This paper addresses the task to devise a statistical estimation procedure in an event where the volume of the array of initial data used in processing is insufficient to correctly determine the parameters of the response function. The object of research is the technology of statistical processing of a small sample of data. The subject of the study is the methods of statistical estimation under conditions of a small sample of initial data. The main direction is to devise a special procedure for statistical processing of a small sample of initial data, which provides a correct statistical estimation of the parameters of the response function. The method for solving the problem is the selection of the most representative orthogonal replica-like subplan from the plan of a complete factorial experiment obtained by artificially orthogonalizing the results of a passive experiment. The necessity and expediency of the proposed procedure is a consequence of the unpredictability and uneven distribution of points in the phase space of coordinates. The result of the implementation of the corresponding procedure is a truncated orthogonal plan of the full factorial experiment, which provides the possibility of independent estimation of all coefficients of the regression polynomial describing the response function. Under conditions of a severe shortage of the number of measurements, the procedure makes it possible to isolate a representative orthogonal replica from the resulting plan of a complete factorial experiment. Using this subplan of the full factorial experiment plan makes it possible to evaluate all the coefficients of the regression polynomial that describes the desired response function. The corresponding computational procedure is based on solving the triaxial Boolean assignment problem

Keywords: statistical data processing, small sample, artificial orthogonalization, triaxial assignment problem

UDC 519.8
DOI: 10.15587/1729-4061.2023.282130

STATISTICAL PROCESSING OF A SMALL SAMPLE OF RAW DATA USING ARTIFICIAL ORTHOGONALISATION TECHNOLOGY

Lev Raskin
Doctor of Technical Sciences, Professor*
Larysa Sukhomlyn
PhD, Associate Professor
Department of Management
Kremenchuk Mykhailo Ostrohradskyi National University
Pershotravneva str., 20, Kremenchuk, Ukraine, 39600
Viacheslav Karpenko
PhD, Senior Lecturer*
Dmytro Sokolov
Corresponding author
Postgraduate Student*
E-mail: sokolovddd@gmail.com
*Department of Multimedia
and Internet Technologies and Systems
National Technical University «Kharkiv Polytechnic Institute»
Kyrpychova str., 2, Kharkiv, Ukraine, 61002

Received date 10.03.2023
Accepted date 13.06.2023
Published date 30.06.2023

How to Cite: *Raskin, L., Sukhomlyn, L., Karpenko, V., Sokolov, D. (2023). Statistical processing of a small sample of raw data using artificial orthogonalisation technology. Eastern-European Journal of Enterprise Technologies, 3 (4 (123)), 14–21. doi: https://doi.org/10.15587/1729-4061.2023.282130*

1. Introduction

The quality of solving tasks to assess the effectiveness of systems and control over them depends on the accuracy of identifying the state of these systems. In turn, the efficiency of the state assessment is determined by the quantity and quality of the controlled parameters of the system, as well as the level of correctness of the methods used for processing the results of measuring these parameters. At the same time, in all cases it is assumed that there is a sample of initial data necessary, in accordance with the standard requirements of mathematical statistics, for conducting the relevant research. The problem arises if, for objective reasons, these requirements are not met. Conventional technologies for solving the problem are reduced to the use of various types of techniques to reduce the number of estimated parameters of the response function that determines the quality of the system. The actual inefficiency of these techniques leads to the need to devise other technologies, the use of which would solve the problem of the shortage of initial data. Solving this task could make it possible to tackle many real problems of state assessment and control over systems whose operating conditions are not stable. Ignoring this feature of actual arrays of initial data can lead to unacceptably gross errors in assessing the state of objects and managing them. These circumstances determine the high relevance of research aimed at devising reliable methods for statistical processing of real source data under conditions of their scarcity when the use of conventional technologies is not entirely correct.

2. Literature review and problem statement

To solve the problems of assessing the state of objects based on the results of statistical processing of the initial data, depending on the task set, methods of multivariate discriminant analysis can be used [1]. An obvious drawback of the multivariate discriminant analysis method is as follows. The identification of the state is based on the dissection of the phase space of coordinates corresponding to the controlled

parameters of the system with the help of hyperplanes. The results obtained in this case are easily interpreted if the dimensionality of the phase space is not large [2]. However, if otherwise, the correctness of the decisions made is compromised. Clustering methods [3, 4] confidently solve the problem of state estimation based on the results of measuring a set of controlled parameters [5, 6]. A significant drawback of the method is the inability to take into account the continuity of the processes of transition of the system from one state to another, which is typical for many real objects, and it is unacceptable not to take this into account (for example, in medicine when assessing the condition of patients).

A fundamentally different approach to assessing the state of objects is implemented in expert systems for identifying the state based on the results of measuring the values of controlled parameters. Such systems work as follows [7, 8]. It is assumed that the system can be in one of the set of (H_1, H_2, ..., H_m) states, and n controlled parameters (x_1, x_2, ..., x_n) are used to evaluate the state. The range of possible values for each of the controlled parameters is divided into m intersecting subintervals. For each pair (x_j, H_i), the $\mu(x_j/H_i)$ function is introduced, the value of which is interpreted as the degree of confidence that, based on the result of monitoring the parameter x_j, it is possible to make a decision regarding the system's stay in the H_i state. Various systems of this type differ from each other in the technology of production rules for making decisions about the state of the system.

Expert systems of this kind work successfully if the set of possible states and the number of controlled parameters are small. However, as the values of m and n increase, the total number of necessary production rules N grows rapidly in accordance with the relation $N=m^n$, which in many practical situations significantly complicates the application of this procedure [9]. In addition, the design of the inference mechanism does not provide for the need and usefulness of taking into account the influence of the interaction of factors [10]. Another issue arises when using one of the most powerful parametric identification methods – regression. The meaning and mathematical tools of the corresponding procedure is to calculate some reasonably chosen numerical characteristics, the analysis of which makes it possible to confidently identify the state of the system [11, 12]. To implement this method, regression analysis technology is used.

It is clear that the number of coefficients of the desired regression polynomial to be estimated depends on the number of factors and the order of interactions taken into account and grows rapidly with this order. The natural approach to solving the problem in this case is to reduce the number of factors [13, 14]. To this end, firstly, it is possible to statistically estimate the level of correlation between factors and exclude one of them for each pair of strongly correlated ones [15]. However, the appropriate cut-off level is not discussed. Secondly, it is possible to estimate the informational value of controlled factors, for example, according to the Kullback criterion [16], and remove uninformative ones from the resulting equation. A structural shortcoming of the method is the dependence of the result of evaluating the information content on the order in which its constituent elements are included in the calculation formula. A common drawback of these approaches is their unpredictable a priori efficiency. Another, more promising approach is based on the method of artificial orthogonalization of a passive experiment proposed in [17, 18]. In accordance with this method, an r-factor orthogonal plan is formed based on the results of a passive r-factor experiment. The standard processing of this orthogonal plan makes it possible to estimate all the coefficients of the regression equation and remove the insignificant ones [19, 20]. Under conditions of a small sample of initial data, in some cases it may be useful to use any of the fractional replicas of the full factorial experiment [21, 22]. A significant strengthening of this idea is to find and devise a procedure for extracting an orthogonal representative replica-like truncated plan from the full plan of this experiment, which could be used to estimate all the coefficients of the regression equation.

3. The aim and objectives of the study

The aim of this study is to devise a method for statistical analysis of a small sample of initial data when the available volume of this sample is not sufficient to apply standard methods of mathematical statistics. The practical usefulness of the method is determined by the possibility of its use in a deadlock situation when the actual amount of initial data does not allow the use of conventional approaches of statistical analysis.

To achieve the goal, the following tasks were set:

– to devise a method of initial selection from the plan of a complete factorial experiment of an orthogonal replica-like subplan;

– to devise a method for optimizing the initial choice to obtain the most representative orthogonal subplan.

4. The study materials and methods

The object of our research is the regression method of statistical analysis. The situation is considered when the insufficient available amount of initial data excludes the possibility of using standard methods of mathematical statistics for estimating the coefficients of the desired regression polynomial. The proposed method for solving the problem is based on the idea of extracting an orthogonal representative replica-like subplan from the plan of a complete factorial experiment. To implement the method, a computational technology for solving triaxial Boolean assignment problems is used.

5. Results of investigating the problem of statistical processing of a small sample of initial data

5. 1. Devising a method for choosing an initial replica-like subplan from the plan of a full factorial experiment

Consider the problem of choosing an orthogonal subplan from the plan of a full factorial experiment. Let $N>2^r$ experiments be obtained as a result of an r-factor passive experiment, which are summarized in a matrix containing N rows and 2^r columns. This two-dimensional matrix is converted to a three-dimensional one containing $p=r/3$ rows, columns, and columns. Each section of this matrix (row, column, or front) is a square matrix with p^2 elements. Each element of the resulting three-dimensional matrix will be assigned a number (i, j, k), $i=1, 2, ..., p$, $j=1, 2, ..., p$, $k=1, 2, ..., p$, which determines the row, column, and column for this element. Note that a three-dimensional cube of a full factorial experiment containing p^3 elements (cubes) can, in general, be transformed into a three-dimensional parallelepiped with $i=1, 2, .., m$ rows, $j=1, 2, ..., n$ columns, and $k=1, 2, ..., p$ columns.

In this case, the values of m, n and p must be chosen so that the equality $mnp=r^3$ holds true. For each specific set of numbers i, j, k, we assign the value C_{ijk} equal to the number of experiments that fell inside the cube with this number. The task is to select an orthogonal subplan from the obtained plan of the full factorial experiment, which has the highest representativeness.

Let us proceed to the description of the formal procedure for constructing the required orthogonal subplan. The key element of this procedure is the solution of the triaxial assignment problem, the mathematical model of which takes the following form [23]: find a Boolean set $x=\{x_{ijk}\}$ that maximizes the function:

$$L(x)=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p} C_{ijk}x_{ijk}, \quad (1)$$

and satisfies the constraints:

$$\sum_{k=1}^{p} x_{ijk}=1,\ i=1,2,..,m,\ j=1,2,..,n,$$

$$\sum_{i=1}^{m} x_{ijk}=1,\ j=1,2,..,n,\ k=1,2,..,p, \quad (2)$$

$$\sum_{j=1}^{n} x_{ijk}=1,\ i=1,2,..,m,\ \ k=1,2,..,p.$$

The solution of problem (1), (2) defines a plan in which the selected elements of the matrix $\{C_{ijk}\}$ are located one by one in each one-dimensional section and their sum is maximum.

To solve this problem, in the three-index matrix $\{C_{ijk}\}$ we calculate:

$$\left|C_j^{ik}\right|=\sum_{i=1}^{m}\sum_{k=1}^{p} C_{ijk},\ j=1,2,..,n, \quad (3)$$

$$\left|C_i^{jk}\right|=\sum_{j=1}^{n}\sum_{k=1}^{p} C_{ijk},\ i=1,2,..,m, \quad (4)$$

$$\left|C_k^{ij}\right|=\sum_{i=1}^{p}\sum_{j=1}^{n} C_{ijk},\ k=1,2,..,p. \quad (5)$$

Now we introduce a special transformation of the matrix $\{C_{ijk}\}$ according to the formula:

$$C_{ijk}^{(0)}=C_{ijk}+\alpha_i+\beta_j+\gamma_k, \quad (6)$$

so that:

$$\left|C_i^{(0)jk}\right|=\left|C_j^{(0)ik}\right|=\left|C_k^{(0)ij}\right|=0. \quad (7)$$

The matrix $\left\{C_{ijk}^{(0)}\right\}$, obtained as a result of transformation (6), will be termed normalized, and the coefficients α, β, γ will be termed normalizing.

To find the coefficients α_i, β_j, γ_p, we build a system of linear algebraic equations:

$$\sum_{k=1}^{p}\left(C_{ijk}+\alpha_{ij}+\beta_{jk}+\gamma_{ik}\right)=0,\ i=1,2,..,m,\ j=1,2,..,n,$$

$$\sum_{i=1}^{m}\left(C_{ijk}+\alpha_{ij}+\beta_{jk}+\gamma_{ik}\right)=0,\ j=1,2,..,n,\ k=1,2,..,p, \quad (8)$$

$$\sum_{j=1}^{n}\left(C_{ijk}+\alpha_{ij}+\beta_{jk}+\gamma_{ik}\right)=0,\ i=1,2,..,m,\ \ k=1,2,..,p.$$

The solution of this system of equations can be easily obtained and takes the form:

$$\alpha_{ij}=-\frac{\left|C_{ij}^{k}\right|}{p}+\frac{\left|C_i^{kj}\right|}{np},\ i=1,2,..,m,\ j=1,2,..,n,$$

$$\beta_{jk}=-\frac{\left|C_{jk}^{i}\right|}{m}+\frac{\left|C_j^{ik}\right|}{mp},\ j=1,2,..,n,\ k=1,2,..,p,$$

$$\gamma_{ik}=-\frac{\left|C_{ik}^{j}\right|}{n}+\frac{\left|C_k^{ij}\right|}{mn}-\frac{\left|C_0\right|}{mnp},\ k=1,2,..,p,\ i=1,2,..,m.$$

Then:

$$C_{ijk}^{(0)}=C_{ijk}-\frac{\left|C_{ij}^{k}\right|}{p}-\frac{\left|C_{jk}^{i}\right|}{m}-\frac{\left|C_{ik}^{j}\right|}{n}+$$
$$+\frac{\left|C_i^{jk}\right|}{np}+\frac{\left|C_j^{ik}\right|}{mp}+\frac{\left|C_k^{ij}\right|}{mn}-\frac{C_0}{mnp}, \quad (9)$$

$$i=1,2,..,m,\ j=1,2,..,n,\ k=1,2,..,p.$$

$$\left|C_{ij}^{k}\right|=\sum_{k=1}^{p} C_{ijk};\ \left|C_{jk}^{i}\right|=\sum_{i=1}^{m} C_{ijk};\ \left|C_{ik}^{j}\right|=\sum_{j=1}^{n} C_{ijk};$$

$$\left|C_i^{jk}\right|=\sum_{j=1}^{n}\sum_{k=1}^{p} C_{ijk};\ \left|C_j^{jk}\right|=\sum_{i=1}^{m}\sum_{k=1}^{p} C_{ijk};$$

$$\left|C_k^{ij}\right|=\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ijk}.$$

If $m=n=p$, then:

$$C_{ijk}^{(0)}=C_{ijk}-\frac{\left|C_{ij}^{k}\right|+\left|C_{jk}^{i}\right|+\left|C_{ik}^{j}\right|}{n}+\frac{\left|C_i^{jk}\right|+\left|C_j^{ik}\right|+\left|C_k^{ij}\right|}{n^2}-\frac{C_0}{n^3}. \quad (10)$$

To solve problem (1), (2), (8), a special method [16] is used, the computational complexity of which grows rapidly with increasing problem dimensionality. At the same time, the matrix $\left\{C_{ijk}^{(0)}\right\}$ determined as a result of applying transformation (6) can be productively used to obtain an approximate solution to this problem. The computational procedure contains $n-1$ iterations of the same type. Let $q-1$ iterations of this procedure be performed. At the next q-th iteration, the following steps are performed step by step:

Step 1. A normalizing transformation is carried out according to formula (6).

Step 2. For each $k O\{1, 2, ..., k_{q-1}\}$, the two-index problem is sequentially solved: find the matrix $x=(x_{ijk})$ that maximizes:

$$L(x)=\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ijk}^{(q-1)}x_{ijk}, \quad (11)$$

and satisfies the limitations:

$$\sum_{i=1}^{m} x_{ijk}=1,\ j=1,2,..,n,$$

$$\sum_{j=1}^{n} x_{ijk}=1,\ i=1,2,..,m. \quad (12)$$

Step 3. Let $\widehat{x_k}=\left\{\widehat{x_{ijk}}\right\}$ be the solution to the problem for a specific value of k.

By sorting out k, we choose the best solution to problem (11), (12), that is, we find:

$$\max_{k\in k_{q-1}}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}C_{ijk}\widehat{x_{ijk}}\right\}=\sum_{i=1}^{m}\sum_{j=1}^{n}C_{ijk^*}x_{ijk^*}. \quad (13)$$

Thus, the index k^* determines the number of the two-dimensional section in which the solution to problem (11), (12) gives the best result.

Step 4. The two-dimensional section corresponding to the found value of k is excluded from the matrix $\left\{C_{ijk}^{(0)}\right\}$.

Step 5. The matrix $\left\{C_{ijk}^{(0)}\right\}$, is corrected by introducing a ban in the columns corresponding to non-zero elements of the plan $\left\{x_{ijk^*}\right\}$, according to the rule:

$$C_{ijk}^{(q)}=\begin{cases}C_{ijk}^{(q-1)},(i,j)\notin \mathbf{M}_q,\\ -R,(i,j)\in \mathbf{M}_q,\end{cases}$$

where $\mathbf{M}_q=\left\{(i,j),x_{ijk^*}^{(q)}=1\right\}$, R is a large number.

As a result of this correction, all $\left\{C_{ijk}^{(q)}\right\}$, matrix elements for which $(i,j)OM_q$ cannot be included in subsequent solutions to problem (11), (12).

Step 6. Index k^* completes the set of indices of excluded two-dimensional sections.

In this case, after the $(n-1)$ iteration of the described procedure, the set K_{n-1} of excluded sections will contain $n-1$ components.

Therefore, the remaining two-dimensional matrix is added to those found in previous iterations.

The resulting matrix $x=\{x_{ijk}\}$ is an approximate solution to the original problem (1), (2).

Let us illustrate the technology of finding a truncated orthogonal plan by solving a simple problem. Let's introduce a 2×2×2 matrix $\{C_{ijk}\}$ by specifying two square matrices of its sections for $k=1$ and $k=2$:

$$\left\{C_{ij1}\right\}\begin{pmatrix}C_{121} & C_{221}\\ C_{111} & C_{211}\end{pmatrix}=\begin{pmatrix}9 & 16\\ 8 & 13\end{pmatrix};$$

$$\left\{C_{ij2}\right\}\begin{pmatrix}C_{122} & C_{222}\\ C_{112} & C_{212}\end{pmatrix}=\begin{pmatrix}4 & 10\\ 9 & 11\end{pmatrix}.$$

Let's calculate the values of the components in ratio (10):

$$\left|C_{ij}^{k}\right|=\sum_{k=1}^{p}C_{ijk};\ \left|C_{11}^{k}\right|=8+9=17;\ \left|C_{11}^{k}\right|=13+11=24;$$

$$\left|C_{12}^{k}\right|=9+4=13;\ \left|C_{22}^{k}\right|=16+10=26;$$

$$\left|C_{jk}^{i}\right|=\sum_{i=1}^{m}C_{ijk};\ \left|C_{11}^{i}\right|=8+13=21;$$

$$\left|C_{21}^{i}\right|=9+16=25;\ \left|C_{12}^{i}\right|=9+11=20;$$

$$\left|C_{22}^{i}\right|=4+10=14;\ \left|C_{ik}^{j}\right|=\sum_{j=1}^{n}C_{ijk};$$

$$\left|C_{11}^{j}\right|=9+8=17;\ \left|C_{21}^{j}\right|=16+13=29;$$

$$\left|C_{12}^{j}\right|=9+4=13;\ \left|C_{22}^{j}\right|=10+11=21;$$

$$\left|C_{i}^{jk}\right|=\sum_{j=1}^{n}\sum_{k=1}^{p}C_{ijk};\ \left|C_{1}^{jk}\right|=8+9+9+4=30;$$

$$\left|C_{2}^{jk}\right|=13+16+11+10=50;$$

$$\left|C_{j}^{ik}\right|=\sum_{i=1}^{m}\sum_{k=1}^{p}C_{ijk};\ \left|C_{1}^{ik}\right|=8+13+9+11=41;$$

$$\left|C_{2}^{ik}\right|=19+6+4+10=39;$$

$$\left|C_{k}^{ij}\right|=\sum_{i=1}^{m}\sum_{j=1}^{n}C_{ijk};\ \left|C_{1}^{ij}\right|=8+13+9+16=46;$$

$$\left|C_{2}^{ij}\right|=9+11+4+10=34;$$

$$C_0=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p}=46+34=39+41=30+50=80.$$

Let us now perform a normalization transformation of the elements of the $\{C_{ijk}\}$ matrix:

$$C_{111}^{(0)}=C_{111}-\frac{\left|C_{11}^{k}\right|+\left|C_{11}^{i}\right|+\left|C_{11}^{j}\right|}{n}+\frac{\left|C_{1}^{jk}\right|+\left|C_{1}^{ik}\right|+\left|C_{1}^{ij}\right|}{n^2}-\frac{C_0}{n^3}=$$
$$=8-\frac{17+21+17}{2}+\frac{30+41+46}{4}-\frac{80}{8}=-0.25;$$

$$C_{211}^{(0)}=C_{211}-\frac{\left|C_{21}^{k}\right|+\left|C_{11}^{i}\right|+\left|C_{21}^{j}\right|}{n}+\frac{\left|C_{2}^{jk}\right|+\left|C_{1}^{ik}\right|+\left|C_{1}^{ij}\right|}{n^2}-\frac{C_0}{n^3}=$$
$$=13-\frac{24+21+29}{2}+\frac{50+41+46}{4}-\frac{80}{8}=0.25;$$

$$C_{121}^{(0)}=C_{121}-\frac{\left|C_{12}^{k}\right|+\left|C_{21}^{i}\right|+\left|C_{11}^{j}\right|}{n}+\frac{\left|C_{1}^{jk}\right|+\left|C_{2}^{ik}\right|+\left|C_{1}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=9-\frac{13+25+17}{n}+\frac{30+39+46}{4}-\frac{80}{8}=0.25;$$

$$C_{221}^{(0)}=C_{221}-\frac{\left|C_{22}^{k}\right|+\left|C_{21}^{i}\right|+\left|C_{21}^{j}\right|}{n}+\frac{\left|C_{2}^{jk}\right|+\left|C_{2}^{ik}\right|+\left|C_{1}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=16-\frac{26+25+29}{2}+\frac{50+39+46}{4}-\frac{80}{8}=-0.25;$$

$$C_{112}^{(0)}=C_{112}-\frac{\left|C_{11}^{k}\right|+\left|C_{12}^{i}\right|+\left|C_{12}^{j}\right|}{n}+\frac{\left|C_{1}^{jk}\right|+\left|C_{1}^{ik}\right|+\left|C_{2}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=9-\frac{17+20+13}{2}+\frac{30+41+34}{4}-\frac{80}{8}=0.25;$$

$$C_{212}^{(0)}=C_{212}-\frac{\left|C_{21}^{k}\right|+\left|C_{12}^{i}\right|+\left|C_{22}^{j}\right|}{n}+\frac{\left|C_{2}^{jk}\right|+\left|C_{1}^{ik}\right|+\left|C_{2}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=11-\frac{24+20+21}{2}+\frac{50+41+34}{4}-\frac{80}{8}=-0.25;$$

$$C_{122}^{(0)}=C_{122}-\frac{\left|C_{12}^{k}\right|+\left|C_{22}^{i}\right|+\left|C_{12}^{j}\right|}{n}+\frac{\left|C_{1}^{jk}\right|+\left|C_{2}^{ik}\right|+\left|C_{2}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=4-\frac{13+14+13}{2}+\frac{30+39+34}{4}+\frac{80}{8}=-0.25;$$

$$C_{222}^{(0)}=C_{222}-\frac{\left|C_{22}^{k}\right|+\left|C_{22}^{i}\right|+\left|C_{22}^{j}\right|}{n}+\frac{\left|C_{2}^{jk}\right|+\left|C_{2}^{ik}\right|+\left|C_{2}^{ij}\right|}{n^2}-$$
$$-\frac{C_0}{n^3}=10-\frac{26+14+21}{2}+\frac{50+39+34}{4}-\frac{80}{8}=0.25.$$

The normalization procedure is completed, and the resulting normalized matrices take the form:

$$\left(C_{ij1}^{(0)}\right)=\begin{pmatrix}0.25 & -0.25\\ -0.25 & 0.25\end{pmatrix};$$

$$\left(C_{ij2}^{(0)}\right)=\begin{pmatrix}-0.25 & 0.25\\ 0.25 & -0.25\end{pmatrix}.$$

At the next 2nd step of the computational procedure, for each of the matrices, $\left(C_{ij1}^{(0)}\right)$ and $\left(C_{ij2}^{(0)}\right)$, the two-dimensional assignment problem (11), (12) is solved. In the example under consideration, the obvious solutions to these two problems are determined by the matrices:

$$\widehat{x_1}=\left\{\widehat{x_{ij1}}\right\}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix};$$

$$\widehat{x_2}=\left\{\widehat{x_{ij2}}\right\}=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}.$$

Since:

$$L_1\left(\widehat{x_1}\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij1}^{(0)}\widehat{x_{ij1}}=0.5,$$

$$L_2\left(\widehat{x_2}\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij2}^{(0)}\widehat{x_{ij2}}=0.5.$$

and $L_1\left(\widehat{x_1}\right)=L_2\left(\widehat{x_2}\right)$, then at the third step of the procedure, any of the obtained solutions can be chosen as the best one. Let us choose $k^*=1$ and, performing step 4, exclude from the matrix $\left\{C_{ijk}^{(0)}\right\}$ the section $\left\{C_{ij1}^{(0)}\right\}$.

At the next step 5, we shall correct the matrix $\left\{C_{ijk}^{(0)}\right\}$. Since for $k^*=1$ the matrix is:

$$\mathbf{M}_1=\left\{(i,j),\widehat{x_{ij1}}=1\right\}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},$$

then:

$$C_{ij2}^{(1)}=\begin{pmatrix}-10 & 0.25\\ 0.25 & -10\end{pmatrix}.$$

The solution to problem (11), (12) for this matrix is:

$$\left\{x_{ij2}^*\right\}=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}.$$

Thus, the solution to the original problem, satisfying the constraints, is obtained. Let's calculate the corresponding value of the objective function:

$$L\left(x^*\right)=\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}C_{ijk}x_{ijk}^*=C_{211}+C_{121}+C_{112}+C_{222}=$$
$$=13+9+9+10=41.$$

Thus, a method for obtaining the initial choice of an orthogonal subplan from the plan of a complete factorial experiment is proposed. The first problem is solved.

5. 2. Optimization method for the initial replica-like subplan

To improve the initial solution, we use a step-by-step procedure. At each step, the following operations are performed. Elements of two subsets $\mathrm{M}_1=\{x_{ijk}:x_{ijk}=1\}$, $\mathrm{M}_0=\{x_{ijk}:x_{ijk}=0\}$ are selected in the matrix $x=\{x_{ijk}\}$. In the subset M_0, an element $x_{i_0j_0k_0}$, is found such that $C_{i_0j_0k_0}=\max\limits_{i_0j_0k_0\in\mathrm{M}_0}\left\{C_{ijk}\right\}$. Next, an attempt is made to include this element in the plan. For this purpose, the following elements are found in the row $\{ij_0k_0\}$, column $\{i_0jk_0\}$, and column $\{i_0j_0k\}$:

– $i_1j_0k_0$ such that $x_{i_1j_0k_0}=1$;
– $i_0j_1k_0$ such that $x_{i_0j_1k_0}=1$;
– $i_0j_0k_1$ such that $x_{i_0j_0k_1}=1$;
– $i_1j_0k_1$ such that $x_{i_1j_0k_1}=0$;
– $i_1j_1k_0$ such that $x_{i_1j_1k_0}=0$;
– $i_0j_1k_1$ such that $x_{i_0j_1k_1}=0$;
– $i_1j_1k_1$ such that $x_{i_1j_1k_1}=1$.

These found elements form a fragment of the general plan $x=\{x_{ijk}\}$. Fragment element values are inverted, i.e., x_{ijk} values equal to 0 are assigned the value 1, and values equal to 1 are assigned the value 0.

The local inversion of the plan carried out in this way naturally changes the value of the objective function of the problem. Let us determine the numerical value of this change. We have:

$$\Delta=C_{i_0j_0k_0}+C_{i_0j_1k_1}+C_{i_1j_0k_1}+C_{i_1j_1k_0}-C_{i_0j_0k_1}-$$
$$-C_{i_0j_1k_0}-C_{i_1j_0k_0}-C_{i_1j_1k_1}. \qquad (14)$$

Fragment elements are included in the task plan if $\Delta>0$. Otherwise, the $x_0j_0k_0$ element is excluded from the set M_0 containing potential improvements to the plan, and the procedure returns to the search for a new value $C_{i_0j_0k_0}=\max\limits_{ijk\in\mathrm{M}_0}\left\{C_{ijk}\right\}$. The solution of the problem ends if the set M_0 turns out to be empty by a certain step.

Let's illustrate the described procedure by making an attempt to improve the plan obtained in the example. Based on the results of the approximate solution to the problem, we shall form $\mathbf{M}_1=\left\{x_{211},x_{121},x_{112},x_{222}\right\}$, $\mathbf{M}_0=\left\{x_{111},x_{221},x_{122},x_{212}\right\}$. Let's find $C_{i_0j_0k_0}=\max\limits_{ijk\in\mathrm{M}_0}\left\{C_{ijk}\right\}=\max\left\{8,16,4,11\right\}=16$.

At the same time $i_0j_0k_0=(2\ 2\ 1)$. Using this element, we form a fragment of the plan for a possible improvement of the approximate solution. We have a fragment:

$$x_{221}=0,\ x_{121}=1,\ x_{211}=1,\ x_{222}=1;$$

$$x_{111}=0,\ x_{122}=0,\ x_{212}=0,\ x_{112}=1.$$

By inverting the values for the elements of the fragment, we get a new plan:

$$x_{221}=1,\ x_{121}=0,\ x_{211}=0,\ x_{222}=0;$$

$$x_{111}=1,\ x_{122}=1,\ x_{212}=1,\ \ x_{112}=0.$$

Let us calculate the corresponding change in the value of the objective function:

$$\Delta=C_{221}+C_{111}+C_{122}+C_{212}-C_{121}-C_{211}-C_{222}-C_{112}=$$
$$=16+8+4+11-9-13-10-9=-2. \qquad (15)$$

The change in the value of the objective function is negative. Consequently, the new plan is worse than the previous one and, thus, the approximate solution to the problem obtained above turned out to be an exact solution.

Let us consider the main properties of a representative orthogonal replica-like plan obtained using the described technology.

Let the number of factors supposedly influencing the efficiency of the system be six and let the results of N experiments be known. We shall assume that after the normalization transformation, all sets of experimental results are reduced to a three-dimensional matrix $\{C_{ijk}\}$ of dimensionality 4×4×4. We further assume that for four sections of this matrix $\{C_{ij1}\},\{C_{ij2}\},\{C_{ij3}\},\{C_{ij4}\}$ problems (11), (12) are solved, the results of which are as follows:

$$\{x_{ij1}\}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix};\ \{x_{ij2}\}=\begin{pmatrix}1&0&0&0\\0&0&0&1\\0&1&0&0\\0&0&1&0\end{pmatrix};$$

$$\{x_{ij3}\}=\begin{pmatrix}0&0&0&1\\0&0&1&0\\1&0&0&0\\0&1&0&0\end{pmatrix};\ \{x_{ij4}\}=\begin{pmatrix}0&0&1&0\\0&1&0&0\\0&0&0&1\\1&0&0&0\end{pmatrix}.$$

The obtained solution has the required property: in each one-dimensional section (in rows, columns, and columns) there is one unit, and the remaining elements are equal to zero.

The elements of the obtained matrices will be assigned the numbers of their rows in terms of a six-factor experiment. We have:

$$\{n_{ij1}\}=\begin{pmatrix}1&2^{\bullet}&3&4\\5^{\bullet}&6&7&8\\9&10&11^{\bullet}&12\\13&14&15&16^{\bullet}\end{pmatrix};$$

$$\{n_{ij2}\}=\begin{pmatrix}17^{\bullet}&18&19&20\\21&22&23&24^{\bullet}\\25&26^{\bullet}&27&28\\29&30&31^{\bullet}&32\end{pmatrix};$$

$$\{n_{ij3}\}=\begin{pmatrix}33&34&35&36^{\bullet}\\37&38&39^{\bullet}&40\\41^{\bullet}&42&43&44\\45&46^{\bullet}&47&48\end{pmatrix};$$

$$\{n_{ij4}\}=\begin{pmatrix}49&50&51^{\bullet}&52\\53&54^{\bullet}&55&56\\57&58&59&60^{\bullet}\\61^{\bullet}&62&63&64\end{pmatrix}.$$

In the above matrices, elements equal to 1 are marked with a dot.

Let us present the plan of the complete experiment corresponding to the conditions of the problem (Table 1).

Let us now single out from this plan of the full factorial experiment the rows corresponding to the obtained truncated plan (Table 2).

It is easy to check that this plan is orthogonal. Thus, the task of obtaining a representative orthogonal subplan of the plan of the full factorial experiment has been solved.

Table 1

A plan for a complete factorial experiment

N	F_6	F_5	F_4	F_3	F_2	F_1
1	–	–	–	–	–	–
2	–	–	–	–	–	+
3	–	–	–	–	+	–
4	–	–	–	–	+	+
5	–	–	–	+	–	–
6	–	–	–	+	–	+
7	–	–	–	+	+	–
8	–	–	–	+	+	+
9	–	–	+	–	–	–
10	–	–	+	–	–	+
11	–	–	+	–	+	–
12	–	–	+	–	+	+
13	–	–	+	+	–	–
14	–	–	+	+	–	+
15	–	–	+	+	+	–
16	–	–	+	+	+	+
17	–	+	–	–	–	–
18	–	+	–	–	–	+
19	–	+	–	–	+	–
20	–	+	–	–	+	+
21	–	+	–	+	–	–
22	–	+	–	+	–	+
23	–	+	–	+	+	–
24	–	+	–	+	+	+
25	–	+	+	–	–	–
26	–	+	+	–	–	+
27	–	+	+	–	+	–
28	–	+	+	–	+	+
29	–	+	+	+	–	–
30	–	+	+	+	–	+
31	–	+	+	+	+	–
32	–	+	+	+	+	+
33	+	–	–	–	–	–
34	+	–	–	–	–	+
35	+	–	–	–	+	–
36	+	–	–	–	+	+
37	+	–	–	+	–	–
38	+	–	–	+	–	+
39	+	–	–	+	+	–
40	+	–	–	+	+	+
41	+	–	+	–	–	–
42	+	–	+	–	–	+
43	+	–	+	–	+	–
44	+	–	+	–	+	+
45	+	–	+	+	–	–
46	+	–	+	+	–	+
47	+	–	+	+	+	–
48	+	–	+	+	+	+
49	+	+	–	–	–	–
50	+	+	–	–	–	+
51	+	+	–	–	+	–
52	+	+	–	–	+	+
53	+	+	–	+	–	–
54	+	+	–	+	–	+
55	+	+	–	+	+	–
56	+	+	–	+	+	+
57	+	+	+	–	–	–
58	+	+	+	–	–	+
59	+	+	+	–	+	–
60	+	+	+	–	+	+
61	+	+	+	+	–	–
62	+	+	+	+	–	+
63	+	+	+	+	+	–
64	+	+	+	+	+	+

Table 2

Truncated factor plan

N	F_6	F_5	F_4	F_3	F_2	F_1
2	–	–	–	–	–	+
5	–	–	–	+	–	–
11	–	–	+	–	+	–
16	–	–	+	+	+	+
17	–	+	–	–	–	–
24	–	+	–	+	+	+
26	–	+	+	–	–	+
31	–	+	+	+	+	–
36	+	–	–	–	+	+
39	+	–	–	+	+	–
41	+	–	+	–	–	–
46	+	–	+	+	–	+
51	+	+	–	–	+	–
54	+	+	–	+	–	+
60	+	+	+	–	+	+
61	+	+	+	+	–	–

6. Discussion of results of devising a method for statistical processing of a small sample of initial data

The triaxial Boolean assignment problem (1), (2) to (8) was stated and solved. Using the obtained solution, a procedure for extracting the initial truncated orthogonal replica-like subplan (11) to (13) from the plan of the full factorial experiment was developed.

The selected subplan is improved to obtain the most representative plan. To optimize the initial plan, an appropriate iterative computational procedure is proposed and justified, which implements a step-by-step improvement of the plan (14), (15). At the same time, if the full plan of the r-factor experiment has 2^r rows, then the number of rows of the truncated plan is $2^{2^r/3}$. In the considered example, from the full plan of the six-factor experiment, which has 64 rows, a representative orthogonal plan was selected, in which there are 16 rows. The orthogonality of this plan provides the possibility of estimating all coefficients of the complete regression polynomial describing the response function.

A procedure has been devised that makes it possible to select a representative orthogonal replica-like subplan from the plan of a complete factorial experiment. This approach has no analogs and is fundamentally new. Its most important feature is the integrated use of standard methods of statistical data processing and multi-index optimization tools. The structural feature of the plan obtained in this case is that orthogonality allows using it to independently estimate all coefficients of the polynomial response function. Another important feature of the proposed methodology is determined by the possibility of obtaining the most representative plan. The proposed procedure closes the problem of statistical processing of a small sample of initial data.

The practical usefulness of the procedure manifests itself especially demonstratively in a situation where a sample is presented for statistical processing, the volume of which does not allow for the correct estimation of the parameters of the response function. The orthogonal truncated plan obtained using the obtained procedure successfully solves the problem that arises in this case. A certain disadvantage of the proposed method is the use of the hypothesis that the initial data of the problem are determined exactly. Therefore, the most important direction to advance the study is the extension of the method to the case when the initial data of the problem are defined inaccurately, for example, fuzzy.

7. Conclusions

1. The task to organize statistical processing of an array of initial data under conditions of a small sample has been considered. The analysis of known methods for solving the problem was carried out. Taking into account the shortcomings of these methods, a procedure is proposed that ensures the fulfillment of the standard requirement for the ratio between the number of objective function parameters to be estimated and the sample size. The proposed method is based on the artificial orthogonalization of the processed sample of initial data. In this case, the original passive experiment is transformed into an orthogonal plan of a full factorial experiment. The situation is considered when some elements of the obtained orthogonal plan are not informative. To solve the problem in this case, a procedure for the formation of the initial truncated orthogonal plan of the full factorial experiment, which has the maximum representativeness, was proposed. The problem is solved using the technology for solving the triaxial Boolean assignment problem.

2. To obtain the most representative orthogonal subplan, a step-by-step method for successive improvement of the initial plan with an optimality check after each step has been proposed. In order to confirm the results obtained, an example of solving a six-factor data processing problem is given.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

References

1. Shoba, K. (2019). Multiple diskriminant analysis. Arlington: Inst. for statistics Education, 112.
2. Aouati, M. (2017). Improvement of accuracy of parametric classification in the space of N×2 factors-attributes on the basis of preliminary obtained linear discriminant function. EUREKA: Physics and Engineering, 3, 55–68. doi: https://doi.org/10.21303/2461-4262.2017.00362

3. Luna-Romera, J. M., Martínez-Ballesteros, M., García-Gutiérrez, J., Riquelme, J. C. (2019). External clustering validity index based on chi-squared statistical test. Information Sciences, 487, 1–17. doi: https://doi.org/10.1016/j.ins.2019.02.046
4. Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Cluster Analysis. Wiley. doi: https://doi.org/10.1002/9780470977811
5. Aouati, M. (2018). Improving the accuracy of classifying rules for controlling the processes of deculfuration and dephosphorization of Fe-C melt. Technology Audit and Production Reserves, 2 (3 (46)), 10–18. doi: https://doi.org/10.15587/2312-8372.2019.169696
6. Mourad, A. (2017). Parametric identification in the problem of determining the quality of dusulfusation and deposphoration processes of Fe-C alloy. Technology Audit and Production Reserves, 2 (1 (34)), 9–15. doi: https://doi.org/10.15587/2312-8372.2017.99130
7. Uotermen, D. (1989). Rukovodstvo po ekspertnym sistemam. Moscow: Mir, 388.
8. Dzhekson, P. (2014). Vvedenie v ekspertnye sistemy. Moscow: Vil'yams, 624.
9. Dzharratano, D., Rayli, G. (2016). Ekspertnye sistemy. Moscow: Vil'yams, 1152.
10. Gavrilova, T. A., Khoroshevskiy, V. F. (2011). Bazy znaniy intellektual'nykh sistem. Sankt-Peterburg: Piter, 384.
11. Oimoen, S. (2019). Classical Designs: Full Factorial Designs. STAT Center of Excellence, 26. Available at: https://www.afit.edu/stat/statcoe_files/Classical%20Designs-Full%20Factorial%20Designs_Final.pdf
12. Montgomery, D. (2013). Design and analysis of experiments. Wiley.
13. Burman, L. E., Reed, W. R., Alm, J. (2010). A Call for Replication Studies. Public Finance Review, 38 (6), 787–793. doi: https://doi.org/10.1177/1091142110385210
14. Hari, R. (2022). Replication study.
15. Narayan, C., Das, M. (2009). Design and analysis of Experiments. New of Experiments. Wiley, 53–76.
16. Kullback, S. (1959). Information Theory and statistics. Willey.
17. Seraya, O. V., Demin, D. A. (2012). Lineynyy regressionnyy analiz maloy vyborki nechetkikh iskhodnykh dannykh. Problemy upravleniya i informatiki, 4, 129–142.
18. Domin, D., Sira, O., Raskin, L. (2021). Artificial orthogonalization of a passive experiment for a small sample of fuzzy data for constructing regression equations. Available at: https://ingraph.org/en/products/212
19. Raskin, L. G. (1976). Analiz slozhnykh sistem i elementy teorii optimal'nogo upravleniya. Moscow: Sov. Radio, 344.
20. Domin, D. (2013). Artificial orthogonalization in searching of optimal control of technological processes under uncertainty conditions. Eastern-European Journal of Enterprise Technologies, 5 (9(65)), 45–53. doi: https://doi.org/10.15587/1729-4061.2013.18452
21. Adler, Yu. P., Markova, E. V., Granovskiy, Yu. V. (1971). Planirovanie eksperimenta pri poiske optimal'nykh usloviy. Moscow: Nauka, 282.
22. Domin, D., Sira, O., Raskin, L. (2021). Technology for constructing regression equations for a small sample of passive experiment data. Available at: https://ingraph.org/en/products/211
23. Raskin, L. G. (1982). Mnogoindeksnye zadachi lineynogo programmirovaniya. Moscow: Radio i svyaz', 246.