(x) и L(x) с dfm, dfs и dfl степенями свободы соответственно. В таблице 2 приведены значения констант, инициирующих функции при единичном числе степеней свободы. Инициализация коэффициентов полиномов осуществлялась регрессионными коэффициентами, полученными по выборочным значениям медианы, стандартного отклонения и коэффициента асимметрии первичных измерений у детей каждого полугода жизни.



а) девочек



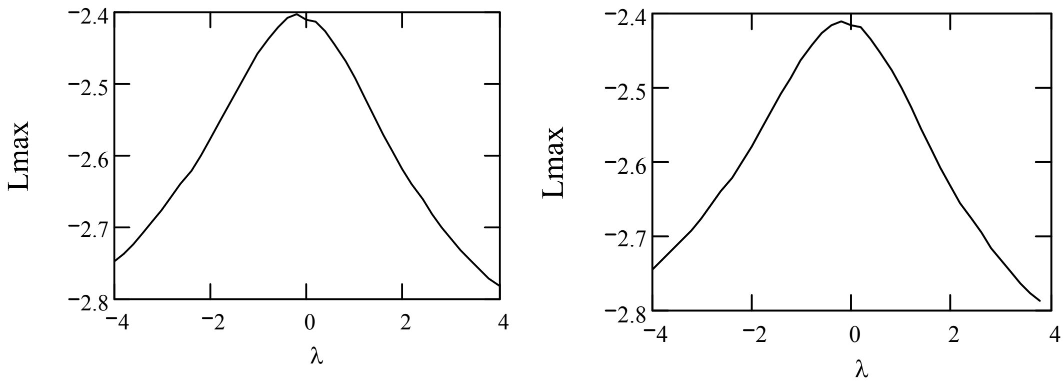
б) мальчиков

Рис. 4. Возрастные зависимости степени трансформации Бокса-Кокса для значений массы тела

Зависимости логарифмической функции максимального правдоподобия от степени трансформации λ массы тела девочек и мальчиков в возрасте от 2 до 14 лет приведены на рисунке 5. Максимум обеих кривых соответствует значению степени трансформации -0,133, которое было использовано для инициализации.

Таблица 2. Постоянные для функций модели M(x), S(x) и L(x) при dfm=dfs=dfl=1

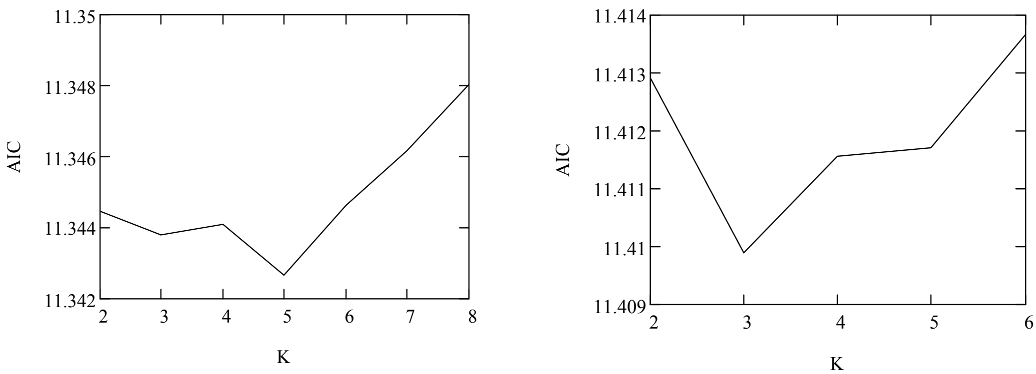
Модель массы	M	S	L
девочек	25,0	0,432	-0,133
мальчиков	24,0	0,434	-0,133



а) девочки б) мальчики

Рис. 5. Зависимость отношения максимального правдоподобия от коэффициента трансформации для показателей массы тела у детей

Для инициализации функций M(x) была определена степень полинома, обеспечивающая минимум критерия Акайке при dfs=dfl=1. Для модели массы тела девочек степень полинома составила dfm=5 и для мальчиков dfm=3. На рисунке 6 приведена зависимость критерия AIC от выбранной степени полинома K регрессии возраста первичных измерений Y.



а) мальчики б) девочки

Рис. 6. Зависимость критерия Акайке от степени K полиномиальной возрастной регрессии массы тела девочек и мальчиков

Простой штраф p=3 был выбран для подгонки модели.

Второй шаг подгонки модели состоял в переборе сочетаний степеней полиномов dfs и dfl с шагом 1 от dfs=dfl=1 до значений, обеспечивающих наименьшее значение GAIC(3) при фиксированном dfm. В таблице 3 и 4 представлены значения критерия GAIC(3) для различных сочетаний dfs и dfl модели девочек и мальчиков.

Таблица 3. Значения критерия GAIC(3) при dfm=6 модели массы тела девочек

	dfs=1	dfs=2	dfs=3	dfs=4	dfs=4	dfs=6
dfl=1	16500	14766	14719	14691	14697	14706
dfl=2	16578	14777	14704	14690	14692	14692
dfl=3	16563	14755	14705	14689	14696	14793
dfl=4	16238	14768	14712	14696	14701	14805

Минимальное значение GAIC(3)=14689 для dfs=4 и dfl=3 было получено для модели массы тела девочек. Минимальное значение GAIC(3)=14895 для dfs=3 и dfl=2 было получено для модели массы тела мальчиков.

Таблица 4. Значения критерия GAIC(3) при dfm=3 модели массы тела мальчиков

	dfs=1	dfs=2	dfs=3	dfs=4
dfl=1	16786	14949	14896	14897
dfl=2	16840	14952	14895	14898
dfl=3	16825	14964	14898	14901

Настройка модели массы тела девочек была проведена при увеличении и уменьшении числа степеней свободы на 1 от первоначально выбранного значения dfm=6 при постоянных выбранных на 2-м шаге значениях dfs=5 и dfl=3, однако, минимизировать значение GAIC(3) больше не удалось. Оценка параметров функций L(x), M(x), S(x) осуществлена максимизацией логарифма максимального правдоподобия по выборочным значениям.

В результате проведенных вычислений была выбрана модель массы тела девочек LMS с 2 степенями свободы для L(x), 6 – для M(x), 4 – для S(x) [18]:

$$M(x) = 14,749 - 3,346x + 1,641x^2 - 0,24x^3 + 0,017x^4 - 4,209 \cdot 10^{-4} x^5;$$

$$S(x) = 0,133 - 0,015x + 3,441 \cdot 10^{-3} x^2 - 1,637 \cdot 10^{-4} x^3;$$

$$L(x) = 0,191 + 0,017x.$$

Настройка модели массы тела мальчиков была проведена при увеличении и уменьшении числа степеней свободы на 1 от первоначально выбранного значения dfm=4 при постоянных выбранных на 2-м шаге значениях dfs=3 и dfl=2. При увеличение dfm на 1 и 2 значение GAIC(3) составило 14895. В результате анализа статистических характеристик трансформированных измерений была выбрана модель массы тела мальчиков с dfm=5 для функции M(x).

Разработанная модель массы тела мальчиков описывается функцией M(x) с 5 степенями свободы, 3 – для S(x) и 2 – для L(x) [18]:

$$M(x) = 11,863 - 0,04x + 0,391x^2 - 0,033x^3 + 1,329 \cdot 10^{-3} x^4;$$

$$S(x) = 0,094 + 7,698 \cdot 10^{-3} x - 1,729 \cdot 10^{-4} x^2;$$

$$L(x) = 0,151 + 2,093 \cdot 10^{-3} x.$$

Для определения оптимального числа степеней свободы моделей массы тела был использован информационный Акайке критерий, обеспечивающий извлечение максимального количества информации из данных. Исследование степени соответствия разработанной модели исходным данным было осуществлено различными статистическими методами.

Из полученных результатов следует, что наибольшие отклонения наблюдаемых и ожидаемых частот встречаемости значений массы тела девочек и мальчиков не превышали 0,05 для всех анализируемых процентилей. Разброс точек носил случайных характер, не наблюдалось выраженной функциональной зависимости и корреляция с возрастом была равна нулю. Для модели массы тела девочек при сравнении по 3-й процентили и у мальчиков по 97-й отмечено преобладание положительных отклонений. Увеличение числа степеней свободы функций моделей на ±1 не привело к уменьшению количества выбросов на графиках сравнения с крайними процентилями.

Поэтому для каждой разработанной модели были рассчитаны различия наблюдаемой частоты встречаемости значений в базисной выборке и ожидаемой, равной порядку квантили, принятой в качестве граничной от 0,05 до 0,95 с шагом 0,05. Максимальная разница составила для моделей: массы тела девочек LMS (2,4,4) – 0,02, массы тела мальчиков LMS (2,4,3) – 0,019.

Согласно выдвинутым при построении модели предположениям, трансформированные по (3) с использованием функций L(x), M(x) и S(x) разработанной модели первичные измерения должны удовлетворять стандартному нормальному распределению. Был проведен анализ формы распределения z-оценок первичных измерений, рассчитанных по полученным моделям. Статистические характеристики полученных z-оценок представлены в таблице 5.

Среднее значение и СКО соответствовали ожидаемым и составляли 0 и 1 соответственно. Медиана совпадала со средним и была равна нулю. Полученные значения находились в симметричных пределах. Значение коэффициента асимметрии статистически значимо и равно 0, а значение эксцесса незначительно превышает 0,5.

Проверка гипотезы о нормальности распределения z-оценок была осуществлена по критерию Колмогорова-Смирнова, с доверительной вероятностью 0,99 нулевая гипотеза была принята. Максимальное расхождение эмпирического и нормального распределений составило для моделей: массы тела девочек LMS (2,4,4) – 0,0211, массы тела мальчиков LMS (2,4,3) – 0,0192. Для всех исследуемых оценка асимметрии по критерию Д’Агостино подтвердила нормальность распределений с уровнем значимости 0,01. Достигнуть статистической значимости эксцесса не удалось. Анализ моделей с увеличением степеней свободы показал, что только выбранные модели имели минимум эксцесса из всех возможных вариантов. Совместная оценка асимметрии и эксцесса по критерию Д’Агостино подтвердила нормальность распределений с уровнем значимости 0,05.

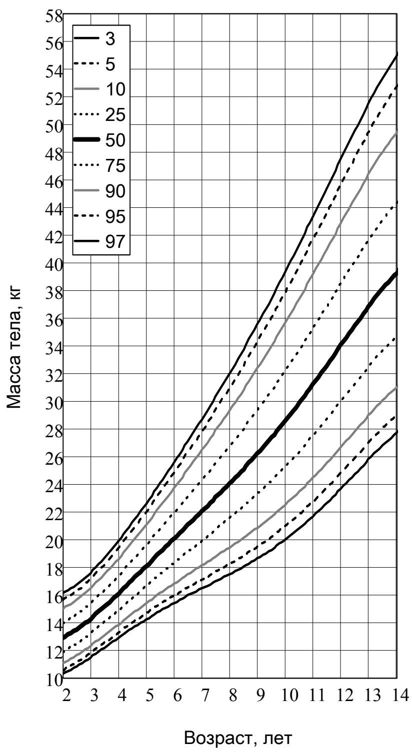
Увеличение степеней свободы на 1 L(x) и S(x) функций моделей не привело к уменьшению амплитуды выбросов и разброса точек. При изменении степени функции M(x) наблюдалось незначительное сглаживание цепочечных графиков и уменьшение максимального значения статистики Колмогорова-Смирнова, при dfm=4 модель LMS (2, 5, 3) составлявшее 0,0181 и dfm=5 модель LMS (2, 6, 3) – 0,0182. Поэтому на основании анализа статистических характеристик трансформированных значений была выбрана модель LMS (2, 5, 3) массы тела мальчиков, вместо LMS (2, 4, 3), оптимальной по информационному GAIC(3) критерию.

Таблица 5. Статистические характеристики значений z-оценок массы детей в возрасте от 2 до 14 лет

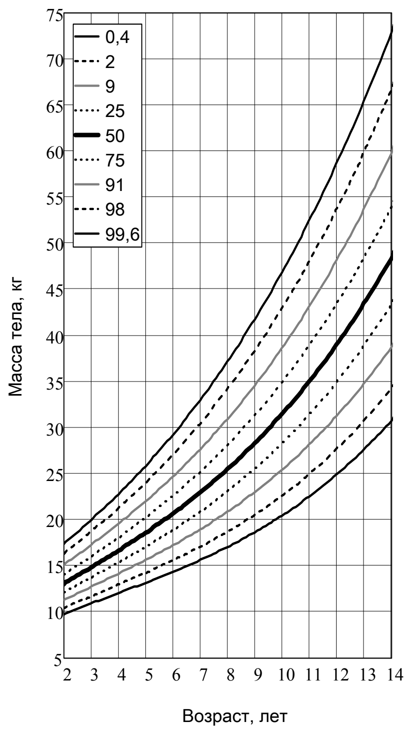
выборка	*среднее*	*СКО*	*медиана*	*максимум*	*минимум*	*асимметрия*	*эксцесс*
девочки (N=4168)	0,00	1,00	0,02	3,67	-3,54	0,03	0,32
мальчики (N=4175)	0,00	1,00	0,00	3,98	-4,51	0,00	0,53

Визуализация процесса настройки модели с помощью цепочечных графиков иллюстративна, но имеет субъективный характер, а использование числовых значений позволяет ускорить процесс принятия решений в пользу той или иной модели, поэтому использования численных оценок статистики Колмогорова-Смирнова и выборочных значений коэффициентов асимметрии и эксцесса при проведении вычислений представляется наиболее оптимальным.

На рисунке 7 представлены разработанные по построенной модели справочные диаграммы массы тела, которые могут быть использованы для мониторинга ожирения среди детской популяции.



А) девочки



Б) мальчики

Рис. 7. Справочные диаграммы развития детей от 2 до 14 лет с применением LMS-метода в различных квантильных режимах

Методы регрессионного анализа могут быть использованы при решении различных практических задач [19-27].

Заключение

В работе проведен анализ математических предпосылок построения полупараметрических моделей. Описано алгоритмическое обеспечение построения LMS-моделей. Рассмотрена практическая задача, связанная с построением LMS-модели мониторинга избыточного веса у детей.

Литература

[1] Орлов А.И. Практическая статистика. – М.: Экзамен. – 2006. – 312 с.

[2] Вентцель Е.С. Теория вероятностей и ее инженерные приложения. – М.: Высшая школа. – 1989. – 282 с.

[3] Афифи А., Эйзен С. Статистиеский анализ. Подход с использованием ЭВМ. – М. Мир. – 1982. – 488 с.

[4] Лемешко Б.Ю. Сравнительный анализ критериев проверки отклонения распределения от нормального закона./ Б.Ю. Лемешко, С.Ю. Лемешко // Метрология. – 2005. – №.2. – С.3.-23.

[5] Кендалл М., Стьюарт А. Теория распределений. – М.: Наука. – 1966. – 587 с.

[6] Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ. – 2006. – 816 с.

[7] Себер Дж. Линейный регрессионный анализ. – М.: Мир. – 1980. – 456 с.

[8] Дейпер Н., Смит Г. Прикладной регрессионный анализ. Т.1. – М.: Финансы и статистика. – 1981. – 392 с.

[9] Cole TJ Smoothing reference centile curves: the LMS method and penalized likelihood/ TJ Cole, PJ. Green//Statistics in Medicine.– 1992. –Vol. 11. –P. 1305-1319

[10] Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, 2003. – 686 с.

[11] Робастность в статистике. Подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П.Рауссеу, В.Штаэль – М.: Мир, 1989. – 512 с.
Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ. – 2006. – 816 с.

[12] ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. – М.: Изд-во стандартов. 2002. – 30 с.

[13] Konishi S. Generalized information criteria in model selection/ S. Konishi, G. Kitagawa//Biometrika. – 1996. – Vol. 4. – P. 875-890.

[14] Van Buuren S. Worm plot: A simple diagnostic device for modeling growth reference curves./ S. Van Buuren, AM Fredriks //Statistics in Medicine. – 2001. –Vol. 20. – P. 1259-1277.

[15] Львович И.Я. Визуализация построения моделей в рамках совершенствования методов интеллектуального анализа данных/ И.Я. Львович, О.В. Минакова// Моделирование систем и информационные технологии: межвузовский сборник научных трудов. Воронеж: АНОО ВИВТ, РосНОУ (ВФ), 2009. – С. 26-29.

[16] Кирьянов Д.В. Mathcad 13. / Д.В. Кирьянов. – СПб.: БХВ-Петербург, 2006. – 608 с.: ил.

[17] Львович И.Я. Определение справочных показателей физического развития детей с применением LMS-метода/ И.Я. Львович, О.В. Минакова, В.П. Ситникова// Вестник ВГТУ. – 2007. - № 10. – С. 96-101.

[18] Klimenko G.Ya. Optimization of medical aid for pregnant women with iron deficiency anemia based on predictive modeling of their health state with due consideration of medical and social risk factors / G.Ya. Klimenko, S.A. Pyataeva, I.Ya. Lvovich, O.N. Choporov, N.V. Naumov. - Lorman, MS, USA. Science Book Publishing House, 2012. - 144 p.

[19] Интегральное оценивание и прогностическое моделирование состояния здоровья беременных, рожениц и родильниц с учетом их медико-социальных характеристик / О.Н. Чопоров, В.П. Косолапов, Н.В. Наумов, Х.А. Гацайниева // Вестник Воронежского института высоких технологий. - 2012. - №9. - С. 91-95.

[20] Моделирование и прогнозирование заболеваемости миомой матки в сочетании с аденомиозом по медико-социальным факторам риска / О.Н. Чопоров, Н.Н. Кудинова, М.В. Фролов, Г.Я. Клименко // Моделирование, оптимизация и информационные технологии. - 2013. - № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Kudinova_soavtori_3_13_1.pdf.

[21] Прогнозирование развития онкологической заболеваемости по индивидуальным медико-социальным факторам риска / О.Н. Чопоров, А.И. Агарков, Г.Я. Клименко, Ю.Ю. Шуршуков // Моделирование, оптимизация и информационные технологии. - 2013. - № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Agarkov_soavtori_3_13_1.pdf.

[22] Бережная Е.В. Оценка риска для здоровья населения г. Воронежа при воздействии химических веществ, загрязняющих атмосферный воздух / Е.В.Бережная // Моделирование, оптимизация и информационные технологии. 2013. № 1. С. 2.

[23] Преображенский Ю.П. Оценка эффективности применения системы интеллектуальной поддержки принятия решений / Ю.П.Преображенский // Вестник Воронежского института высоких технологий. 2009. № 5. С. 116-119.

[24] Львович Я.Е. Принятие решений в экспертно-виртуальной среде / Я.Е.Львович, И.Я.Львович // Воронеж, Издательство "Научная книга", 2010, 139 с.

[25] Гафанович Е.Я. Прогнозирование исходов и выбор рационального лечения артериальной гипертензии с применением математических методов / Е.Я.Гафанович, И.Я.Львович // Вестник Воронежского государственного технического университета. 2013. Т. 9. № 4. С. 84-86.

[26] Калаев В.Н. Регрессионный анализ в биологических исследованиях / В.Н. Калаев, Е.А.Калаева, А.П.Преображенский, О.В.Хорсева // Системный анализ и управление в биомедицинских системах. 2007. Т. 6. № 3. С. 755-759.

[27] Преображенский Ю.П. Применение имитационно-семантического моделирования и полумарковских процессов принятия решений в клинической практике / Ю.П.Преображенский, Н.С.Преображенская // Вестник Воронежского института высоких технологий. 2010. № 6. С. 83-89.

Prof. Igor Lvovich, D. Sc.
Paneuropean University, Bratislava, Slovakia
office@vivt.ru

Prof. Yakov Lvovich, D. Sc.
the honored scientist of the Russian Academy of Natural Sciences
President of Voronezh Institute of High Technologies
office@vivt.ru

Prof. Oleg Choporov, D. Sc.
Voronezh Institute of High Technologies
choporov_oleg@mail.ru

Technique of information database formation for carrying out multilevel monitoring and classificatory-and-forecasting modelling

Методика формирования информационной базы данных для проведения многоуровневого мониторинга и классификационно-прогностического моделирования

O. N. Choporov, A. A. Kurotova, I. I.Manakin
О. Н. Чопоров, А. А. Куротова, И. И. Манакин

Abstract:

The article offers a technique of computer information database formation for carrying out multilevel monitoring and classificatory-and-forecasting modelling. The author considers an algorithm of quality indices transformation in numerical values based on expert assessments, an algorithm of information filtering allowing to exclude doubtful messages, a modified algorithm ZET based on a missing values prediction accounting competent lines and columns of an initial database, an algorithm of the characteristics information value assessment, a procedure of the construction of integrated indices being an additive convolution of several unrelated local components according to their importance. The offered technique is recommended to be used for databases formation for carrying out medical-and-social monitoring and data intellectual analysis.

Аннотация:

Предложена методика формирования компьютерной информационной базы данных для проведения многоуровневого мониторинга и классификационно-прогностического моделирования. Рассмотрен алгоритм преобразования качественных показателей в численные значения, основанный на экспертных оценках; алгоритм фильтрации информации, позволяющий исключить недостоверные сообщения; модифицированный алгоритм ZET, основанный на *предсказании пропущенных значений с учетом компетентных строк и столбцов исходной базы данных; алгоритм оценки информативности характеристик; процедура построения интегральных показателей, являющихся аддитивной сверткой нескольких невзаимосвязанных локальных составляющих с учетом их значимости. Предложенная методика рекомендуется к использованию при формировании баз данных для проведения медико-социального мониторинга и интеллектуального анализа данных.*

Key words:

Database, monitoring, data analysis, classificatory-and-forecasting modelling, expert assessments, blanks filling, information filtering, integrated index.

Ключевые слова:

База данных, мониторинг, анализ данных, классификационно-прогностическое моделирование, экспертные оценки, заполнение пробелов, фильтрация информации, интегральный показатель.

ACM Computing Classification System:

Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

Одной из основных предпосылок, обеспечивающих рациональное планирование и управление является организация интерактивного сбора, поиска, накопления разнородной информации, а также предоставление возможности получения наглядной информации, характеризующей исследуемую систему в реальном масштабе времени, что достигается посредством применения систем мониторинга [9].

При этом мониторинг предлагается рассматривать как многоуровневую систему, например, для системы здравоохранения это может быть региональный (муниципальный) и индивидуальный уровни: на региональном (муниципальном) уровнях анализируются показатели заболеваемости населения, деятельности и ресурсного обеспечения учреждений здравоохранения, а на индивидуальном – медико-социальные факторы риска и состояние здоровья больных [8, 10, 11, 15, 16].

При использовании многоуровневого подхода для системного анализа результатов мониторинга на каждом уровне рассмотрения требуется выбор адекватных методов статистической обработки данных, математического моделирования и принятия решений.

Следует отметить, что точность полученных статистических оценок и адекватность построенных моделей существенно зависит от качества исходных данных, в связи с чем, требуется разработка методики предварительной обработки данных, включающей алгоритмы, направленные на повышение достоверности исходной информационной базы.

Таким образом, организация сбора и предварительной обработки информации является определяющей при проведении многоуровневого мониторинга и классификационно-прогностическом моделировании с точки зрения качества полученных результатов интеллектуального анализа. Объективные оценки не могут быть получены при использовании неполноценного или неполного материала. В связи с этим при сборе первичного материала возникает потребность в предварительном планировании этого процесса, а также в обработке информационной базы для повышения как качества, так и достоверности собранной информации [2, 5, 7, 10].

Этапы формирования информационной базы данных

Как показывает практика, стремление отразить большее количество факторов, характеристик объекта или процесса часто не только не позволяет достичь повышения точности решения поставленной задачи, но и делает модель более громоздкой и трудно воспринимаемой. В связи с этим, уже на первоначальном этапе целесообразно явно четко определить, какие характеристики исследуемого объекта или процесса относятся к существенным, а чем можно пренебречь [6].

Большую часть информации, на основе которой строятся классификационно-прогностические модели, собирают при анализе архивной и текущей информации, получении результатов мониторинга, изучении результатов лабораторных исследований, проведении эксперимента. Показатели, которые измерены в качественной шкале, для дальнейшей статистической обработки требуется преобразовать в количественные оценки. Также необходимо определить, какие параметры управляемые, а какие нет. При этом значение и управляемых, и неуправляемых факторов может меняться со временем.

Таким образом, при формировании компьютерной информационной базы данных для проведения многоуровневого мониторинга и классификационно-прогностического моделирования должны быть реализованы следующие этапы:

1) формирование списка исследуемых показателей, разработка структуры базы данных для мониторинга;

2) проведение сбора фактического материала и заполнение базы данных;

3) преобразование значений качественных характеристик в численные оценки;

4) исключение недостоверных данных (фильтрация информации);

5) заполнение пробелов;

6) оценка информативности и выбор основных контролируемых показателей;

7) разработка интегральных показателей.

На первом этапе разрабатывается структура информационной компьютерной базы. Экспертами определяется перечень показателей для каждого уровня описания исследуемой системы. Например, для системы здравоохранения или отдельных медицинских учреждений на региональном и муниципальном уровнях это показатели распространенности заболеваний, показатели деятельности медицинской службы и ее ресурсного обеспечения; на индивидуальном уровне, для отдельного больного – это данные анамнеза, результаты клинических и лабораторных исследований, а также результаты медико-социального исследования. Следует отметить, что если сбор информации связан с недопустимыми финансовыми или временными затратами, целесообразно сокращение перечня изучаемых показателей, и отбор самых значимых.

Оценка значимости исследуемых показателей на данном этапе выполняется на основе экспертных оценок, обработанных с использованием метода априорного ранжирования.

Следующие этап, связанный с формированием исходной базы данных относится к наиболее трудоемким, так как он связан с процедурами обработкой существенных объемов нормативно-справочной, архивной документации, проведением анкетирования и экспериментальных исследований. Для полноценного анализа динамики анализируемых показателей на региональном уровне и построения прогностических моделей требуются данные, измеренные не менее, чем в 7 временных точках, то есть при регистрации значений показателей по итогам за календарных год, необходимо охватить временной интервал не менее 7 лет.

Необходимое количество объектов базы данных с индивидуальными характеристиками обследуемого контингента населения зависит от требуемой точности и достоверности результатов исследования. Для расчета используется следующее выражение:

$$n = \frac{t^2 \sigma^2}{\Delta^2} \tag{1}$$

где *n* – необходимый объем выборки; *t* – значение *t*-критерия Стьюдента, зависящее от требуемой достоверности (при уровне значимости α=0,05, t=1,96); *σ* – среднеквадратическое отклонение анализируемой характеристики; *Δ* – допустимая ошибка (определяет точность результатов исследования).

Так для получения достоверных данных при уровне значимости α=0,05 и допустимой ошибке в 5%, требуется не менее 400 человек в каждой группе обследованных.

Использование методов статистической обработки информации, построение классификационно-прогностических моделей затруднительно невозможно без использования средств вычислительной техники. Поэтому формирование компьютерной базы данных целесообразно осуществлять с использованием специализированных программ или стандартных электронных таблиц и СУБД. Использование специализированных программ более эффективно, поскольку дает возможность учесть при их создании особенности исследуемого объекта и объединить в рамках одного программного комплекса как средства управления базой данных, так и разработанные модели и алгоритмы. При отсутствии возможности создания специализированных программ возможно использование стандартных электронных таблиц и СУБД (например, MS Access, MS Excel и др.), причем использование СУБД более эффективно, так как при этом существенно облегчается ввод исходных данных, а также имеется возможность формирования сложных запросов для получения требуемой информации.

Алгоритм преобразования качественных показателей в численные оценки

Для последующей обработки значения показателей, представленные в виде смысловых (лингвистических) значении, должны быть преобразованы к численному виду [5-7]. Такое преобразование предлагается осуществлять на основе следующего алгоритма.

1. Показатели, представленные двумя возможными значениями (например, «Да» / «Нет»), преобразуются в 1 и 0.

2. Лингвистические значения показателей упорядочиваются по возрастанию значимости (например, «неудовлетворительно»-«удовлетворительно»-«хорошо»-«отлично»; «легкий»-«средний»-«тяжелый», «нет»-«затрудняюсь ответить»-«да» и т. д.).

Если имеются затруднения или ситуация неоднозначная (например, такая ситуация может возникнуть при оценке семейного положения – «вдовец»-«разведен»-«холост»-«женат»), предлагается к использованию метод априорного ранжирования, который позволяет дать объективную оценку субъективному мнению экспертов (специалистов) [4, 12].

При организации сбора априорной информации m экспертам ($m>7$) предлагаются к заполнению анкеты, в которых требуется дать оценку n значениям показателя с учетом их значимости (при этом наиболее значимому присваивается ранг «1»). В случае, когда эксперты затрудняются с присвоением всем значениям различных рангов, им дается возможность присвоить двум или более различным значениям показателя совпадающие ранги. При наличии совпавших рангов матрицу ранжирования необходимо привести к стандартизованному (нормальному) виду, при котором сумма рангов по каждому столбцу матрицы ранжирования, в котором записаны оценки j-го эксперта ($j=\overline{1,m}$), была бы равна $n(n+1)/2$. Для решения данной задачи всем значениям показателя, которые имеют совпавшие ранги, присваивается ранг, определяемый как среднее значение мест, которые поделили показатели с совпавшими рангами между собой.

На основе матрицы ранжирования осуществляется оценка степени согласованности мнений экспертов с использованием дисперсионного коэффициента конкордации [14]:

$$W=\frac{S\left(d^2\right)}{\frac{1}{12}m^2(n^3-1)-m\sum_{j=1}^{m}T_j}, \tag{2}$$

где

$$S\left(d^2\right)=\sum_{j=1}^{n}\left(\left(\sum_{j=1}^{m}a_{ji}\right)-\frac{1}{2}m(n+1)\right)^2; \tag{3}$$

a_{ji} – сумма рангов i-ого значения показателя;

T_j – показатель связанности рангов, который определяется по формуле

$$T_j=\frac{1}{12}\sum_{i=1}^{n}(t_i^3-t_i); \tag{4}$$

t_j – число повторений в j-том столбце матрицы i-го ранга. В случае, когда в матрице ранжирования отсутствуют совпавшие ранги

$$W = \frac{12S(d^2)}{m^2(n^3-1)}. \tag{5}$$

Значение коэффициента конкордации находится в интервале [0,1]. В случае, когда W=1 эксперты полностью единодушны в своих оценках, при W = 0 согласие экспертов полностью отсутствует.

Для оценки статистической значимости коэффициента конкордации W рассчитывается χ^2 – критерий Пирсона:

$$\chi^2_{рас} = m(n-1)W. \tag{6}$$

В случае, когда табличное (критическое) значение $\chi^2_{табл}$ (при уровне значимости α и числе степеней свободы $f=n-1$) оказывается меньше полученного по формуле (6) (расчетного) $\chi^2_{рас}$, гипотеза о согласованности мнений участников экспертизы принимается.

Полученные таким образом ранги могут быть использованы в качестве численной оценки анализируемого показателя.

3. Если различия двух смежных пар значений анализируемого показателя неравнозначны, выполняется следующий этап преобразования, на котором экспертам необходимо дать оценку, с использованием 5-балльной шкалы, значимости различий $\Delta^j_{i-1,i}$ ($i=\overline{2,n}$, $j=\overline{1,m}$) между смежными («соседними») градациями показателя, которые предварительно должны быть отсортированными в порядке увеличения их значимости.

В случае согласованности мнений экспертов ($\chi^2_{рас} > \chi^2_{табл}$), данные ими оценки усредняются:

$$\Delta_{i-1,i} = \sum_{j=1}^{m} \Delta^j_{i-1,i}, \; i=\overline{2,n} \; . \tag{7}$$

Для вычисления численной оценки значений показателя используется следующее выражение:

$$z_1 = 0, \quad z_i = z_{i-1} + \Delta_{i-1,i}, \quad i=\overline{2,n}, \tag{8}$$

где n – число различных градаций показателя, которые отсортированны в порядке увеличения их значимости.

Для дальнейших расчетов выполняется нормировка полученных на основе выражений (7)- (8) численных оценок:

$$z_i^н = \frac{z_i}{z_n}, \quad i=\overline{1,n} \; . \tag{9}$$

В итоге полученные таким образом численные оценки значений показателя находятся в интервале [0,1], наименее значимому («наихудшему») значению соответствует ноль, а наиболее значимому («наилучшему») – единица.

Алгоритм заполнения пропущенных значений

В ходе заполнения компьютерной информационной базы данных, особенно при использовании архивной информации, возникают «пробелы», которые могут быть обусловлены неполнотой имеющихся данных. Следует отметить, что в большинстве случаев отсутствие даже значения одного показателя приводит к невозможности использования при построении моделей всей остальной информации об объекте моделирования.

Для заполнения пробелов предлагается к использованию модифицированный алгоритм ZET, в основе которого лежат три предположения.

Первое – гипотеза избыточности, которая заключается в том, что в реальных таблицах с данными имеется избыточность, которая проявляется в наличии схожих объектов (строк) и взаимосвязанных свойств (столбцов). В случае, когда избыточность отсутствует (например, в таблице, содержащей случайные числа), отдать предпочтение одному из прогнозов невозможно.

Второе – гипотеза аналогичности, которая заключается в утверждении, что, когда некоторая пара объектов схожа по значениям (n-1) свойств, то она схожа и по n-ому свойству.

Третье – гипотеза локальной компетентности, которая заключается в том, что избыточность имеет локальный характер, то есть для каждого объекта имеется свое подмножество похожих объектов-аналогов и для каждого свойства имеется свое подмножество схожих свойств-аналогов. Если это утверждение не выполняется, то нет смысла привлечения к предсказанию значения некоторого элемента yij информации, содержащейся в строках, не схожих с i-й строкой, и в столбцах, не схожих с j-м столбцом. В получении предсказаний должны использоваться только «компетентные» строки и «компетентные» столбцы, которые для каждого предсказываемого элемента выбираются отдельно. Для использования данного алгоритма, требуется сначала выполнить нормировку всех показателей.

В самом алгоритме можно явно выделить три этапа.

1. Для анализируемого пробела (пропущенного значения) из исходной матрицы, извлекается подмножество «компетентных» строк, а затем для выбранных строк выбирается подмножество «компетентных» столбцов.

2. В автоматическом режиме в выражении, используемом для предсказания пропущенного значения подбираются параметры, при которых ожидаемая ошибка предсказания минимальна.

3. Осуществляется прогнозирование элемента по полученной формуле.

Под «компетентностью» строки i по отношению к строке 1 понимается величина Lil, которая обратно пропорциональна расстоянию между данными строками. Под «компетентностью» j-го столбца по отношению к k-му столбцу Ljk понимается величина, пропорциональная расстоянию между этими столбцами. На основе «компетентных» строк и столбцов строится подматрица, которая должна содержать от 3-х до 7-ми строк и столбцов. При этом компетентная строка (столбец) не должны иметь пропущенного значения в j-м (i-ой) столбце (строке).

Для обеспечения эффективной работы алгоритма предварительно необходимо выполнить нормировку всех показателей. Нормировка представляет собой переход к некоторому единому (стандартизованному) представлению всех признаков, введение новой единицы измерения, которая допускает формальное сопоставление объектов. Как наиболее удобная, рекомендуется нормировка показателей относительно допустимого диапазона изменения их значений.

С использованием зависимостей, существующих между j-ым и всеми остальными (k-ыми) столбцами, на основе уравнений регрессии, в процессе предсказания значения пропущенного значения вырабатываются «подсказки» b(k) = = F(X(k)). При наличии подматрице q+1 столбца, q подсказок усредняются с учетом веса, который пропорционален компетентности соответствующего столбца.

В результате получается «подсказка» (прогнозная величина) b(j), которая порождена избыточностью, содержащейся в столбцах:

$$b(j) = \sum_{k=1}^{q} \left(b(k) \cdot \overset{\alpha}{L}(jk) \right) / \sum_{k=1}^{q} \overset{\alpha}{L}(jk) \qquad (10)$$

Коэффициент α позволяет регулировать влияние на результат предсказания компетентности. Когда значения α небольшие, разница в компетентности мало сказывается на результате, когда коэффициент α принимает большие значения, более компетентные столбцы гораздо больше влияют на результат по сравнению с другими. В выборе параметра α заключается суть второго этапа (подбор формулы для прогнозирования): для всех известных элементов j-ого столбца осуществляется предсказание при разных значениях α. После этого выбирается значение α, при котором удалось достигнуть минимальной ошибки прогноза. Затем с использованием формулы (10) с выбранным значением α осуществляется прогноз b(j) для значения пропущенного элемента.

Алгоритм заполнения пропущенного значения с использованием зависимости между строкой i и всеми s другими (1-ыми) строками (l=1,2,…,s) аналогичен вышеописанному. Прогнозирование осуществляется по формуле

$$b(i) = \sum_{l=1}^{s} \left(b(l) \cdot \overset{\alpha}{L}(il) \right) / \sum_{l=1}^{s} \overset{\alpha}{L}(il) \qquad (11)$$

Здесь выбор параметра α осуществляется на основе всех известных элементов i-ой строки при достижении минимального значения ошибки их предсказания. Для получения общего прогноза y'ij значения пропущенного элемента yij выполняется усреднение прогнозов b(i) и b(j).

Процедура исключения недостоверных данных

Кроме «пробелов» серьезным фактором, оказывающим влияние на результаты моделирования, является присутствие недостоверных данных, которые могут появиться из-за различных субъективных и объективных причин, к которым можно отнести неточность измерения, ошибки в исходной документации, погрешность результате сбоя аппаратуры, ошибки при внесении данных и др.). Для исключения недостоверных данных предназначен этап фильтрации информации, при реализации которого контролируется не только выход показателей за допустимые границы, но и недопустимые сочетания нескольких признаков.

Главным эвристическим правилом, используемым при решении задачи информационной фильтрации, является выбор объектов (информационных сообщений), с наиболее типичным набором характеристик для данной ситуации [10].

Вся исходная информация представляется в виде множества объектов:

$$P_{исх} = \{p_n\}, \quad n = \overline{1, N_{исх}} \tag{12}$$

Для каждого объекта имеется набор показателей, характеризующих его:

$$p_n \to F_n = \{f_n^1, f_n^2, \dots, f_n^i, \dots, f_n^I\}, \tag{13}$$

где $i = \overline{1, I}$ – индекс показателя, $n = \overline{1, N_{исх}}$ – номер объекта.

Решением задачи фильтрации является отбор из исходного множества объектов с оценкой достоверности их характеристик w_n ($n = \overline{1, N}$) выше некоторой заданной величины w_0. Показатели, содержащие смысловые (лингвистические) значения на основе экспертных оценок преобразуются в численные оценки.

Для объектов, характеристики которых представлены в виде численных значений, решение задачи информационной фильтрации заключается в отображении множества объектов (12) во множество оценок достоверности характеристик исходных объектов

$$W = \{w_n\}, \quad n = \overline{1, N_{исх}} \tag{14}$$

и формирование множества $P \subseteq P_{исх}$, включающего отобранные (отфильтрованные) объекты, достоверность характеристик которых больше величины w_0.

В большинстве случаев различные показатели измеряются в разных единицах. Для повышения эффективности работы алгоритма фильтрации требуется выполнить нормировку всех исследуемых показателей. Рекомендуется использовать нормировку относительно допустимого диапазона значений показателей.

Оценка степени достоверности характеристик объекта при решении задачи (12)-(14) основывается на концепции типичности, т.е. считается, что достоверность w_n характеристик объекта тем выше, чем его характеристики типичнее для данной группы объектов. Так как сведения о пациентах $\{p_n\}$ представлены в виде численных значений, может быть использован геометрический подход, при котором объекты с их характеристиками рассматриваются как «созвездия» в i-мерном пространстве признаков [10]. При этом способ решения выбирается в зависимости от априорных сведений о мере «засоренности» исходной выборки объектов $P_{исх}$.

В случае, когда априорно известно о малой «засоренности» выборки $P_{исх}$, правомерно предположение о том, что объекты p_n сгруппированы симметрично относительно некоторого центра тяжести (объекта со средними характеристиками), и наиболее достоверные сообщения с большей вероятностью располагаются на наименьшем расстоянии от этого гипотетического обобщенного объекта p_0 с набором характеристик $p_0 \to F_0 = \left\{ f^1_0, f^2_0, \ldots, f^i_0, \ldots, f^I_0 \right\}$.

Решение получают путем вычисления значений вектора расстояний $S = \{S_1, S_2, \ldots, S_n, \ldots, S_{N_{\text{исх}}}\}$ от объекта p_n до обобщенного объекта p_0 с использованием адекватной по отношению к представленным сведениям метрики, например, с помощью евклидовой.

Для определения степени достоверности характеристик объекта p_n используется следующее выражение:

$$w_n = S_{\min} / S_n, \tag{15}$$

где $S_{\min} = \min\limits_{\forall n} S_n$.

В случае, когда выборка $P_{исх}$ значительно «засорена», более правомерно предположение об асимметричном распределении объектов и понятие обобщенного объекта p_0 не может адекватно представить выборку $P_{исх}$. При этом предлагается другая процедура вычисления степени достоверности характеристик объектов: с использованием адекватной метрики, выбранной исследователем, рассчитываются значения вектора суммарных расстояний $S = \{S_1, S_2, \ldots, S_n, \ldots, S_{N_{\text{исх}}}\}$ от каждого объекта до всех остальных, и так же, по формуле (15), вычисляется степень достоверности характеристик объекта.

Следует учесть, что объем множества $P \subseteq P_{исх}$ отфильтрованных объектов существенно зависит от выбранного значения w_0.

Формирование интегрального показателя

Следующий этап формирования информационной компьютерной базы данных состоит в выборе одного или набора наиболее информативных контролируемых показателей, которые характеризуют состояние моделируемой системы.

Информативность отбираемых показателей предлагается измерять как сумму коэффициентов парной корреляции по модулю [3, 7], или других показателей связи (коэффициентов взаимной информации, коэффициентов взаимной сопряженности и т. п.). Следует отметить, что у данного подхода имеется недостаток: значения коэффициентов представляют собой конкретные реализации вероятностного процесса, и, соответственно, небольшая разница их значений может иметь случайный характер.

С учетом этого обстоятельства, определению информативного признака как наиболее «влиятельного» признака, в наибольшей степени соответствует дискретный показатель. Предложено два подобных показателя: количество значимых связей данного признака с другими и число связей в дендрите, который строится для рассматриваемой системы признаков. Данные показатели лучше будут характеризовать уровень связанности признака и его информативность [7].

В случае, когда отсутствует отдельный показатель, позволяющий адекватно описать состояние ситуации на индивидуальном или региональном уровне, а по набору показателей оценка затруднена, строится интегральный показатель, представляющий собой свертку нескольких не связанных между собой локальных составляющих с учетом степени их значимости.

При построении интегрального показателя сначала формируется список характеристик объекта исследования. Далее с использованием метода априорного ранжирования [4, 12] и метода «дискретных корреляционных плеяд» [10] осуществляется отбор минимального количества наиболее значимых, но не связанных друг с другом показателей, отражающих ситуацию.

Для каждого показателя осуществляется нормировка или разрабатывается система балльных оценок [1].

Для расчета интегрального показателя предлагается следующая формула:

$$\text{ИП} = \sum_{\text{i}=1}^{\text{N}} w_i X_i^{\text{н}} \tag{16}$$

где N – количество показателей, включенных в интегральный;

w_i – значимость (вес) i-го показателя,

$X_i^{\text{н}}$ – нормированное значение (балльная оценка) i-го показателя.

Для определения значимости каждого показателя, включенного в интегральный, используется метод априорного ранжирования [4, 12].

Значения весов w_i определяются по формуле

$$w_i = \frac{m \cdot (n+1) - \sum_{j=1}^{m} r_{ij}}{\sum_{i=1}^{n} \left(m \cdot (n+1) - \sum_{j=1}^{m} r_{ij} \right)}, \; i = \overline{1,n}. \tag{17}$$

где r_{ij} ($j = \overline{1,m}$) – ранг, выставленный j-м экспертом.

С учетом того, что $\sum_{i=1}^{n} w_i = 1$, а показатели, включенные в ИП, приведены к нормированному виду или оценены с использованием k-балльной шкала, максимальное значение интегрального показателя соответствует верхней границе нормировки (обычно равной +1) или максимальному баллу (k), а минимальное значение равно нижней границе нормировки (обычно это 0) или минимальному баллу (обычно +1).

Представленный интегральный показатель дает возможность провести комплексную оценку состояния моделируемой системы любого уровня.

Формирование интегрального показателя

Таким образом, для повышения качества информационной базы для многоуровневого мониторинга и классификационного моделирования, процесс ее формирования должен включать следующие этапы: формирование списка исследуемых показателей, разработка структуры базы данных для мониторинга; проведение сбора фактического материала и заполнение базы данных; преобразование значений качественных характеристик в численные оценки; исключение недостоверных данных (фильтрация информации); заполнение пробелов; оценка информативности и выбор основных контролируемых показателей, отражающих исследуемую систему; разработка интегральных показателей.

Предложенная методика была апробирована при формировании баз данных для проведения медико-социального мониторинга и интеллектуального анализа данных.

Литература

[1] Использование балльной оценки для формирования интегрального показателя состояния здоровья населения / Г.Я. Клименко, И.Э. Есауленко, В.П. Косолапов и др. // Бюллетень НИИ соц. гигиены, экономики и управления здравоохранением им. Н.А. Семашко. – Москва, 2003, С. 18-22.

[2] Клименко Г.Я. 42. Методика и результаты преобразования лингвистических характеристик в численные оценки факторов риска / Г.Я. Клименко, В.П. Косолапов, О.Н. Чопоров // Журн. «Консилиум», – Воронеж, 2001, №4. С. 25-28.

[3] Кудинова Н.Н. Результаты анализа значимости медико-социальных факторов риска развития миомы матки в сочетании с аденомиозом / Н.Н. Кдинова, П.Е. Чесноков, О.Н. Чопоров // Вестник Воронежского института высоких технологий. – 2013. - №11. – С. 202-206.

[4] Львович Я.Е. 62. Моделирование биотехнических и медицинских систем / Я.Е. Львович, М.В. Фролов // Под ред. В.Н. Фролова: учеб. пособие. – Воронеж: Изд-во ВГТУ, 1994.

[5] Методика 67. преобразования качественных характеристик в численные оценки при обработке результатов медико-социального исследования / О.Н. Чопоров, А.И. Агарков, Л.А. Куташова, Е.Ю. Коновалова // Вестник Воронежского института высоких технологий. – Воронеж, 2012. - №9. – С. 96-98.

[6] Методы 68. предварительной обработки информации при системном анализе и моделировании медицинских систем / Л.А. Куташова, О.Н. Чопоров, Н.В. Наумов, А.И. Агарков // Врач-аспирант. – 2012. – № 6.2 (55). – С. 382-390.

[7] Методы 69. предварительной обработки информации при системном анализе и моделировании медицинских систем / О.Н. Чопоров, Н.В. Наумов, Л.А. Куташова, А.И. Агарков // Врач-аспирант. – № 6.2 (55). – 2012. – С. 382-390.

[8] Моделирование и прогнозирование заболеваемости с миомой матки в сочетании с аденомиозом по медико-социальным факторам риска / О.Н. Чопоров, Н.Н.Кудинова, М.В. Фролов, Г.Я. Клименко // Электронный научный журнал «Моделирование, оптимизация и информационные технологии». – 2013. - № 3. http://moit.vivt.ru/wp-content/uploads/2014/01/Kudinova_soavtori_3_13_1.pdf.

[9] Оптимизация управления функционированием медицинских систем различного уровня / О.Н. Чопоров, И.Я. Львович, К.А. Разинкин, А.А. Рындин // Системы управления и информационные технологии. – 2013. – Т. 53. - №3. – С. 100-104.

[10] Проблемы 80. здравоохранения промышленно-развитого региона в современных условиях / И.Э. Есауленко, Г.Я. Клименко, В.Н. Созаева, О.Н.Чопоров. – Воронеж: Изд-во ВГУ, 1999. – 263 с.

[11] Прогнозирование развития онкологической заболеваемости по индивидуальным медико-социальным факторам риска / О.Н. Чопоров, А.И. Агарков, Г.Я. Клименко, Ю.Ю. Шуршуков // Электронный научный журнал «Моделирование, оптимизация и информационные технологии». – 2013. - № 3. http://moit.vivt.ru/wp-content/ uploads/2014/01/ Agarkov_soavtori_3_13_1.pdf

[12] Управление 95. в биологических и медицинских системах: Учеб. пособие / О.В. Родионов, Е.Д. Федорков, В.Н. Фролов, М.В. Фролов. Под ред. д-ра техн. наук, проф. Я.Е. Львовича. – Воронеж: ВГТУ, 2002. – 342 с.

[13] Чопоров О.Н. 98.Методы анализа значимости показателей при классификационном и прогностическом моделировании / О.Н. Чопоров, А.Н. Чупеев, С.Ю. Брегеда // Вестник Воронежского государственного технического университета. – 2008. – Т.4, №.9. – С. 92-94.

[14] Юнкеров В.И. Математико-статистическая обработка данных медицинских исследований / В.И. Юнкеров, С.Г. Григорьев. – СПб.: ВМедА, 2002. – 266 с.

[15] Classification of territorial units of region on the level of desease of the adult female population the myoma of the uterus and the endometriosis on the basis of geoinformation technologies / E.N. Korovin, N.N. Kudinova, M.V. Frolov, O.N. Choporov // Information Technology Applications. – Slovakia, 2013. - № 4. – C. 74-81.

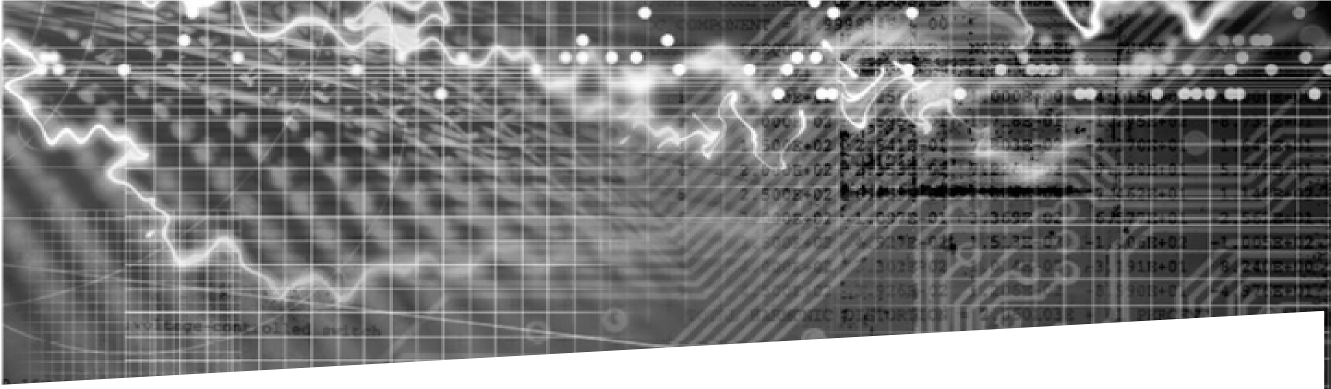
[16] Working out of information subsystem of forecasting of malignant new growths development and health state of oncological patients upon medical and social risk factors / O.N. Choporov, A.I. Agarkov, G.Ja. Klimenko, V.G. Medintsev // Information Technology Applications. – Slovakia, 2013. - № 4. – C. 41-54.

....

Prof. Oleg Choporov, D. Sc.
Voronezh Institute of High Technologies
choporov_oleg@mail.ru

Alena Kurotova
postgraduate University of Economics, Bratislava, Slovakia

Ivan Manakin
postgraduate Voronezh State Medical Academy named after N.N. Burdenko



The possibilities of improvement wireless coverage inside buildings

Возможности улучшения покрытия беспроводных сетей внутри зданий

I. Y. Lvovich, A. P. Preobrazhensky
И. Я. Львович, А. П. Преображенский

Abstract:

The analysis approaches that can be used for analysis of electromagnetic environment in the premises on the basis of computer modeling is given. The advantages of simulation modeling are indicated. The algorithm of operation of the system improving the coverage of wireless networks is given. The basic steps of the optimization algorithm are shown. The composition of a software product using the developed algorithm is described. The program is designed on the basis of a modular and hierarchical structure, the characteristics of the modules are described.

Аннотация:

Проведен анализ подходов, которые могут быть использованы для анализа электромагнитной обстановки в помещении на основе компьютерного моделирования. Отмечены достоинства имитационного моделирования. Приведен алгоритм работы системы улучшения покрытия беспроводных сетей. Указаны основные шаги алгоритма оптимизации. Описан состав программного продукта. Программа сформирована на основе модульно-иерархической структуры.

Key words:

Wireless network optimization, information transfer, model, coverage.

Ключевые слова:

Беспроводная сеть, оптимизация, передача информации, модель, покрытие.

ACM Computing Classification System:

Wireless devices, Wireless integrated network sensors, Reconfigurable computing

Введение

Задача, связанная с передачей информации по беспроводному каналу во многих случаях решается на основе численных методов. Это определяет трудности формирования требуемого покрытия беспроводной связи в помещении. В работе рассматривается один из возможных подходов к улучшению беспроводной связи на основе соответствующих алгоритмов.

В связи с тем, что все более в помещениях используют беспроводные сети, к настоящему времени было создано большое число программных средств для того, чтобы осуществлять прогнозирование качества принятых сигналов и производительности сети [1, 2]. В качестве основы используют или лучевые модели, численные модели, эвристические прогнозы и статистические модели.

1. Используемые подходы

Рассмотрим некоторые из применяемых подходов. В работе [3], авторами был разработан эвристический способ для проведения прогнозирования распространения электромагнитных волн, в рамках которого можно проводить разработку и оптимизацию сети Wi-Fi при определенных требованиях к покрытию с минимальным числом точек доступа.

При этом существует необходимость к снижению излучения электромагнитных волн в беспроводных системах связи [4]. Разработаны базовые принципы международной безопасности, например [5] и ICNIRP (Международной комиссией по защите от неионизирующих излучений) [6] были созданы и распространены нормы по ограничению воздействия таких излучений на людей.

Вследствие этого важно проводить оценку характеристик распространения электромагнитных волн в беспроводных системах связи с высокой точностью при заранее известных перспективах развития компьютерных сетей.

Кроме того, работы, проводящиеся по ограничению потребления энергии в беспроводных системах связи [7], опасения по поводу возможного вредного воздействия воздействие на человека источников электромагнитного излучения привели к ситуации, когда планировщики сети должны брать на себя ответственность за уровень электромагнитной обстановки в помещении.

Тем не менее, многие исследования по-прежнему посвящены изучению простых характеристик при воздействия СВЧ-излучения в различных средах или для разных технологий [8], или для MIMO [9] сетей и терминалов, без того чтобы осуществить попытки по фактическому сокращению или минимизации воздействия таких видов излучений [10, 11]

Есть подходы, в которых исследователи пытаются предсказать или смоделировать влияние до развертывания сети. Например, в [12], предлагают метрику оценивать по воздействию на окружающую среду при развертывание сети. При этом можно проводить оптимизацию базовой станции на основе метрики [13], которая учитывает не только характеристики освещения, трафика и экономической эффективности, но и того, каким образом воздействует на окружающую среду электромагнитное поле, есть работы, связанные с применением методов искусственного интеллекта [14].

На основе компьютерного моделирования можно не только проводить разработку нового оборудования для вычислительных сетей, но и осуществлять проектирование этих сетей.

При разработке модели, которая аппроксимирует свойства и характеристики рассматриваемой сети, стремятся к тому, чтобы она позволяла решать задачи, связанные оптимизацией и управлением этой сетью.

Среди различных подходов, связанных с моделированием компьютерных сетей можно выделить имитационное моделирование. Имитационная модель представляет собой логико-математическое описание объекта, в рамках которого можно проводить экспериментирование на компьютере с целью осуществления процессов проектирования, анализа, а также проведения оценок того, как функционирует объект.

В рамках имитационного моделирования рассматриваются модели, которые описывают процессы таким образом, каким они бы происходили на практике. Построенную модель можно рассматривать во времени как при одном испытании, так и для определенного их множества. Понятно, что результаты определяются исходя из случайного характера процессов. Исходя из этих данных, есть возможность получения устойчивой статистики.

2. Особенности имитационного моделирования

Имитационное моделированию применяют для случаев, когда:

- эксперименты на реальном объекте являются дорогостоящими или их трудно провести;

- трудно сформировать аналитическую модель в связи с тем, что необходимо учитывать множество факторов: зависимость от времени, нелинейность процессов, влияние случайных факторов.

Среди видов имитационного моделирования можно выделить такие: агентное моделирование, дискретно-событийное моделирование, системная динамика.

Среди преимуществ имитационного моделирования можно отметить:

1. Выигрыш по стоимости. Проведя машинный эксперимент можно оценить те затраты, которые получит компания, проведя какие-либо шаги по реорганизации компьютерной сети.

2. Выигрыш по времени. Имитационная модель дает возможности оценки степени оптимальности каких-либо изменений в компьютерной сети, буквально за минуты.

3. Повторяемость процессов. На основе имитационной модели есть возможности проведения довольно большого количества экспериментов при различных параметрами, для того чтобы найти наилучший вариант.

4. Высокая точность. При использовании обычных расчетных математических методов не всегда можно учесть важные характеристики в системе, а на основе имитационного моделирования процессы описываются в естественном виде.

5. Хорошая наглядность. В рамках имитационной модели можно визуально наблюдать процесс работы системы во временной области, результаты выдаются в графическом виде.

6. Универсальность. На основе имитационного моделирования могут быть решены задачи из самых разных областей.

Одной из важных задач при обеспечении правильной работы беспроводной сети является создание минимально необходимого уровня поля в различных помещениях. На рисунке 1 приведен предлагаемый алгоритм работы системы улучшения покрытия беспроводных сетей, созданный в рамках имитационного моделирования.

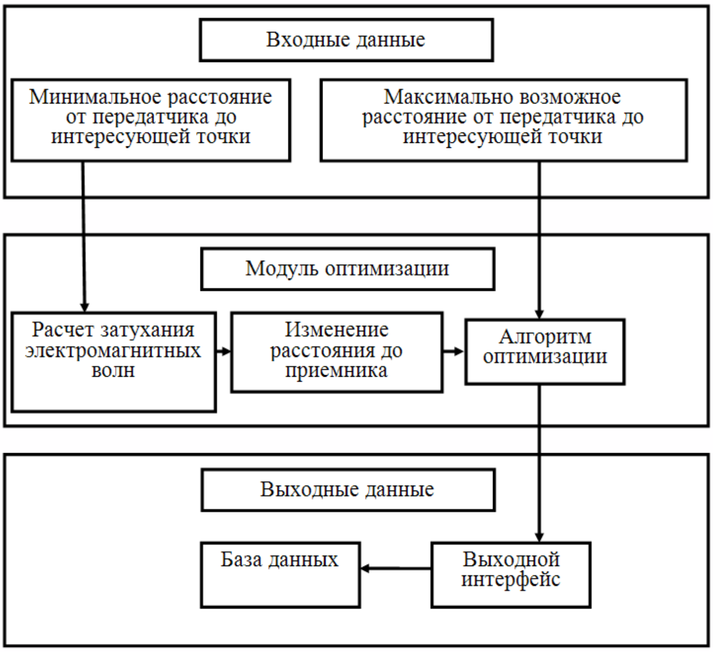


Рис. 1. Алгоритм работы системы улучшения покрытия беспроводных сетей

На рисунке 2 приведен алгоритм оптимизации. Метрика может быть рассчитана различными способами. Нами предлагается метрика, которая основывается на отношении суммы взвешенных уровней электромагнитного поля к расстояниям до точек наблюдения.

Рис. 2. Основные шаги алгоритма оптимизации

В работах [15-32] приведены алгоритмы, которые могут быть использованы при оценке характеристик распространения электромагнитных волн в помещениях.

3. Особенности программного продукта

Опишем состав программного продукта. Исполняемый файл программы может быть расположен на любом разделе жесткого диска или съемного носителя.

В состав проекта входят такие файлы:

1.Файл программного модуля главной формы – определяет функциональность главной формы и содержит вычислительные процедуры;

2. Файл программного модуля формы вывода информации о программе – определяет функциональность формы вывода информации о программе;

3. Файл проекта – осуществляет связь всех файлов, из которых состоит приложение;

4. Файл вычислительного модуля - включает подпрограммы определения максимального потока в сети и минимального разреза;

5. Файл программного модуля формы вывода графического изображения потока или разреза - определяет форму вывода потока или разреза;

6. Файл программного модуля формы вывода описания программы – определяет форму для вывода описания программы (помощь).

Кратко опишем структуру программы. В программе задается максимальное число вершин сети. Определяются одномерные и двумерные массивы, в которых хранится информация о составляющих сети.

Программа сформирована на основе модульно–иерархической структуры. Связь модулей происходит с использованием простого параметра. Все необходимые данные модуль принимает и возвращает в форме параметров вызова, а эти данные являются простыми (неструктурными) переменными.

Программные модули располагаются на различных уровнях иерархии. Модули верхних уровней проводят управление работой модулей нижних уровней. Вышестоящий модуль производит передачу управления модулю более низкого уровня, а когда тот заканчивает работу, он возвращает управление вызвавшему его модулю.

Заключение

В работе рассмотрен алгоритм, позволяющий проводить оптимизацию покрытия беспроводных систем связи в помещениях. Даны основные шаги алгоритма оптимизации. Описан программный продукт, использующий такой алгоритм.

Литература

[1] Ji, Z., B.-H. Li, H.-X. Wang, H.-Y. Chen, and T. K. Sarkar, Efficient ray-tracing methods for propagation prediction for indoor wireless communications," IEEE Antennas and Propagation Magazine, Vol. 43, No. 2, 41-49, April 2001.

[2] Torres, R., L. Valle, M. Domingo, and M. Diez, CINDOOR: an engineering tool for planning and design of wireless systems in enclosed spaces," IEEE Antennas and Propagation Magazine, Vol. 41, No. 4, 11-22, September 1999.

[3] Plets, D., W. Joseph, K. Vanhecke, E. Tanghe, and L. Martens, Coverage prediction and optimization algorithms for indoor environments," EURASIP Journal on Wireless Communications and Networking, Special Issue on Radio Propagation, Channel Modeling, and Wireless, Channel Simulation Tools for Heterogeneous Networking Evaluation, Vol. 1, 2012, http://jwcn.eurasipjournals.com/content/2012/1/123

[4] Plets, D., W. Joseph, E. D. Poorter, L. Martens, and I. Moerman, Concept and framework of a self-regulating symbiotic network," EURASIP Journal on Wireless Communications and Networking, Vol. 2012, No. 340, 2012, http://jwcn.eurasipjournals.com/content/2012/1/340.

[5] IEEE Std C95.1, IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic Fields, 3 kHz to 300 GHz," 1999.

[6] ICNIRP, // Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic Fields (up to 300 GHz)," Health Physics, Vol. 74, No. 4, 494-522, April 1998.

[7] Deruyck, M., E. Tanghe, W. Joseph, and L. Martens, // Modelling and optimization of power consumption in wireless access networks," Elsevier Computer Communications (Special Issue: European Wireless 2010), Vol. 34, No. 17, 2036-2046, November, 2011.

[8] Foster, K. R., Radiofrequency exposure from wireless LANs using Wi-Fi technology," Health Physics, Vol. 92, 280-289, 2007.

[9] Perentos, N., S. Iskra, A. Faraone, R. McKenzie, G. BitBabik, and V. Anderson, // Exposure compliance methodologies for multiple input multiple output (MIMO) enabled networks and terminals," IEEE Transactions on Antennas and Propagation, Vol. 60, 644-653, 2012.

[10] Golovinov S.O., Preobrazhenskii A.P., Lvovich I.Y. Modeling the millimeter wave propagation in urbanized areas based on a combined algorithm // Telecommunications and Radio Engineering. – 2013. – T. 72. – № 2. – C. 139-145.

[11] Баранов А.В. Проблемы функционирования mesh-сетей // Вестник Воронежского института высоких технологий. – 2012. № 9. С. 49-50.

[12] Russo, P., G. Cerri, and V. Vespasiani, A numerical coefficient for evaluation of the environmental impact of electromagnetic .elds radiated by base stations for mobile communications," Bioelectromagnetics, Vol. 31. – 613-621, 2010.

[13] Cerri, G., R. De Leo, D. Micheli, and P. Russo, Base-station network planning including environmental impact control," IEE Proceedings | Communications, Vol. 151. – No. 3, 197-203, June 2004.

[14] Koutitas, G., Green network planning of single frequency networks," IEEE Transactions on Broadcasting, Vol. 56 – №. 4. – 541-550, December 2010.

[15] Львович Я.Е. Экспертно-оптимизационное моделирование кластерного разделения объектов сетевой системы / Я.Е. Львович, С.О. Сорокин // Вестник Воронежского института высоких технологий. – 2014. – № 13. – С. 49-52.

[16] Преображенский А.П. Об оценке характеристик беспроводной связи в помещении / А.П. Преображенский // Вестник Воронежского института высоких технологий. – 2014. – № 13. – С. 40-41.

[17] Данилова А.В. Об оценке характеристик беспроводной связи в помещении / А.В. Данилова, А.Г. Юрочкин // Вестник Воронежского института высоких технологий. – 2014. – № 13. – С. 113-115.

[18] Блохина Т.В. Исследование рассеяния электромагнитных волн на объекте при условии помех / Т.В.Блохина, Е. Ружицки // Вестник Воронежского института высоких технологий. – 2014. – № 12. – С. 47-50.

[19] Сапрыкин А.А. Характеристики высокочастотных mesh-сетей / Т.В. Блохина, Е. Ружицки // Вестник Воронежского института высоких технологий. – 2014. – № 12. – С. 116-118.

[20] Моргунов В.С. Современные методы расчета распространения радиосигналов в помещениях / В.С.Моргунов // Вестник Воронежского института высоких технологий. – 2014. – № 12. – С. 136-139.

[21] Комков Д.В. Характеристики радиопланирования при проектирования беспроводных систем связи / Д.В.Комков // Моделирование, оптимизация и информационные технологии. – 2013. – № 2. – С. 3.

[22] Мотин Д.Ю. О моделировании покрытия зоны обслуживания в беспроводной системе связи / Д.Ю. Мотин // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – С. 13.

[23] Баранов А.В. Методы анализа распространения и дифракции электромагнитных волн в беспроводных сетях / А.В. Баранов // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – С. 11.

[24] Болучевская О.А. Свойства методов оценки характеристик рассеяния электромагнитных волн / О.А. Болучевская, О.Н. Горбенко // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. – С. 4.

[25] Винюков М.С. О производительности компьютерной сети / М.С. Винюков, К.Ю. Гордиевская // Моделирование, оптимизация и информационные технологии. – 2013. – № 3. – С. 7.

[26] Жулябин Д.Ю. Модели каналов для беспроводных систем связи / Д.Ю. Жулябин // Моделирование, оптимизация и информационные технологии. – 2014. – № 1. – С. 1.

[27] Головин А.А. Предложения по разработке подсистем анализа и синтеза элементов систем радиосвязи / А.А. Головин, Я.А. Мишин, Э.С. Зацепин // Моделирование, оптимизация и информационные технологии. – 2014. – № 2. – С. 3.

[28] Головин А.А. Проблемы оценки потоков данных в компьютерной сети / А.А. Головин, Д.В. Завьялов // Моделирование, оптимизация и информационные технологии. –2014. – № 3. – С. 2.

[29] Горбенко О.Н. О моделировании сенсорных сетей / О.Н. Горбенко, А.А. Рожкова // Моделирование, оптимизация и информационные технологии. – 2014. – № 4. – С. 9.

[30] Горбенко О.Н. Свойства подходов, связанных с управлением электрическими сетями / О.Н. Горбенко, А.А. Рожкова // Моделирование, оптимизация и информационные технологии. – 2014. – № 4. – С. 20.

[31] Ермолова В.В. Архитектура системы обмена сообщений в немаршрутизируемой сети / В.В. Ермолова, Ю.П. Преображенский // Вестник Воронежского института высоких технологий. – 2010. – № 7. – С. 79-81.

[32] Преображенский А.П.О возможностях ускорения вычислений при решении задач / А.П. Преображенский // Вестник Воронежского института высоких технологий. – 2014. – № 12. – С. 67-68.

....

Prof. Igor Lvovich, D. Sc.
Paneuropean University, Bratislava, Slovakia
office@vivt.ru

Doc. Andrew Preobrazhensky
Voronezh Institute of High Technologies
app@vivt.ru

Cybernetic approach to the analysis of contemporary economic systems

Кибернетический подход к анализу современной экономической системы

A. A. Voronov
А. А. Воронов

Abstract:

The possibility of application of cybernetic approach related to the analysis of modern economic system is analyzed in the article. It is concluded that in the context of serious external influences on the economic system (political, economic impacts, and sanctions) when searching for the necessary managerial decisions it's convenient and effective to apply the black box method which allows to find the way out without studying the internal structure of the system, and analyzing only the input data and the resulting response.

Аннотация:

В статье анализируется возможность применения кибернетического подхода применительно к анализу современной экономической системы. Делается вывод, что в условиях серьезных внешних воздействий на экономическую систему (политические, экономические воздействия, санкции) при поиске необходимых управленческих решений удобным и эффективным будет применение метода «черного ящика», позволяющего находить выход не изучая внутреннее устройство системы, а анализируя только входные данные и возникающий отклик.

Key words:

Cybernetic approach, economic system, management, black box, safety.

Ключевые слова:

Кибернетический подход, экономическая система, управление, «черный ящик», безопасность.

ACM Computing Classification System:

Statistical timing analysis, Probability and statistics, Probabilistic reasoning algorithms, Information theory.

Введение

В последнее время внимание практиков и ученых-экономистов всего мира притягивает проблема развития современной экономики. Мы видим, как периодически проходят различные форумы, семинары, конференции, иные формы обмена опытом и попытки найти выходи из сложившихся проблем.

Хотим мы этого или не хотим, но следует признать, что мировая экономическая система находится сегодня в неустойчивом состоянии и неизвестно, что ее ожидает завтра. Может быть, впервые за последние 10 лет граждане российского и ведущих европейских государств стали ощущать на себе влияние нестабильности: снизилось качество жизни, упали доходы населения, существенно повысилась стоимость практически любых товаров. Общество, сегодня, как никогда стало зависимым от цен на энергоресурсы, на высокотехнологичные товары, технологии, продукты питания. С целью как-то отвлечь населения от внутренних проблем, администрации большинства государств проводят неясную политику по поиску какого-то внешнего агрессора, виновного во всех бедах, причем данная политика исходит не самостоятельно, а от «большого брата», находящегося далеко от нас и которому чужды любые успехи партнеров-конкурентов. К чему может привести подобная мировая практика до конца неясно, ясно одно, что не к лучшему и пока этого худшего не произошло необходимо выстроить четкий алгоритм нахождения необходимых решений.

Современное состояние экономической системы можно охарактеризовать как своеобразное воздействие на ее безопасность, вызывающее кризисные явления и неустойчивость ее функционирования. В этой связи налицо актуальность и важность изучения процессов управления в экономических системах, поиск точек минимизации состоянии неустойчивости и эффективных решений, направленных на повышение устойчивости анализируемой системы и ее стабильности.

1. Теоретический подход к исследованию экономической системы

Характеристика современного состояния экономики заключается в признании слабой предсказуемости производственно-экономических кризисов. Это свойство обнаруживается в невозможности точно спрогнозировать траекторию развития кризисного явления, ни при каком сколь угодно глубоком знании его морфологии, ни при каком сколь угодно длительном наблюдении за его развитием. Трудности научного прогнозирования возникают не потому, что не хватает логических, математических или каких-либо других методов, а из-за неопределенности относительно того, что следует прогнозировать, и саморегулирования кризисного процесса [1]. В свою очередь, построение алгоритма защиты системы влечет за собой необходимость правильного и эффективного управления ею. Как отмечают исследователи, основные функции системы управления безопасностью (и, соответственно, устойчивостью функционирования) должны состоять в оценке степени критичности ситуации, связанной с тем или иным видом нарушений безопасности, оценке уровня риска и поддержке принятия решений в зависимости от сложности ситуации. В процедуре принятия решений может возникать ряд трудностей, потому что зачастую невозможно формирование полного списка угроз безопасности, оценка критичности возникшей ситуации, прогнозирование развития, в случае негативного воздействия. Следовательно, одной из основных проблем является неполнота исходных данных о состоянии системы защиты в совокупности с множеством возможных угроз и дестабилизирующих факторов [2].

Стабильность в любой системе управления, как правило, связана не с самим объектом управления, а с его моделью, сформированной с учетом информационных потоков, характеристики управленческого персонала и окружающих воздействий. В этом смысле понимание хозяйствующего субъекта (предприятия, организации) как системы удобно характеризовать с позиций единства, целостности его элементов и, одновременно, их относительной самостоятельностью и способностью к саморазвитию. Системный подход дает возможность выявить все взаимосвязи элементов с целым, а целого – с его элементами. Кроме того, системный подход делает акцент на несводимость характеристик системы к сумме качеств ее элементов [3].

Очевидным является факт, что неустойчивость экономической системы требует немедленного вмешательства в систему контроля и управления. Можно предположить, что достаточно эффективным в данном случае может стать кибернетический метод, смысл которой можно свести к исследованию того общего, что есть в закономерностях, лежащих в основе процессов управления в различных средах, условиях, областях. Процессы управления, изучаемые в кибернетике, протекают в объектах, которые называются сложными динамическими системами. Управление всегда предполагает информационные процессы. Поэтому кибернетика есть вместе с тем наука об информации, об информационных системах и процессах. Здесь следует обратить внимание на то обстоятельство, что именно процесс получения, владения и обработки достоверной информации хозяйствующими и иными субъектами является определяющим в обеспечении их устойчивости.

Попробуем поразмышлять о теоретической применимости кибернетического метода применительно к экономической системе в общем и далее переложить полученное на современную реальность. Если рассматривать экономическую систему в целом, то следует отметить, что данная система не отличается простотой, поскольку функционирование как единичных элементов системы (например, хозяйствующих субъектов), так и самой экономической системы в целом возможно не только по линейному сценарию (развитие системы линейно зависит от развития ее элементов), но и характеризуется взаимной обратной связью (колебания общей системы вызывает отклик у ее элементов), что неизбежно требует эффективного регулирования и управления.

Кибернетический подход вносит в процесс управления экономическими системами различные понятия: регулирование, самоорганизация, цель, регулирование, эффективность, устойчивость и т. д.

Если на примере машинного производства рассматривать процесс создания какой либо машины, то следует отметить, что он характеризуется формирование более сложных элементов из более простых, постепенно, до тех пор, когда не будет достигнут конечный результат. В кибернетическом процессе сам процесс идет в обратном направлении: от комплексных целостностей и их функционирования к постепенному изучению составляющих элементов и их соединению, на которых основываются их функции. И только постепенно, кибернетический анализ позволяет прийти от целого к части, от системы к элементу. Для таких и подобных случаев так называемый метод черного ящика является как раз адекватным методом исследования, анализа, синтеза и научного предсказания [4, С. 36].

2. Метод «черного ящика»

Цель любого теоретического анализа экономической системы – попытаться найти причины связей между элементами и от анализируемых явлений и особенностей их функционирования идти к сущности. В этом плане мы должны отметить, что частичные результаты (пусть даже они могут не иметь большого внешнего значения) могут оказаться полезными в будущем и прежде всего тогда, когда задача состоит в том, чтобы исследовать определенные только в той мере, в какой это необходимо для практического использования каких либо экономических субъектов.

Метод «черного ящика» является одним из наиболее распространенных методов кибернетического анализа. Если рассматривать экономическую систему как кибернетическую (а это так и есть), то здесь выгодно использовать отношения «оригинал-модель», соответственно похожесть поведение позволяет моделировать развитие системы. Под понятием «черного ящика» уместно признавать систему, в которой внешнему наблюдению доступны лишь входные и выходные величины, а сам метод заключается в изучении свойств системы на основании знания и сопоставления ее входов и выходов при проведении экспериментального изучения, позволяющего в дальнейшем построить модель системы и предсказать ее поведение при любых заданных входах [5, С. 344].

С другой стороны, под «черным ящиком» понимается объект исследования, внутреннее устройство которого неизвестно. Понятие «черный ящик» предложено У.Р. Эшби. В кибернетике оно позволяет изучать поведение систем, то есть их реакций на разнообразные внешние воздействия и в тоже время абстрагироваться от их внутреннего устройства [6].

По существу, «черным ящиком» является любой объект, о котором можно судить лишь на основании изучения его внешних свойств. Метод «черного ящика» особенно важен для изучения поведения сложных систем, так как зачастую для них ответить на вопрос, как будет вести себя системы в тех или иных условиях, можно будет лишь на основании изучения ее поведения в каких-то других условиях, создание которых доступно при экспериментировании, или изучения характера поведения в прошлом. Приоритетное внимание, при этом, уделяется поиску оптимальных условий. Такая цель является одной из наиболее распространенных научно-технических задач. Подобные задачи возникают в тот момент, когда установлена возможность проведения процесса и необходимо найти наилучшие (оптимальные) условия его реализации. В широком смысле подобные задачи являются оптимизационными [7].

«Черный ящик» представляет собой сложную гомоморфную модель кибернетической системы, в которой соблюдается разнообразие. Он только тогда является удовлетворительной моделью системы, когда содержит такое количество информации, которое отражает разнообразие системы. Можно предположить, что чем большее число возмущений действует на входы модели системы, тем большее разнообразие должен иметь регулятор.

Всякая реальная система бесконечно сложна. Поэтому любое ее описание носит приближенный, а стало быть, модельный характер. Вид модели зависит от целей, для которых она создается. Существуют различные варианты модельного описания систем.

Современная экономическая система представляет собой нечто комплексное и цельное и выделенное из окружающей среды. Экономическая система и внешняя среда взаимодействуют между собой. В системологии используется представление о входах и выходах системы. В рассматриваемом случае вход системы – это воздействие на экономическую систему со стороны внешней среды, а выход – это воздействие, оказываемое системой на окружающую среду. В этой связи достаточно на примере модели «черного ящика», не изучая ее внутреннее функционирование описывать ее внешние взаимодействия.

Изменение внутренних свойств кибернетической системы может происходить за счет изменений либо в объекте управления, либо в каком–нибудь из элементов системы управления. Однако данная система есть нечто целое, действующее во внешней среде и обладающее определенными свойствами, и поэтому всякое изменение во внешней среде или изменение внутренних свойств влияет в той или иной степени на работу системы. Очевидно, что без среды нет системы и оценить воздействие этой среды на систему, выявить источник негативного воздействия и найти пути минимизации указанного воздействия и есть задача прикладной кибернетики.

3. Применение метода «черного ящика» для реальной экономической системы

Рассмотрим коротко упрощенное состояние современной экономической системы. Какие основные ее особенности (на выходе «черного ящика») можно отметить? Прежде всего, это относительная нестабильность, зависимость от ряда внешних факторов: политическая обстановка, цены на энергоносители, внутренняя политика в государстве, нормативно-правовая база и т. д. Возникает вопрос, а как откликается экономическая система на эти внешние факторы. Нам следует сделать исследовательский шаг, дополнительно определить организацию системы и выявить ее определяющий регулятор. Очевидно, что в нашем случае таким определяющим регулятором является само государство и та политика, которую оно проводит. Очевидным является то, что государство стремится сохранить существующие экономические отношения и обеспечить их устойчивость, а вместе с этим и сохранение условий их существования. Если же в ходе обострения противоречий между структурой экономической системы и ее функцией (производственные отношения) это единство будет нарушено, наступить временное состояние неустойчивости, в течение которого системы стремится к новому состоянию относительной устойчивости.

В результате принимаемых действий (если они достаточно верные) происходит реорганизация системы, которая влияет и на само государство [4, С. 67].

Нельзя в данном случае оставить без внимания и обратную связь в кибернетике, изученный известным ученым С. Биром, который можно охарактеризовать как механизм, позволяющий выходным показателям системы корректировать ее входные параметры. Одной из важных особенностей управления с обратной связью С. Бир называет невозможность выхода параметров за установленные границы, так как при отклонении от равновесного состояния (желаемых показателей системы) начинают действовать регуляторы, возвращающие показатели к норме. Таким образом, обратная связь является прекрасным аппаратом контроля и воздействия на систему.

Обратная связь постоянно воздействует на систему. Воздействие есть средство изменения существующего состояния системы путем возбуждения силы, позволяющей это сделать. Действие обратной связи может превзойти существующий вход в зависимости от места, времени, формы, интенсивности, содержания и длительности воздействия. Тот субъект, кто решает возникшую проблему, должен вмешиваться в существующее состояние, чтобы выполнить свою задачу. Воздействие может заставить систему пройти ее критическую точку и прекратить работу или работать быстрее. Для специалистов, решающих проблему, по определению, нет таких частей системы, которые были бы свободны от дефектов. Причина неправильного функционирования системы может быть заключена в любой подсистеме, поэтому проблема не может быть решена о тех пор, пока не установлено местонахождение причины неправильного функционирования [8, С. 108].

Другой важной особенностью можно считать тот факт, что регулятор с обратной связью может компенсировать влияние на систему возмущений, причины возникновения которых могут быть неизвестны. В этом случае он все равно выполняет функцию корректировки, заложенную в него [9].

Заключение

Попытаемся подытожить немного из сказанного на практическом примере. В настоящее время мы имеем в России достаточно нестабильную экономическую систему – это выход «черного ящика», своеобразный отклик экономической системы на воздействие – нестабильность, озабоченность общества, снижение его жизненного уровня, социально-экономические проблемы в стране и т. д.

В качестве входа в «черный ящик» можно указать приблизительные или некоторые причины этой негативности – низкие цены на энергоносители, природные нефте и газоресурсы, сложная политическая обстановка, внешние санкции и т. д. – это вход «черного ящика». Наша задача в целом - не выявлять причинно-следственную связь между входом и выходом (она здесь подразумеваетсяч интуитивно), а анализируя внешние и внутренние характеристики (вход и выход) попытаться найти управленческий выход из сложившейся ситуации.

Что же делается сегодня для исправления кризисной ситуации? Первоначальной мерой в указанном направлении стал принятый Правительством РФ так называемый план первоочередных мероприятий по обеспечению устойчивого развития экономики и социальной стабильности в 2015 году (утв. Распоряжением от 27.01.2015 г. № 98-р), согласно которому в качестве стабилизационных мер выступает достижение сбалансированности рынка труда, достижение положительных темпов роста и макроэкономической стабильности в среднесрочной перспективе, содействие развитию малого и среднего предпринимательства за счёт снижения финансовых и административных издержек, поддержка импортозамещения и экспорта по широкой номенклатуре несырьевых, в том числе высокотехнологичных, товаров, создание возможностей для привлечения оборотных и инвестиционных ресурсов с приемлемой стоимостью в наиболее значимых секторах экономики, а также снижение напряженности на рынке труда и поддержку эффективной занятости. Казалось бы, эти действия (план) и есть формальный результат как бы кибернетического анализа системы на примере «черного ящика». Тогда данный план должен являться искомым решением проблемы, однако, можно предположить, что из-за недостаточного анализа всех возмущающих воздействий и откликов, а также информационных потоком данное решение может быть не вполне комплексным, поскольку не до конца ясно, а за счет чего будет достигаться данный результат по минимизации кризисных явлений? Наше мнение сводится к тому, что составляющие элементы экономической системы (хозяйствующие субъекты) будут вынуждены (как это бывало и раньше) варьировать своим управленческим аппаратом и выработкой оптимальных решений в процессе своего выживания, а общая система может им только обеспечить поддержку (ну или хотя бы не мешать).

Следует отметить наличие разницы между теоретическими рассуждениями и реальной действительностью, поскольку при разработке конкретной модели сложно заложить все критерии и компоненты сложившейся ситуации. Поэтому, в любом случае, предлагаемые, на основе применения кибернетического метода решения, в целом будут приблизительными. В этой связи целесообразным будет являться комплексное применение теоретического анализа и практических, эмпирических данных. И очевидным является то, что сегодня, для того, чтобы обеспечить эффективность и стабильность экономической сферы необходимо не только проводить глубокий теоретический анализ, но и выискивать эффективных производственников, менеджеров, иных специалистов, создавать необходимые условия для производственной активности. Только тогда будет достигнут необходимый результат.

Литература

[1] Новосельцев В.И. Методологические аспекты изучения кризисов в социальных и экономических системах // Моделирование, оптимизация и информационные технологии. Электронный научный журнал. – 2013. – № 2 / http://moit.vivt.ru.

[2] Львович Я.Е. Принятие решений в условиях дестабилизации системы /Львович Я.Е., Сахаров Ю.С., Яковлев Д.С // Вестник Воронежского института высоких технологий. – 2013. – №11. – С. 114-116.

[3] Воронов А.А. Использование элементов системного и кибернетического подходов в сфере управления хозяйствующими субъектами // Современные проблемы экономики и менеджмента: Материалы Международной научно-практической конференции. – Воронеж, ВГУ, 2014. – С. 18-21.

[4] Клаус Г. Кибернетика и общество: Монография. – Москва: Издательство «Прогресс», 1966. – 432 с.

[5] Основы кибернетики. Теория кибернетических систем / Под ред. К.А. Пупкова. Учеб. пособие для вузов. – М., «Высшая школа», 1976. – 408 с.

[6] Тестирование по стратегии черного ящика / http://www.sbup.com/wiki/Тестирование_по_стратегии_чёрного_ящика

[7] Кривошеев И. А. «Черный ящик», как основа решателя задач оптимизации параметров ГТД / И.А. Кривошеев, Ю.А. Хохлова, Р.А. Завьялов // Молодой ученый. – 2011. – №10. – Т. 1. – С. 76-81.

[8] Обратная связь и «черный» ящик в кибернетике / http://cyber.econ.spbu.ru/zhukova/Zhukova.htm

[9] Оптнер С.Л. Системный анализ для решения проблем бизнеса и промышленности / Станнфорд Л. Оптнер; Пер. с англ., вступ. ст. С.П. Никанорова. – 2-е изд. – М.: Концепт, 2003. – 206 с.

D.J.S., Cand. Sc. (Physics and Mathematics) Alexander Voronov

Voronezh Institute of High Technologies

a.a.voronov@mail.ru

Instructions for Authors

…………………………………………………………………………………………………..

International scientific journal Information Technology Applications jointly issued by Faculty of Informatics of Paneuropean University and Civil association EDUCATION-SCIENCE-RESEARCH in Bratislava, offers space to publish:

Scientific articles	in the range of 20 standard pages (there is possible to place up at the most 1800 characters including character spacing on the one page of A4 format). Reports in the range of 5 standard pages
Discussions	in the range of 2 standard pages
Information	in the range of 1 standard page

The journal presents practical and theoretical knowledge about the use of information technology mainly in the field of economy, business, law, media, psychology, education, power engineering and public administration and next. It is written in Slovak, English, Russian and Czech language. The journal is published biannualy. Contributions will be accepted only in electronic form in doc or docx format on vvv.esr@gmail.com in the form of **author's surname.doc (docx)**. Main requirement of accepting the contribution is its originality. Another Condition for publishing the contribution is the positive attitude of editorial board and two independent reviewers.

The contribution must be written in MS WORD, Times New Roman font, single spacing of the lines, A4 page format, 2.5 cm margins, not to number the pages according to the following structure:

Required part

1. Tittle of the contribution in English language: font size 16, bold, center alignment. *Omit line*

2. Name and surname of the author (or authors separated by hyphen): font size 14, italic, center alignment. *Omit line*

3. Abstract: font size 12, bold, left alignment. The text of the Abstract written in English on a new line, range of 250-300 words, font size 11, justified alignment: *scientific goal/methods, conclusions* according to http://info.emeraldinsight.com/authors/guides/abstract.html.

4. Key words: font size 12, bold, left alignment. Text written in English on a new line, font size 11, justified alignment, range of 3-5 key words (separated by comma). *Omit line*

5. ACM Computing Classification System: font size 12, bold, left alignment. To adduce the classification codes (font size 11, separated by comma) on the same line according to http://www.acm.org/about/class/2012. *Omit line*

6. Dividing the contribution to a clearly defined parts (Introduction, Conclusion) and to numbered chapters (1, 2, ...) and subchapters (1.1, 1.1.1, 1.1.2, ..., 2.1, 2.1.1, 2.1.2, 2.2, ...): font size 12, bold, justified alignment. Introduction, Conclusion – bold, not to number; Chapters, Subchapters – to number, bold

Tables, graphs a pictures to put straight into the text and to mark by sequential number and description (font size 11, italic, left alignment) - Examples:

Table 1: title.
Graph 1: title.
Picture 1: title.

Below it to mention the source – *Source: source name* (font size 11, italic, left alignment)

Bibliographic references adduced in the text according to STN ISO 690 standard and international standars in a form (name of the author, year of issue)

Citations appear in the text, in direct citation is necessary to add the page number.

7. Literature: font size 12, bold, left alignment. Literature List alphabetized on new lines, font size 12, justified alignment, with all identification data according STN ISO 690 standard if it concerns *book, chapter of book, contribution from almanac, arrticle from journal, internet documents* (http://owl.english.purdue.edu/owl/resource/560/01).

8. Author’s address: font size 12, bold, left alignment, address placed on a new line, font size 12, left alignment seriately according to *Name and surname of the author, degree, address of the institute , e-mail*.

Optional part

9. Tittle of the contribution in other language: font size 16, bold, center alignment. *Omit line*

10. Abstract: according to the point 3 (but written in other language).

11. Key words: according to the point 4 (but written in other language).

VORONEZH INSTITUTE OF HIGH TECHNOLOGIES (THE RUSSIAN FEDERATION)

ВОРОНЕЖСКИЙ ИНСТИТУТ ВЫСОКИХ ТЕХНОЛОГИЙ

Voronezh Institute of High Technologies is a multi-sectoral and multilevel educational institution providing basic and professional training of highly professional personnel. VIHT Diploma means an up-to-date educational level of a leading higher school in Black Earth Region of the Russian Federation satisfying the world quality standards in information technologies.
Voronezh Institute of High Technologies is certified according to management quality system of an international standard ISO 9001-2008.

The quality of international programmes and the staff professional competence let the institute to be internationally certified, that gives the graduates an opportunity to get diplomas of the internationally-recognized format - the certificates of international organizations: IES, ECDL, MICROSOFT, CISCO, 1C, PRINCE2. The institute of high tech nologies is a partner of the National Centre of managers' certification. The graduates obtain a state-approved diploma.

HIGHER EDUCATION

BACHELOR DEGREE

- **Workforce Management**
  - Personnel planning and marketing
- **Management**
  - Tourism Management
  - Tourist Services Management
  - Small Businesses Management
- **Informatics and Computing Engineering**
- **Information Systems and Technologies**

SPECIALIST'S DEGREE (FIVE-YEAR EDUCATION)

- **Fire Safety**

MASTER DEGREE

- **Management**
  - √ Business Development Management
- **Informatics and Computing Engineering**
  - √ Information and Software of Automated Systems
  - √ Computer Network and Telecommmunications

POST-GRADUATE DEGREE

- **Informatics and Computing Engineering**
- **Education and Pedagogical Sciences**

TWO DIPLOMAS PROGRAMME

Pan-European University is one of the most prestigious universities in Europe. It offers an opportunity to get European higher education in Voronezh Institute of high technologies, as well as the chance to acquire a unique learning and work experience in an international environment which is highly essential for successful people.

VOCATIONAL TRAINING

- **Programming in Computing Systems**
- **Economics and Accounting** *(branch-wise)*
  - √ Social-and-Cultural Training
  - √ Bank Training
  - √ Manufacturing

CHILDREN'S COMPUTER SCHOOL

- √ Computer Literate
- √ Computer Graphics, Design, Animation
- √ Hardware and Computer Software (CISCO IT Essentials)
  Computer Hardware and Software (CISCO IT Essentials)
- √ Programming

SUPPLEMENTARY EDUCATION

- √ Management Business School
- √ Specialists Training in the field of Telecommunications
- √ CISCO Networking Academy
- √ Training and Certification on the leading international educational programs of IT-education
- √ School of Computer Design

VOCATIONAL RETRAINING AND REFRESHER COURSES

- √ Training programs for educators, included in the regional bank of training programs
- √ Fire-Technical Basics
- √ Traffic Security
- √ Corporate training programs, seminars, trainings
- √ e-commerce

394043, Russia, Voronezh, Lenina Street, 73a
Tel: +7-800-55-56-054
ihfo@vivt.ru
www.vivt.ru

Available on the App Store

Google play

License 90Л01 0001022 от 05.03.2014 (№ 0954),
Gosudarsvennoy accreditation certificate № 90А01 от 09.08.2012 (№ 0179)