DOI: https://doi.org/10.15407/kvt192.03.027

UDC 004.942: 614.7

M.Yu. ANTOMONOV, DSc (Biology), Professor,
Chief Researcher of the Laboratory of Epidemiological
Research and Medical Informatics
e-mail: antomonov@gmail.com
State Institution "O.M. Marzieiev Institute for Public Health
of the National Academy of Medical Sciences of Ukraine"
50, Popudrenko str. Kyiv, 02660

INFORMATION TECHNOLOGY FOR CONSTRUCTING THE COMPOSITE INDICES FOR DATA OF DIFFERENT TYPES USED IN MEDICAL AND ENVIRONMENTAL STUDIES

Introduction. *Information technologies used in medical and environmental researches often deal wiht huge amounts of information processing. These technologies allow us to identify and investigate previously hidden dependencies and interactions in complex environmental, medical and biological systems, and on the other hand, it is accompanied by the analysis of large data sets, some of which (sometimes most of them) have an uninformative (noisy) character. One of the ways of solving this problem are the methods of constructing composite indices (CI), i.e. complex indicators, which allow to perform an integral assessment of the state and functioning of ecological, medical and biological systems.*

The purpose of the paper *is to develop a generalized information technology for constructing composite indices for different types of data used in medical and environmental studies.*

Results. *Medical and ecological researches include two main components: analysis of the states of both human health and of the environment; in solving such problems it is necessary to evaluate and analyze the state of the bioobject according to the data of different types: quantitative, rank, binary and qualitative variables. The developed general information technology is oriented on supporting the solution of a wide range of medical and hygienic tasks and integrates various approaches to processing and analysing of data of different types. Proposed technology consists of four stages: the formation and initial analysis of an initial indicators set, the calculation and normalization for obtainig unnamed equivalents, the actual design of the composite indices, and their verification. The implementation of this technology makes it possible to compare data of different dimensions, determine the significance of specific characteristics in a general research totality, to evaluate the integral state and to classify the research objects.*

Conclusion. *The proposed information technology for the construction of composite indices based on data of different types: quantitative, rank, binary and qualitative variables, is an effective tool for determining and comparing the state of bioobjects of different nature, and its use makes it possible to avoid mistakes in the incorrect application of mathematical methods for processing medical and ecological information.*

Keywords: *information technology, composite indicators, processing medical and ecological quantitative, rank, binary and qualitative variables.*

INTRODUCTION

At present, information technologies used in medical and environmental researches often deal with huge amounts of information processing. The reasons are obvious: an increase in the complexity of research tasks, the expansion of the capabilities of recording technology, the possibilities of Internet technologies for the transmitting digital information, and the program and technical capabilities of computers used for data processing and analysis. On the one hand, this allows us to identify and investigate previously hidden dependencies and interactions in complex environmental, medical and biological systems, and on the other hand, this is accompanied by the analysis of large data sets, some of which (sometimes the majority) have an uninformative (noisy) character . These problems cause development of such methods of the initial convolution, compression of the initial numerical arrays, in the application of which the loss of information would be minimal [1].

In mathematical statistics, traditional methods of factor, discriminant, cluster and regression analysis have been developed years ago and successfully used for these tasks. Modern software products make it easy enough to apply the relatively newer methods of Data Mining: multidimensional scaling, multifactorial dimensional reduction (MDR), decision trees, graphical methods for representing and classifying multidimensional data [2].

One of the most popular and rather effective ways of solving this problem are the methods of constructing composite indices (CI) that is complex indicators, which allow to astimate integrally the state and functioning of ecological, medical and biological systems [3].

Such evaluation has certain advantages: the complexity of the obtained information, the ease of use. It can serve as a tool for accounting, analysis and planning, an indicator of the state and the criterion of comparative evaluation, an indicator of the effectiveness of decisions taken and the completeness of their implementation, it can also serve as the basis for selecting the possible measures and indicators of expected results in the future.

PROBLEM STATEMENT

Medical and ecological researches include two main components: analysis of the state of both human health and environment.

A comprehensive assessment of human health is done at three levels: individual, group and population.

At the individual level, the following calculated indicators are used: those characterizing the functioning of individual physiological systems of the body (cardiovascular, respiratory, etc.), generalizing indices characterizing the coordination of two or more of these functional systems, and integral indices describing the complex interrelationships of organs and systems (an index of somatic health, adaptation potential, etc.).

The appropriate formulas are used to calculate these indices, in which the variables are a limited list of anthropometric and functional indices: height, weight, chest circumference, respiratory hold-up time, heart rate (before and after the activity), arterial and systolic pressure, recovery time, dynamometry, etc. Often, the same indicators are included in the formulas of different indices, which causes their interdependence. The range of changes in the indices is most

varied, and criterial scales of health assessment respectively. Their orientation is also uncertain: the increasing of some indices indicates an improvement in the state of the body, others — on the contrary, a deterioration. Sometimes a certain range of "norm" is postulated, and the deviation in any direction is considered as a negative phenomenon [4–8].

Group comprehensive health assessment is meant to evaluate and monitor the status of identified groups of individuals, in particular children's groups. It is based on the definition of the health group of each individual using at least four criteria: the presence or absence of chronic diseases at the time of examination; the level of the functional state of the body basic systems; the degree of the body resistance to unfavourable effects; the level of development achieved and the degree of its harmony. Group complex health indicators are descriptive expert characteristics [9].

For the comprehensive assessment of population health, morbidity and prevalence are usually used, on the basis of which standardized coefficients, primary and general morbidity rates, indices and other characteristics are calculated. Such secondary indicators are the result of a simple ratio of the values of some indicators to the values of others, these indicators are analyzed using graphical and descriptive methods [10]. Among the methods that aggregate the initial population indicators, the method of "per-centil-profile", the method of the sum of places, methods using age-specific disease rates can be identified [11–13]. All these comprehensive population indicators in the mathematical aspect are simply calculated ratios or additive scores of points. When studying the level of regional demographic development approximation of the preference function by a linear regression model is proposed for constructing a composite index [14], and also composite indicators are used to analyze the dynamics of the health status of the population using mathematical models [15].

Environmental indicators often describe the state of the water and air environment. As an integrated index of drinking water quality, the most obvious formula is the sum of the concentrations of all contaminants (x_i) normalized to their "safe" (x_0) value (to the maximum permissible concentration — MPC) [16]. There are a number of similar indicators (for example, the total index of chemical contamination [17], the combinatorial pollution index [18, 19]), which combine a given number of primary characteristics.

For a comprehensive assessment of air pollution, basically, the following characteristics are used:

$$\mathrm{CIAP} = \sum_{i=1}^{n} \left(\frac{x_i}{x_0}\right)^{k_i}, \quad \mathrm{TAPI} = \sum_{i=1}^{n} \frac{x_i}{x_0 k_i}, \quad \mathrm{P} = \sqrt{\sum_{i=1}^{n} \left(\frac{\mathrm{x_i}}{\mathrm{x_0} k_i}\right)^2},$$

where CIAP — the complex index of atmospheric pollution [20]; TAPI — total air pollution index [21]; P — Pinigin` indicator of atmospheric pollution [22]. The dimensionless constant, which depends on the hazard class of the *i*th substance (k_i), assumes values of 1.5; 1.3; 1.0; 0,85 at the calculation of IAP, values of 0,8; 0.9; 1.0; 1.1 at the calculation of TAPI and values of 2,0; 1,5; 1,0; 0,8 at the calculation of P respectively, for substances of 1-, 2-, 3-, and 4th hazard classes.

Thus, the analysis showed the existence of many proposed complex indicators used in medical and environmental studies; methods and formulas for their calculation are very diverse. However, their mathematical contents are reduced mainly to the standardization of characteristics (mostly by dividing by some

"norm") and subsequent summarizing (sometimes with calculated or predetermined coefficients). The ranges of their variation are specific for each indicator. Also, their breakdown by gradation (criteria) and their subsequent verbal quality assessment are specific and different. Each of the complex indicators is intended for use in its narrowly restricted field of research.

All these shortcomings make it possible to talk about the expediency and topicality of developing an unified information technology (IT) for designing composite indicators for medical and ecological researches that would be applicable to combining private characteristics of various types of data of the health status and quality of the environment, would have a standard and adequate mathematical completeness and would be easy to use [23].

The purpose of the paper is to develop a generalized information technology for constructing composite indices for different types of data used in medical and environmental studies.

INFORMATION TECHNOLOGY OF CONSTRUCTION OF COMPOSITE INDICES FOR QUANTITATIVE VARIABLES

In mathematical statistics, the original data are classified as belonging to one of four types of scales: quantitative, rank, binary and labels. Information technology for the design of CI for each of these types of data has both common and distinctive stages. Common are the initial and final stages, the specific — the actual processing of data and technology for the formation of CI.

Taking into account the variety of research tasks, normative or methodical regulations and expert opinion, the proposed IT design of CI on quantitative indicators unites various effective and acceptable approaches at all its stages.

The first stage is the formation and initial analysis of the set of initial indicators.

When *forming the initial list of indicators* it is necessary to be guided by the following principles: *informativeness* — indicators should characterize the most significant properties of the object under study; *completeness* of the description — the totality of the recorded characteristics must be exhaustively and comprehensively described; *uniquality* — each indicator characterizes only one characteristic; *measurement capabilities* — indicators can be recorded; *representativeness* — they must reflect the immanent qualities of the object; *non redundancy* — the characteristics should not be interrelated. An important feature of the chosen variables is their acceptability — the necessity of matching variables to quality in which the CI is non-contradictory.

The evaluation of the informative value of quantitative indicators is determined by the research task and, accordingly, the statistical methods used in this process. Thus, in the analysis of variance, the most informative variables are those that have the greatest variance, when comparing samples — those for which the criteria of difference are the greatest, in the correlation analysis those having the highest correlation coefficients with the resulting index, in discriminant analysis those the most reliably entering the classification function, in descriptive statistics those having the greatest (least) variability of the index, etc.

Decrease in the number of indicators. In the process of collecting initial data, the problem of their redundancy often arises. The solution of this problem consists in selecting only significant features by some criteria, for example, by the threshold value of information content.

The second stage is the calculation and normalization of unnamed equivalents.

At this stage, the choice of the *"basis"* of the indicators is performed at first. As a rule, quantitative variables have different units of measurement and for their transfer to a dimensionless scale it is necessary to standardize them relatively to a certain "basis". As a "basis," it is advisable to use the parameters of the original data array, such as the arithmetic mean, the smallest or largest values of the sample, if they correspond to the notion of the "ideal" of the characteristic. "Basis" can be configured as the upper or lower limit of the confidence intervals of the means, i.e. as a limiting value of the average totality. As the "basis" you may take the values of the characteristics in the control group (if the study consists of comparing the experimental and control groups) or in the same basic group under more favorable conditions (for example, the values of the indicators in youth when analyzed in old age). "Basis" may be physiological norms or MPCs, or can be given by the objectives of the study (for example, as a desired idealized result).

The procedure for choosing a "basis" should be carried out not formally, but using the whole set of priori knowledge. Within the framework of one research, it is necessary to use a unified approach to its basis selection.

The next step is *to calculate the dimensionless equivalents* of the original quantitative variables, i.e. there is a transition to some uniform description for all characteristics, regardless of the units in which they are measured. To obtain dimensionless equivalents, it is necessary to use values of the same dimensionality (kg; cm; mm Hg, etc.).

The simplest way to obtain a dimensionless quantity (d_1) is to calculate the ratio of the value of the initial indicator (x) to its "basis" (x_0). For this purpose, it is possible to calculate the relative deviation: $d_2 = (x - x_0) / x_0$, which can be expressed as percentage. It is also simple to use the matching procedure with the sweep of the sample:

$$d_3 = (x - x^-) / (x^+ - x^-) .$$

As limiting values (x^- and x^+), the real minimum and maximum values in the sample, or the lower and upper limits of the confidence interval of the arithmetic mean may be used.

When calculating dimensionless equivalents useful method is to compare initial values with variability indices, in this case the significance of indicators is indirectly taken into account (if assuming that the variability is greater, the less is the significance of the indicator). In particular, when using the arithmetic mean error S as variability indices, we have $d_4 = x / S$. If the standard deviation as the variability characteristic is used, then obtain $d_5 = x / \sigma$. The standardized deviations (d_6) or the Student's coefficients ($d_7 = t$) may be used:

$$d_6 = \frac{x - \bar{x}}{\sigma}, \qquad d_7 = t = \frac{x - \bar{x}}{S}.$$

In the last step of this stage, *the normalization of dimensionless equivalents* is performed. Best of all the normalization is to be performed so as new variables have firm limits of the change, for example, changed between 0 and 1. The transition from the initial data or their equivalents to the

normalized variables can be described by different functions. Their formation or choice is determined solely by the research objectives and the art of the data processor.

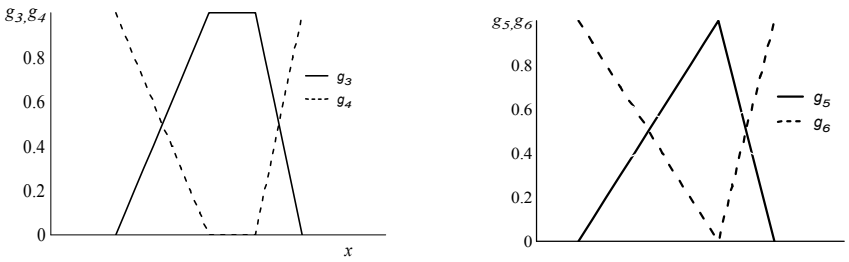
By using different transition functions, you can focus either on variables close to the average value, or on the maximum / minimum values. At the same time, the greatest "weight" is assigned to either "good" for the biosystem to indicators close to the "ideal", or, on the contrary, the most "pathological". In fact, the normalization operation is equivalent to an expert evaluation of the significance of the data, with the difference that these "weights" are not equal to discrete values, but are described by a continuous function.

There should be complete clarity about what is "good" and what is "bad" concerning the state of the object. If the value of "1" is chosen as a favorable state of the bioobject, then for each particular registered index of the state of this object the normalization function should be chosen so that the best values for the biosystem are close to 1, and the worst values are close to 0.

The normalizing linear functions throughout the sample over (the entire span of the sample) are the simplest. They are: an increasing linear function (g_1), which is equal to 1 for $x = x^+$ and 0 for $x = x^-$, and a decreasing linear function (g_2) equal to one for $x = x^-$ and 0 for $x = x^+$:

$$g_1 = d_3 = (x - x^-) / (x^+ - x^-), \qquad g_2 = (x^+ - x) / (x^+ - x^-).$$

If the "basis" of the indicator is not equal to a strictly fixed value and the values in a certain range (from x_0^- up to x_0^+) can be considered normal, then it makes sense to assign the "best" value of the normalized variable to this entire range, for example, $g = 1$.



a) b)

Fig. 1. Normalization by piecewise-linear functions "truncated pyramid" and "inverse truncated pyramid" (a) and "pyramid" and "reverse pyramid" (b)

In this case, the normalization function (g_3) will be written in the form of a "truncated pyramid" (Fig. 1a):

$$g_3 = \begin{cases} \dfrac{x - x^-}{x_0^- - x^-}, & \text{if} \quad x^- \le x \le x_0^-, \\ 1, \quad \text{if} \quad x_0^- < x < x_0^+, \\ \dfrac{x^+ - x}{x^+ - x_0^+}, & \text{if} \quad x_0^+ \le x \le x^+. \end{cases}$$

If the range of the "basis" is considered to be the worst for the biosystem, the normalization function g_4 will be a mirror-symmetric of the function g_3 and will be written as an "inverted truncated pyramid":

$$g_4 = \begin{cases} \dfrac{x_0^- - x}{x_0^- - x^-}, & \text{if} \quad x^- \le x \le x_0^-, \\ 0, \quad \text{if} \quad x_0^- < x < x_0^+, \\ \dfrac{x - x_0^+}{x^+ - x_0^+}, & \text{if} \quad x_0^+ \le x \le x^+. \end{cases}$$

If the "basis" does not have a range of variation and is strictly fixed, then $x_0^- = x_0^+ = x_0$ the functions g_3 and g_4 degenerate in the g_5 ("pyramid") and g_6 ("reverse pyramid") functions, consisting of two segments (Fig. 1b).

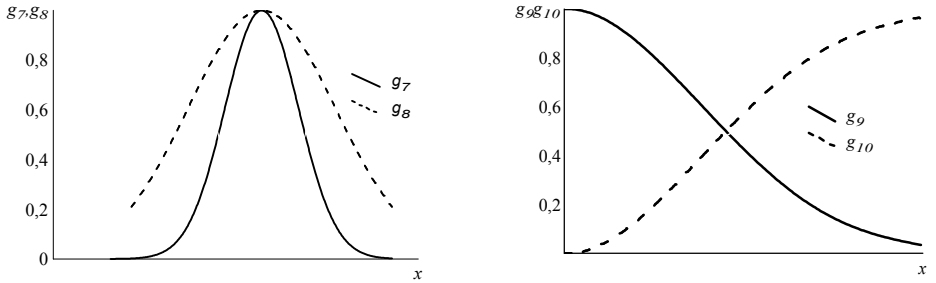
$$g_5 = \begin{cases} \dfrac{x - x^-}{x_0 - x^-}, & \text{if} \quad x^- \le x \le x_0, \\ \dfrac{x^+ - x}{x^+ - x_0}, & \text{if} \quad x_0 < x \le x^+. \end{cases}$$

$$g_6 = \begin{cases} \dfrac{x_0 - x}{x_0 - x^-}, & \text{if} \quad x^- \le x \le x_0, \\ \dfrac{x - x_0}{x^+ - x_0}, & \text{if} \quad x_0 < x \le x^+. \end{cases}$$

However, it is more reasonable to use nonlinear transformations to normalize physiological parameters. Nonlinear functions do not have fracture points; they can be formed in such a way that they do not reach the limit (anomalous) values, but only aspire to them asymptotically; their description is more compact and aesthetic. It is best to use exponential functions for nonlinear normalization.

For example, if the maximum "weight" (equal to 1) is given to the value of "optimum", and the extreme values correspond to unsatisfactory estimates of the biosystem state, and the further from the "optimum", the worse, then as the normalizing one can use the symmetric unimodal function "bell" (Fig. 2a):

$$g_7 = \exp\left(-a\,(x - x_0)^2 \right).$$



a) b)

Fig. 2. Normalization with the help of the unimodal "bell" (a) and "semi bell" function with asymptotic approximation for the maximum values (b)

The width of the "bell" is determined by the value of the parameter a: in Fig. 2a, the parameter a in the function g_7 is half the width of the function g_8. The value of the parameter a can be specified through the variability of the original array, then the "bell" covers a certain number of standard deviations (σ).

For example, for $a = 1/(2\sigma^2)$ the function g_7 can be written in the form of a Gaussian curve:

$$g_7 = \exp\left(-(x_i - x_0)^2 / 2\sigma^2 \right).$$

If x^- is the best value for the biosystem, then the value of the normalizing equivalent is 1, and its values with increasing x asymptotically tend to 0 ("the more, the worse"), the normalization functions are logically chosen as a "semi bell" (Fig. 2 b):

$$g_9 = \exp\left(-a(x - x^-)^2 \right).$$

Examples of indicators for which this normalization can be used are the availability of bilirubin and creatinine in the blood, the availability of protein, sugar, leukocytes and erythrocytes in the urine, etc.

If, conversely, the minimum value of the original array corresponds to the exact minimum value of the equivalent ($g(x = 0) = 0$), and the maximum value of x is assigned to equivalent value close to 1 (the larger, the better), the normalization function can be written in the form of an increasing S-shaped function (Fig 2 b):

$$g_{10} = 1 - \exp\left(-a(x - x^-)^2 \right)$$

This normalizing function is acceptable for a lot of parameters: blood oxygen tension, visual keenness, intellectual development, minute breathing volume, vital capacity of the lungs.

The third stage is the actual construction of the composite indicators. When combining the characteristics in the CI, it is necessary to take into account their medical and biological meaning and clearly understand the goals and objectives of the study. It is recommended to combine variables related to the same body system (i.e., cardiovascular, respiratory, etc.) or taking into account the specificity of the study (in psychophysiological or psychological testing).

The generalization of the standardized equivalents can be carried out:

- additively as the arithmetic mean or as a weighted average with "weights" w_i (G_1);

- multiplicatively as a simple product or a product with corresponding power "weights" that characterize their relative importance (the so-called "production functions"); using the geometric mean, or the formula for calculating the probability of independent events:

$$G_1 = \left(\sum_{i=1}^{n} w_i g_i \right) / \left(\sum_{i=1}^{n} w_i \right), \; G_2 = 1 - \prod_{i=1}^{n} (1 - g_i) .$$

For multiplicative convolutions, it is necessary to ensure that the products do not have 0 or close to 0 values. If these values cannot be eliminated at the normalization stage, then it is necessary to apply preliminary averaging in these formulas.

To construct composite indices, *the use of regression models* is possible. In this case, the direct use of the initial variables for CI not standardized (dimensionless) equivalents is possible, which greatly simplifies CI construction and practical use. With this approach, the choice of informative variables takes place automatically (assuming that the informativeness in this case is determined by the reliability of the coefficients), the significance and action direction of the initial characteristics action is determined by the β-coefficients of the model, the adequacy of the CI is easily calculated (i.e., by the Fisher criterion).

The only fundamental problem is the lack of the function empirical values, which are required in regression analysis for constructing models. However, it can be solved if the values of the CI calculated by one of the above methods or expert estimates are used as the output function. At the same time, it is natural that the list of variables included in the regression model should differ from the set of characteristics by which the external function (CI) was calculated. If these values are enclosed in the interval [0,1], then the range of the regression functions will also be between "0" and "1".

After calculating the CI for any of the variants, it is possible to successively combine the CI in a community of a higher level, taking into account the hierarchy of the organization of the biosystems. For example, it is possible to obtain separate CIs for the state of various physiological systems (blood, cardiovascular, respiration, etc.) for specific characteristics, and then to combine them into a general "super comprehensive" index of the physical state. Further, these assessments, for example, physical, mental and social status, can be combined into an even more general indicator of the individual` health state.

With complex indicators, you can carry out any mathematical processing — statistical analysis, comparison, the identification of dependence on time and other factors, etc.

At **the final, fourth stage**, verification of the CI takes place, i.e. verification of the correctness and adequacy of its obtaining. This check can be carried out by experts using the selected assessment procedures. For a qualitative evaluation of the result, the expert's opinion can be expressed in the categories "capable", "corresponding", "effective" or vice versa. It is advisable to choose "contrasting" examples for testing. In this case, the worst object, in the expert opinion, should be expressed by the minimum value of the CI, the best object should be expressed by the maximum value. If some object is "ideal" by the expert`s opinion, then after evaluation it should receive the greatest value on condition of CI correct construction. Naturally, the importance of the expert evaluation increases if there are several experts and/or they compare several CIs obtained by different methods or with some other CI, already recognized.

Mathematical verification of CI can be done by various methods, for example, by calculating the correspondence of the CI to the entire spectrum of variables that are included in its construction. To do this, you can use pair and multiple correlation analysis, differentiation criteria, etc.

Naturally, if the result of verification is unsatisfactory, it requires the recalculation of the entire design of the CI.

The effectiveness of CI use increases significantly if it is given the practical meaning, understandable to any user. For example, the range of CI possible changes can be divided into several gradations, each of them will have a verbal (semantic) assessment. The number of gradations can be determined arbitrarily. The division into two gradations is acceptable for strict selection (as "suitable"/"unsuitable"). Three or four grades correspond to the traditional division into "bad", "satisfactory", "good", "excellent". Gradation limits are either simply established, based on the convenience of classification, or are calculated mathematically. In this case, there can be a simple partitioning into equal intervals (i.e., the range from 0 to 1 as 0-0.33, 0.34-0.66, 0.67-1). Either there is partitioning by statistical methods based on empirical data: using sigma deviations (when there are three, four, five ranges) or quartiles (when there are four ranges, each containing an equal area of the CI normal distribution).

As a result, the researcher gets a tool with which he can evaluate the significance of a set of informative characteristics expressed by one number. This is the verbal integral evaluation (IE).

Upon completion of the work on the CI and upon obtaining the IO, their visualization and implementation of these results are necessary. Such block is similar to the final stage in the performance of any scientific work.

COMPOSITE INDICES FOR RANK AND BINARY VARIABLES

For rank variables a technology developed for quantitative variables can be implemented with the exception of options for calculating group parameters and normal distribution parameters. More over, instead of the arithmetic mean for rank variables, it is customary to use the median.

When forming the composite indicators for rank variables, it is easiest to use the method of direct points evaluation. In this case, all significant characteristics of the object are normalized in the same interval, as a rule, between "0" and "1". At the same time, the highest value is assumed maximum "good", for example "1", and "0" is regarded "bad". Next, we compare all the attributes by their significance for the object` integral evaluation and introduce points (measures, multipliers) of this significance (weights) for each of the attributes.

The total complex score of the whole object is obtained by adding up the points relating to the particular characteristics. In this case, as in the case of quantitative variables, it is possible to normalize the final CI from "0" to "1" dividing it into the maximum possible value of the sum of points.

The option of **binary variables** is the simplest to calculate: the informative indicators are highlighted, the criterion of assigning the indicator to the "necessary quality" is formulated, as a result of which a list of "unidirectional" binary features is compiled, and weights are assighed to these characteristics. If the research task is to classify an object, the decisive rule for this classification is additionally established. Mostly, this

rule consists of comparing the sum of points (weights) of characteristics taken into consideration with the previously chosen limit.

CONSTRUCTION OF COMPOSITE INDICES FOR QUALITY VARIABLES (MARKERS)

When obtaining a CI for qualitative variables, the main difficulty is in their transforming into a quantitative scale (digitization). This is done with the help of expert evaluation. If the characteristics are few (up to ten), the ranking of characteristics can be used, followed by normalization of the sum of ranks, a direct score or pair comparison.

In the latter case, it is best to apply the hierarchies method of T. Saati, which is well described in literature [24]. It should be noted that the expert, comparing n characteristics, actually holds $n(n - 1)/2$ comparisons. If the number of characteristics is estimated by the ten, then to facilitate the work of experts, it is recommended to use the step-by-step calculation of the significance by the following algorithm.

Step 1. *Partition of the array of analyzed characteristics* according to semantic contents into m groups, each of them contains n_j characteristics.

Step 2. *Determining the importance* of the ith attribute in each jth group by the kth expert (w_{ijk}) by the Saati method using matrix procedures (calculation of vector eigenvalues (λ_{ijk}) and their normalization):

$$w_{ijk} = \frac{\lambda_{ijk}}{\sum_{i=1}^{n_j} \lambda_{ijk}}.$$

Step 3. *Determining the significance of the signs*, which (like the eigenvectors values) depend on the dimension of the feature matrices in each group (n_j). The more features, the less these values are obtained on average. To compensate for this effect, it is recommended to use a correction coefficient with the appropriate normalization:

$$g_j = \frac{g'_j}{\sum_{j=1}^{m} g'_j},$$

Step 4. *Comparing the significance of the feature groups by experts* for the constructed CI and calculating the normalized coefficients of the significance of the groups:

$$v_{jk} = \frac{v'_{jk}}{\sum_{j=1}^{m} v'_{jk}}.$$

Step 5. *Calculating the coherence of expert assessments*. Since the expert evaluation assumes the activity of a group of experts, it is necessary to evaluate the consistency of their estimation, mostly with the help of correlation analysis. Since the values obtained by the Saati method are quantitative, the correlation can be calculated by the Pearson' formula. If the assessments of some experts do not correlate with the opinion of the rest of the group, these experts are mostly excluded from the group.

To assess the consistency of the results suggested by the *k*th expert when filling in the matrix of paired comparisons (the lack of "logical chains"), the index for consistency (IC):

$$IC_{kj} = (\lambda_{j\max} - n_j)/(n_j - 1).$$

An expert evaluation is considered agreed if the IC < 0.1.

Step 6. *Assigning the classification ratings to experts*. The skill levels of experts may be different. Therefore, it is advisable to assign to each of them a certain standardized qualification coefficient e_k, based on the length of service in this subject area, academic degree and rank, position, etc.

Step 7. *Calculating the total significance of each sign* carried out by the formula:

$$W_{ij} = w_{ijk} v_{jk} g_j e_k .$$

Since normalization was performed at each stage of calculations and the "weights" of experts and groups in the formula have already been taken into account, therefore the sum of the significances of all characteristics is "1", and they can be compared regardless of their belonging to the group.

Thus, the developed technology allows to design a complex indicators for data of any kind of scales (relations, ranks, binary variables and markets). The technology has been tested on a variety of medical, biological and environmental data sets.

CONCLUSIONS

The developed generalized information technology is oriented at supporting the solution of a wide range of problems of medical and ecological researches and integrates various approaches to the processing and analysis of data of different types. Proposed technology consists of four stages: the formation and initial analysis of an initial indicators set, the calculation and normalization for obtainig unnamed equivalents, the actual design of the composite indices, and their verification. The implementation of this technology makes it possible to compare data of different unnamed, determine the significance of specific characteristics in a general research totality, to evaluate the integral state and to classify the research objects.

The use of the proposed modifications of information technology for constructing composite indices based on data of different types: quantitative, rank, binary and qualitative variables, allows to avoid mistakes in the incorrect application of mathematical methods for processing medical and environmental information.

REFERENCES

1. Suter, E, et al. Indicators and Measurement Tools for Health Systems Integration: A Knowledge Synthesis. *International Journal of Integrated Care*, 2017; 17 (6): 4, 1–17. DOI: https://doi.org/10.5334/ijic.3931
2. T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning / data mining, inference, and Prediction. Second edition, 12th printing 2017, 745 p.
3. Antomonov M.Yu., Voloshchuk E.V. Constructing integral indicators of quantitative characteristics using one-dimensional and multidimensional statistical methods. *Kibernetika i vyčislitel'naâ tehnika*. 2012. Iss. 167. P. 61–68 (in Russian).
4. Mikheienko O.I. Integrated method for assessing the health of the human body. *Pedahohika, psykholohiya ta medyko-biolohichni problemy fizychnoho vykhovannya i sportu*. 2011. Iss. 6. P. 93–101 (in Ukrainian).

5. Apanasenko G.L. Diagnosis of individual health. *Gigiyena i sanitariya*. 2004. Iss. 1. P. 55–58 (in Russian).
6. Merkov A.M., Polyakov L.E. Sanitary statistics (manual for doctors). Moscow: Meditsina, 1974. 384 p. (in Russian).
7. Bulich E.G., Muravov I.V. Human health: The biological basis of vital activity and motor activity in its stimulation. Kiev: Olimpiyskaya literatura, 2003. 424 p. (in Russian).
8. Apanasenko G.L. The book is about health. Kiev: Medkniga, 2007. 132 p. (in Russian).
9. Bezruk V.V. Anthropometry. Assessment of physical development of children. Methods of evaluation: methodical instructions for practical classes for students of the third year of medical faculty (specialty "medical psychology"). Chernivtsi, 2008. 19 p. (in Ukrainian).
10. Verevina M.L., Rusakov N.V., Zhukova T.V, Gruzdeva O.A. Assessment of the incidence of the population, depending on living conditions. *Gigiyena i sanitariya.* 2010. Iss. 1. P. 21–25. (in Russian)
11. Bolshakov A.M., Krutko V.N. Integral health indicators and complex systems for their evaluation. *Gigiyena i sanitariya*. 2011. Iss. 6. P. 51 52 (in Russian).
12. Medic V.A., Tokmachev M.S. Manual on Health and Health Statistics. Moscow: Meditsina, 2009. 527 p. (in Russian).
13. Shekera O.G. Health: Basic terms and indicators. *Zdorov□ya suspil□stva*. 2011. Iss. 1. P. 26–31 (in Russian).
14. Krivova O.A., Kozak L.M. Comprehensive assessment of regional demographic development. *Kibernetika i vyčislitel'naâ tehnika*. 2015. Iss. 182. pp. 70–84. (in Russian).
15. Rogozinskaya N.S., Kozak L.M. Mathematical models for the dynamics of statistical indicators for the study of the health status of the population in terms of cancer incidence. *Kibernetika i vyčislitel'naâ tehnika.* 2011, Iss 166. P. 85–96. (in Russian).
16. GOST 2874-82 Drinking water. Hygienic requirements and quality control. — Enter. 85-01-01. Moscow: Izdatel'stvo standartov, 1985. 6 p. (in Russian).
17. Turbinsky V.V., Maslyuk A.I. The risk to the public health of the chemical composition of drinking water. *Hygiene and sanitation*. 2011. Iss. 2. P. 23–27. (in Russian).
18. Gnevashev M.V. Statistical methods for assessing the state of water bodies on a set of ecosystem indicators for water protection purposes. Ekaterinburg, 2006. 42 p. (in Russian).
19. Belogokrov V.P., Lozannsky V.R., Pesina S.A. Application of generalized indicators for assessing the level of contaminated water bodies. Integrated assessment of surface water quality. StPb .: Gidrometeoizdat, 2001. 34 p. (in Russian).
20. Index of atmospheric pollution (IZA) URL: http://moreprom.ru/article.php?id=56. [Last accessed: 08.06.2018] (in Russian).
21. Kakareka S.V. Estimation of total air pollution. *Geografiya i prirodnyye resursy.* 2012. Iss. 2. P. 14–20 (in Russian).
22. Pinigin M.A. Hygienic basis for assessing the degree of air pollution. *Hygiene and sanitation.* 1993. Iss. 7. P. 4–8 (in Russian).
23. Antonomov M.Yu. Mathematical processing and analysis of medico-biological data 2 ed. — Kiev: MEDC "Medinform", 2018. 579 p. (in Russian)
24. Saati T.L. Adoption of decisions. The method of analyzing hierarchies. Moscow: Radio i svyaz', 1989. 316 p. (in Russian).

Resieved 11.06.2018

ЛИТЕРАТУРА

1. Suter, E, et al. Indicators and Measurement Tools for Health Systems Integration: A Knowledge Synthesis. *International Journal of Integrated Care*, 2017; 17(6): 4, 1–17. DOI: https://doi.org/10.5334/ijic.3931
2. T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning/ data mining, inference, and Prediction. Second edition, 12th printing 2017, 745 p.
3. Антомонов М.Ю., Волощук Е.В. Конструирование интегральных показателей количественных признаков с помощью одномерных и многомерных методов статистики. Кибернетика и вычислительная техника. 2012. Вып. 167. С. 61–68.

4. Міхеєнко О.І. Комплексна методика оцінки рівня здоров'я організму людини. *Педагогіка, психологія та медико-біологічні проблеми фізичного виховання і спорту*. 2011. № 6. С. 93–101.
5. Апанасенко Г.Л. Диагностика индивидуального здоровья. *Гигиена и санитария*. 2004. № 1. С.55–58.
6. Мерков А.М., Поляков Л.Е. Санитарная статистика (пособие для врачей). М. : Медицина, 1974. 384 с.
7. Булич Э.Г., Муравов И.В. Здоровье человека: Биологическая основа жизнедеятельности и двигательная активность в ее стимуляции. К.: Олимпийская литература, 2003. 424 с.
8. Апанасенко Г.Л. Книга о здоровье. К.: Медкнига, 2007. 132 с.
9. Безрук В.В. Антропометрія. Оцінка фізичного розвитку дітей. Методи оцінки : методичні вказівки до практичних занять для студентів III курсу медичного факультету (спеціальність „медична психологія"). Чернівці, 2008. 19 с.
10. Веревина М.Л.,. Русаков Н.В,. Жукова Т.В, Груздева О.А. Оценка заболеваемости населения в зависимости от условий проживания. *Гигиена и санитария*. 2010. №. 1. С. 21–25.
11. Большаков А.М., Крутько В.Н. Интегральные индикаторы здоровья и комплексные системы для их оценки. *Гигиена и санитария*. 2011. № 6. С. 51–52.
12. Медик В.А., Токмачев М.С. Руководство по статистике здоровья и здравоохранения. М.: Медицина, 2009. 527 с.
13. Шекера О.Г. Здоров'я: Основні терміни і показники. *Здоров'я суспільства*. 2011. №1. С. 26–31.
14. Кривова О.А., Козак Л.М. Комплексная оценка регионального демографического развития. *Кибернетика и вычислительная техника*, 2015, вып. 182. С. 70–84.
15. Рогозинская Н.С., Козак Л.М. Математические модели динамики статистических показателей для исследования состояния здоровья населения по онкозаболеваемости. *Кибернетика и вычислительная техника*, 2011, вып 166. С. 85–96.
16. ГОСТ 2874-82 Вода питьевая. Гигиенические требования и контроль за качеством. — Введ. 85-01-01. Москва : Изд-во стандартов, 1985. — 6 с.
17. Турбинский В.В., Маслюк А.И. Риск для здоровья населения химического состава питьевой воды. *Гигиена и санитария*. 2011. №. 2. С. 23–27]:
18. Гневашев М.В. Статистические методы оценки состояния водных объектов по комплексу экосистемных показателей для целей водоохраны. Екатеринбург, 2006. 42 с.
19. Белогокров В.П., Лозаннский В.Р., Песина С.А. Применение обобщенных показателей для оценки уровня загрязненных водных объектов. Комплексные оценки качества поверхностных вод. СПб.: Гидрометеоиздат, 2001. — 34с.
20. Индекс загрязнения атмосферы (ИЗА) URL: http://moreprom.ru/article.php?id=56. [Дата обращения: 08.06.2018]
21. Какарека С.В. Оценка суммарного загрязнения атмосферного воздуха. *География и природные ресурсы*. 2012. № 2. С. 14–20.
22. Пинигин М.А. Гигиенические основы оценки степени загрязнения атмосферного воздуха. *Гигиена и санитария*. 1993. №7. С. 4–8.
23. Антомонов М.Ю. Математическая обработка и анализ медико-биологических данных -2 изд. — Киев: МИЦ «Мединформ»", 2018. — 579с.
24. Саати Т. Л. Принятие решений. Метод анализа иерархий. — М.: Радио и связь, 1989. — 316 с.

Получено 11.06.2018

М.Ю. Антомонов, д-р.біол. наук, проф.,
голов. наук. співроб. лаб. епідеміологічних досліджень
і медичної інформатики,
e-mail: antomonov@gmail.com
ДУ «Інститут громадського здоров'я
ім. А.Н. Марзєєва НАМН України»,
Україна, 02660, м. Київ, вул. Попудренко, 50

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КОНСТРУЮВАННЯ КОМПЛЕКСНИХ ПОКАЗНИКІВ ДЛЯ РІЗНИХ ТИПІВ ДАНИХ В МЕДИКО-ЕКОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Вступ. *Одним з ефективних шляхів вирішення проблеми оброблення і аналізу величезних обсягів інформації медичних та екологічних досліджень є застосування інформаційних технологій на основі методів конструювання композитних індексів, що дозволить виконувати інтегральне оцінювання стану екологічних, медичних і біологічних систем.*

Метою статті є *розроблення узагальненої інформаційної технології конструювання комплексних показників для різних типів даних, які використовуються в медико-екологічних дослідженнях.*

Результати. *Медико-екологічні дослідження мають два основних складника: аналіз стану здоров'я людини і навколишнього середовища. У разі розв'язання таких завдань необхідно здійснювати оцінювання та аналіз стану біооб'єкту за даними різних типів: кількісними, ранговими, бінарними і якісними змінними. Розроблену узагальнену інформаційну технологію орієнтовано на підтримку розв'язання широкого кола завдань медико-екологічних досліджень, тому ця технологія інтегрує різні підходи до оброблення і аналізу даних різного типу. Виконання чотирьох етапів запропонованої технології (а саме формування та первинний аналіз комплексу вихідних показників, розрахунок та нормування безрозмірних еквівалентів, конструювання комплексних показників і їх верифікація) дозволяє проводити порівняння даних різної розмірності, визначати значущість конкретних характеристик в загальній дослідницькій сукупності, оцінювати інтегральний стан і здійснювати класифікацію об'єктів дослідження.*

Висновки. *Запропонована інформаційна технологія конструювання комплексних показників за даними різних типів: кількісними, ранговими, бінарними і якісними змінними, є ефективним інструментом для порівняльного аналізу стану біооб'єктів різної природи, її використання дозволяє уникнути помилок некоректного застосування математичних методів оброблення медико-екологічної інформації.*

Ключові слова: *інформаційна технологія, композитні показники, оброблення медичних і екологічних кількісних, рангових, бінарних і якісних змінних.*

М.Ю. Антомонов, д-р биол. наук, проф.,
глав. науч. сотр. лаб. эпидемиологических исследований
и медицинской информатики,
e-mail: antomonov@gmail.com
ГУ «Институт общественного здоровья
им. А.Н. Марзеева НАМН Украины»,
Украина, 02660, г. Киев, ул. Попудренко, 50

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ КОНСТРУИРОВАНИЯ КОМПЛЕКСНЫХ ПОКАЗАТЕЛЕЙ ДЛЯ РАЗНЫХ ТИПОВ ДАННЫХ В МЕДИКО-ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Одним из эффективных путей решения проблемы обработки и анализа больших объемов информации медицинских и экологических исследований является применение информационных технологий на основе методов конструирования композитных индексов, что позволяет выполнять интегральную оценку состояния экологических, медицинских и биологических систем.

Целью статьи является разработка обобщенной информационной технологии конструирования комплексных показателей для различных типов данных, используемых в медико-экологических исследованиях.

Медико-экологические исследования имеют два основных компонента: анализ состояния здоровья человека и окружающей среды. В случае решения таких задач необходимо осуществлять оценку и анализ биообъекта по данным различных типов, т.е. количественными, ранговыми, бинарными и качественными переменными. Разработанная обобщенная информационная технология ориентирована на поддержку решений широкого круга задач медико-экологических исследований, поэтому она интегрирует различные подходы к обработке и анализу данных разного типа. Выполнение четырех этапов предлагаемой технологии (а именно формирование и первичный анализ комплекса исходных показателей, расчет и нормирование безразмерных эквивалентов, конструирование комплексных показателей и их верификация) позволяет проводить сравнение данных различной размерности, определять значимость конкретных характеристик в общей исследовательской совокупности, оценивать интегральное состояние и осуществлять классификацию объектов исследования.

Предложенная информационная технология конструирования комплексных показателей по данным различных типов: количественным, ранговым, бинарным и качественным переменным, является эффективным инструментом для сравнительного анализа состояния биообъектов различной природы, ее использование позволяет избежать ошибок некорректного применения математических методов обработки медико-экологической информации.

Ключевые слова: *информационная технология, композитные показатели, обработка медицинских и экологических количественных, ранговых, бинарных и качественных переменных.*