UDC 519.85
DOI: 10.15587/2706-5448.2026.363099

Lev Raskin,
Oksana Sira,
Larysa Sukhomlyn,
Vitalii Vlasenko,
Ihor Pryshchepa

DEVELOPMENT OF REGRESSION TECHNOLOGY FOR ASSESSING THE STATE OF SEMI-MARKOV SYSTEMS UNDER CONDITIONS OF A SMALL SAMPLE OF INITIAL DATA

The object of research is to assess the state of semi-Markov systems in conditions of a small sample of initial data.

This paper addresses the problem of assessing the state of a multi-element multifactorial stochastic object under conditions of a small sample of fuzzy initial data. Known methods for solving similar problems are ineffective in practical conditions of a small sample of initial data. A real possibility for identifying the relationship between explanatory and explained variables lies in the use of artificial orthogonalization of the results of a passive experiment. In this case, a full factorial experiment design is constructed, the most important property of which is orthogonality. This makes it possible to independently evaluate all the coefficients of the regression equation, which determine the degree of influence of factors and all their interactions on the value of the explained variable. This is achieved through the development of a technology for artificial orthogonalization of a passive experiment and a method for processing the resulting full factorial experiment design. The resulting heterogeneity of the full design is mitigated by finding a truncated representative orthogonal subdesign. The research resulted in a method that enables the calculation of all coefficients of the full regression equation with a small sample of initial data. This method provides a more accurate solution to the problem compared to known methods. The universality of the method lies in the fact that it is implemented in the same way for any set of objective functions. The ability of this method to overcome the computational complexity arising from the large dimensionality of relevant problems allows it to be applied in a wide range of practical areas. The proposed methodology is illustrated step by step by solving an example.

Keywords: *semi-Markov systems, regression polynomial, small sample, artificial orthogonalization, representative subdesign.*

Received: 19.02.2026
Received in revised form: 13.05.2026
Accepted: 29.05.2026
Published: 05.06.2026

How to cite
Raskin, L., Sira, O., Sukhomlyn, L., Vlasenko, V., Pryshchepa, I. (2026). Development of regression technology for assessing the state of semi-Markov systems under conditions of a small sample of initial data. Technology Audit and Production Reserves, 3 (2 (89)), 113–120. https://doi.org/10.15587/2706-5448.2026.363099

1. Introduction

A fundamental feature of mathematical models of real technical, economic, military, and other systems is that they are not Markov models. This means that the distribution laws for the random duration of their stay in each of the possible states can be arbitrary. Among the wide range of specific problems in the research of semi-Markov systems, the nontrivial task of assessing the influence of factors and their interactions on the state of a multifactor system occupies a special place. A rigorous mathematical methodology for solving such problems does not exist, although some significant results are known.

In particular, work [1] presents the results of research on optimizing RFSSW process parameters for AA2024-T3 alloy using analysis of variance (ANOVA), machine learning, and multicriteria optimization. It is shown that the use of ANOVA allows for a quantitative assessment of the influence of individual process factors on the process output characteristics and the identification of statistically significant parameters, while integration with the NSGA-II algorithm enables an effective search for trade-off solutions. However, issues related to the correct allocation of factor contributions in the presence of correlations and complex nonlinear interactions, as well as the unambiguous interpretation of factor influences outside the scope of ANOVA, remain unresolved. Furthermore, implementing the method requires information on the estimation errors of the controlled parameters. Obtaining this information is unattainable in practice.

The paper [2] provides an overview of modern mathematical methods for feature selection, including filtering, wrapper, and embedded methods, as well as approaches based on sensitivity analysis. It is shown that the use of variation-oriented methods, such as the sensitivity analysis of variance (Sobol Indices), allows for a quantitative assessment of the contribution of factors and their interactions to response variation, even in nonlinear models. However, issues related to the computational complexity of such methods for a large number of factors, as well as the ambiguity in the distribution of contributions between correlated features, remain unresolved.

The paper [3] presents the research results of the statistical properties of the SHAP method and its relationship with Functional ANOVA Decomposition. It has been shown that the contribution of factors to the model can be represented as an orthogonal decomposition of the response function into components corresponding to individual factors and their interactions, which provides a theoretically sound interpretation of factor significance. However, issues related to the computational complexity of calculating contributions for a large number of factors, as well as the ambiguity of the distribution of contributions under statistically dependent variables, remain unresolved. That is, for the correct implementation of the method, the controlled parameters

ISSN-L 2664-9969; E-ISSN 2706-5448

must be independent. Verifying the degree to which this requirement is met is impossible.

Among the issues arising when solving the problem of studying semi-Markov systems, a significant one is the problem of the high dimensionality of the observation space under conditions of a small sample of initial data. Various approaches to overcoming this difficulty are considered in the modern literature.

The paper [4] considers the problem of online diagnostics of hidden states using high-dimensional spectroscopic data. The authors note that the initial observation space contains a large number of interconnected features, which significantly complicates the diagnosis of the system state. To overcome the dimensionality issue, latent factor decomposition is used. The original observation space is transformed into a latent low-dimensional representation reflecting the system's key dynamic modes. This reduces computational complexity and improves the robustness of state estimation. A limitation of this approach is the potential loss of informative local features during projection into the latent space. Furthermore, the quality of the diagnostics depends significantly on the correct choice of the latent factor structure and the number of hidden components. Insufficiently representative data can degrade the interpretability of the results.

In [5], the authors study the problem of diagnosing the state of an HSMM under conditions of incomplete and multivariate observations. The authors propose using compressed sensing and sparse reconstruction as the primary mechanism for combating dimensionality. This approach is based on the assumption that the actual dynamics of the system have a sparse structure and can be reconstructed from a limited number of observations. This significantly reduces the number of required measurements and decreases the computational complexity of diagnostics. A disadvantage of this method is its strong dependence on the sparsity hypothesis. If the actual data structure is not sparse, the quality of reconstruction and diagnostics deteriorates sharply. Furthermore, the compressed sensing method is sensitive to noise and requires solving computationally complex optimization problems.

In [6] devoted to diagnosing the degradation states of complex technical systems based on Hidden Semi-Markov Models. The authors demonstrate that an increase in the number of hidden states and diagnostic features leads to a rapid increase in the computational complexity of statistical inference procedures. Gibbs sampling, probabilistic state aggregation, and hidden-state compression are used to reduce dimensionality and stabilize diagnostics. This approach approximates posterior distributions without explicitly calculating the full state space, reducing computational burden. Limitations of the method include the high computational complexity of Monte Carlo procedures, slow convergence of sampling algorithms, and sensitivity to the choice of state aggregation structure. Furthermore, probabilistic state space compression can lead to the loss of information about rare degradation modes.

Therefore, the problem of studying semi-Markov systems, which operational efficiency depends on a large number of factors and their interactions, remains relevant.

The object of this research is to assess the state of semi-Markov systems with a small sample of initial data.

The aim of research is to develop a mathematical framework for estimating the coefficients of a regression polynomial that relates the values of influencing factors and their interactions to the system state. To achieve this aim, the following objectives must be solved:

1) transform the results of a passive experiment to assess the system's state into an orthogonal design for a full factorial experiment;

2) develop a method for selecting a representative truncated orthogonal subdesign from the full factorial experiment design, ensuring the determination of the significance level of the selected group of factors for assessing the system's state.

2. Materials and Methods

Numerous problems of assessing and predicting the state of systems are solved using a specially developed response function. This function establishes the dependence of the system's output variable on the values of the input variables that determine the mode, operating conditions, and values of the system's controlled parameters. The set of possible factor values forms a factor space. A fairly common situation is when a multivariate regression polynomial is used to describe the response function [7–9]. The response function coefficient values are estimated using the least-squares method [10] using the measured factor values and the corresponding response function value in a set of experiments.

The following scientific methods were used in the research:

– the least-squares method;

– methods for solving Boolean mathematical programming problems.

The least-squares method was used to calculate the regression polynomial coefficients.

The method of sequentially enumerating designs of the optimization problem, improved at each step, results in a representative truncated orthogonal design. In the simplest special case, when the regression polynomial is linear, it takes the form

$$y = a_0 + a_1 x_1 + .. + a_m x_m. \quad (1)$$

Let's write the necessary relations for calculating the coefficients of the response function (1) in matrix form. For this purpose, it is possible to introduce the matrix H and the vectors A and Y:

$$H = \begin{pmatrix} x_{11}\ x_{12}\ x_{1m} \\ x_{21}\ x_{22}\ x_{2m} \\ \cdots\cdots\cdots \\ x_{n1}\ x_{n2}\ x_{nm} \end{pmatrix},\ A = \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_m \end{pmatrix},\ Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}. \quad (2)$$

Next, it is possible to introduce the least-squares functional [10]

$$J = (HA - Y)^T (HA - Y), \quad (3)$$

minimizing it with respect to A yields the vector of estimates

$$\hat{A} = (H^T H)^{-1} H^T Y. \quad (4)$$

This relation can be obtained by differentiating (3) with respect to the vector A. To differentiate the scalar function (3) with respect to the vector A, it is possible to introduce the following useful technique. Using (3), it is possible to define an auxiliary scalar function

$$\hat{J}(t) = J(A + ht) = [H(A + ht) - Y]^T [H(A + ht) - Y].$$

Next, it is possible to find

$$\frac{d\hat{J}(H)}{dt}\bigg|_{t=0} = \frac{d}{dt}\left\{\left[H(A+ht)-Y\right]^T\left[H(A+ht)-Y\right]\right\} =$$

$$= \frac{d}{dt}\left\{\left[Y-H(A+ht)\right]^T\left[Y-H(A+ht)\right]\right\} =$$

$$= \frac{d}{dt}\left\{\left[Y^T - A^T H^T - h^T H^T t\right]\left[Y - HA - Hht\right]\right\} =$$

$$= \frac{d}{dt}\Big[Y^T Y - Y^T HA - Y^T Hht - A^T H^T Y + A^T H^T HA +$$

$$+ A^T H^T Hht - h^T H^T Yt + h^T H^T HAt + h^T H^T Hht^2\Big]\Big|_{t=0} =$$

$$= Y^T Hh + A^T H^T Hh - h^T H^T Y + h^T H^T HA.$$

Since $Y^T Hh, h^T H^T Y, A^T H^T Hh, h^T H^T HA$ are scalars, then $h^T H^T Y = Y^T Hh$ and $h^T H^T HA = A^T H^T Hh$. Then

$$\left.\frac{dJ(t)}{dt}\right|_{t=0} = 2(Y^T Hh + A^T H^T Hh) = 2(-Y^T H + A^T H^T H)h.$$

Thus, the desired derivative $dJ(A)/dA$ is a matrix operator acting on h. Therefore

$$\frac{dJ(A)}{dA} = 2(-Y^T H + A^T H^T H). \quad (5)$$

Equating this ratio to zero leads to an equation for A. So $-Y^T H + A^T H^T H = 0$. From this $A^T H^T H = Y^T H$, or, transposing, $H^T HA = H^T Y$, from which, it is possible to obtain the desired estimate for vector A

$$A = (H^T H)^{-1} H^T Y. \quad (6)$$

The described technology is implemented in a standard manner and enables the analytical dependence of the response function values on the values of factors influencing the system's state. It should be noted that the successful implementation of this technology depends significantly on the dimensionality of the response function, and in many practical problems, the corresponding complexity becomes difficult to overcome.

3. Results and Discussion

A possible method for solving these problems with a small sample of initial data is as follows.

For each factor of the regression polynomial, the boundaries of the range of possible values and their levels are determined. In the simplest case, there are two: lower and upper. For each factor, scaling is performed using the formula

$$\hat{x}_j = \frac{2x_j - (x_j^{\max} + x_j^{\min})}{x_j^{\max} - x_j^{\min}} \in [-1;1].$$

According to this: $\hat{x}_j \min = -1, \hat{x}_j \max = 1$.

As a result, in all experiments, all factors will take on extreme values. Then, the design matrix, for example, in a three-factor two-level experiment, will have the form shown in Table 1.

Table 1

Design of a complete three-factor experiment

No.	$F_1F_2F_3$	F_2F_3	F_1F_3	F_1F_2	F_3	F_2	F_1	y
0	–	+	+	+	–	–	–	y_0
1	+	+	–	–	–	–	+	y_1
2	+	–	+	–	–	+	–	y_2
3	–	–	–	+	–	+	+	y_3
4	+	–	–	+	+	–	–	y_4
5	–	–	+	–	+	–	+	y_5
6	–	+	–	–	+	+	–	y_6
7	+	+	+	+	+	+	+	y_7

Let x_{ej} be the value of the j-th factor in the i-th experiment $i = 0,1,..,N-1$, $j = 1,2,..,m$. Important properties of the resulting experimental design should be noted [11]:

1. Symmetry relative to the center of the experiment

$$\sum_{i=0}^{N-1} x_{ij} = 0, \; j = 1,2,..,m.$$

2. Normalization conditions are met

$$\sum_{i=0}^{N-1} x_{ij}^{\;2} = 2^m, \; j = 1,2,..,m.$$

3. Rotatability – the accuracy of predicting the response function value at equal distances from the center of the experiment (0; 0) is the same.

4. Orthogonality of the column vectors of the design matrix.

This last property is crucial; it determines the potential feasibility of constructing an experimental design with a small sample of initial data. The orthogonality of the design matrix radically simplifies the procedure for calculating the coefficients of the regression polynomial. Taking (1) into account, the planning matrix is written column-wise

$$H = \begin{pmatrix} x_{10} & x_{11}...x_{1m} \\ x_{20} & x_{21}...x_{2m} \\ ..... \\ x_{n0} & x_{n1}....x_{nm} \end{pmatrix} = (x_0 x_1 ... x_m),$$

in this case

$$H^T = \begin{pmatrix} x_0^{\;T} \\ x_1^{\;T} \\ .... \\ x_m^{\;T} \end{pmatrix},$$

then:

$$H^T H = \begin{pmatrix} x_0^{\;T} \\ x_1^{\;T} \\ .... \\ x_n^{\;T} \end{pmatrix} (x_0^{\;\sigma} x_1^{\;\sigma} ... x_n^{\;\sigma}) =$$

$$= \begin{pmatrix} x_0^{\;T}x_0 & x_0^{\;T}x_1....x_0^{\;T}x_m \\ x_1^{\;T}x_0 & x_1^{\;T}x_1....x_1^{\;T}x_m \\ ..... \\ x_m^{\;T}x_0 & x_m^{\;T}x_1....x_m^{\;T}x_m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ & 1 \\ ..... \\ 0 & 1 \end{pmatrix},$$

$$(H^T H)^{-1} = \frac{1}{2^{m+1}} \begin{pmatrix} 1 & 0 \\ & 1 \\ ..... \\ 0 & 1 \end{pmatrix}. \quad (7)$$

Substituting (7) into (4) yields

$$\hat{A} = (H^T H)^{-1} H^T Y = \frac{1}{2^{m+1}} \begin{pmatrix} 1 & 0 \\ & 1 \\ ..... \\ 0 & 1 \end{pmatrix} H^T Y =$$

$$= \frac{1}{2^{m+1}} H^T Y = \frac{1}{2^{m+1}} \begin{pmatrix} x_0^{\;T} \\ x_1^{\;T} \\ .... \\ x_m^{\;T} \end{pmatrix} Y = \frac{1}{2^{m+1}} \begin{pmatrix} x_0^{\;T}Y \\ x_1^{\;T}Y \\ .... \\ x_m^{\;T}Y \end{pmatrix}. \quad (8)$$

Thus, the calculation relationship for calculating the coefficient a_j of the regression polynomial (1) is

$$a_j = \frac{1}{2^{m+1}} \sum_{i=0}^{n-1} x_{ij} y_i, \; j = 0,1,..,m. \quad (9)$$

ISSN-L 2664-9969; E-ISSN 2706-5448

An orthogonal design of a full n-factorial design in a 2^m-dimensional factor space corresponds to 2^m hypercube vertices. To implement the standard technology for finding the regression polynomial coefficients, it is necessary to conduct the appropriate number of experiments in each of the subspaces adjacent to the corresponding vertices of the m-dimensional hypercube. However, implementing such an active experiment is not always feasible. In practice, to solve problems of assessing the condition of objects during their operation, the results of passive experiments are used as available initial data. The main drawback of such experiments is the difficulty of achieving a satisfactory ratio between the number of polynomial coefficients and the number of experiments. In a situation where the degrees of influence of factors $x_1, x_2, .., x_m$ and their interactions on the value of the output variable are not known in advance, the required excess of the number of experiments over the number of estimated polynomial parameters is practically impossible to achieve [12]. Potential improvements in this situation can be achieved in various ways [13].

The first approach is to calculate the correlation matrix between the polynomial parameters. If the correlation level for a given pair of factors is high, one of these factors can be discarded.

The second approach is as follows. The set of observed factor values is divided into two subsets corresponding to two different states of the object (H_1, H_2). The measured factor values in each subset are used to calculate histograms, followed by a smooth approximation. The resulting function $f(x)$ for each factor x is normalized using the formula

$$\hat{f}(x)=\frac{f(x)}{\int_0^\infty f(x)\mathrm{d}x}. \quad (10)$$

The function $\hat{f}(x)$ has all the properties of a distribution density (it is non-negative and its integral is equal to 1).

Let densities $\hat{f}(x_j/H_1)$ and $\hat{f}(x_j/H_2)$ be defined for some x_j, corresponding to the states H_1 and H_2 of the object. Then, the Kullback-Leibler measure [14], calculated using the formula

$$\tau=\int_0^\infty \hat{f}(x_j/H_1)\log\frac{\hat{f}(x_j/H_1)}{\hat{f}(x_j/H_2)}\mathrm{d}x_j. \quad (11)$$

The shortcomings of measure (11) are obvious. First, the parameter τ can take values from the interval $[0,\infty]$, meaning this value of τ is not standardized, which reduces its informative value. However, a more important drawback is the second drawback: measure (11) is asymmetric, meaning its value depends on the order in which the densities $\hat{f}(x_j/H_1)$ and are $\hat{f}(x_j/H_2)$ entered into formula (11). Moreover, to compare the informativeness of factors x_1 and x_2 to distinguish states H_1 and H_2, any of the following options can be chosen:

$$\tau_{11}=\int_0^\infty \hat{f}(x_1/H_1)\log\frac{\hat{f}(x_1/H_1)}{\hat{f}(x_1/H_2)}\mathrm{d}x_1,$$

$$\tau_{12}=\int_0^\infty \hat{f}(x_2/H_2)\log\frac{\hat{f}(x_1/H_2)}{\hat{f}(x_1/H_1)}\mathrm{d}x_1,$$

$$\tau_{21}=\int_0^\infty \hat{f}(x_2/H_1)\log\frac{\hat{f}(x_2/H_1)}{\hat{f}(x_2/H_2)}\mathrm{d}x_2,$$

$$\tau_{22}=\int_0^\infty \hat{f}(x_2/H_2)\log\frac{\hat{f}(x_2/H_2)}{\hat{f}(x_2/H_1)}\mathrm{d}x_2.$$

It is clear that the calculated values of τ_{11}, τ_{12}, τ_{21}, τ_{22} can differ unpredictably, resulting in different estimates of the informativeness of factors x_1 and x_2.

A possible alternative measure of the informativeness of controlled factors when identifying states H_1 and H_2 is

$$\varsigma=1-\int_0^\infty\left[\hat{f}(x/H_1)\hat{f}(x/H_2)\right]^{\frac{1}{2}}\mathrm{d}x. \quad (12)$$

The value of the measure ζ is zero if the densities do not intersect, i. e., $\hat{f}(x/H_1)\cdot\hat{f}(x/H_2)=0$, $x\in[0;\infty]$, and is equal to 1 if these densities coincide, i. e., if $\hat{f}(x/H_1)=\hat{f}(x/H_2)$, $x\in[0;\infty]$. Moreover, the value monotonically increases as the length of the interval of coincidence of the densities decreases.

Both of these approaches to assessing the informativeness of controlled factors are appropriate; however, their results may not be sufficient for correctly assessing the significance of factors remaining after screening out those with little informativeness.

Significantly more radical approaches to the problem of reducing the number of estimated coefficients of a regression polynomial involve using the results of active experiments, as follows.

Problem 1. A procedure for artificially orthogonalizing a passive experiment is therefore considered. First, the factor values measured during the passive experiment are scaled. Then, using (4), the coefficients A of the regression polynomial are calculated. The completeness of the vector in each specific case is determined by the ratio between the number of available experiments and the number of estimated parameters of the regression polynomial, which must be no less than the required number [15, 16]. The resulting regression polynomial is then used to calculate the estimated values of the response function at the points in the factor space corresponding to the vertices of the orthogonal hypercube. The set of these values determines the vector Y. This preparatory work enables the use of specific techniques for reducing the dimensionality of the problem, determined by the properties of orthogonal designs.

Problem 2. The simplest of these techniques involves using replicas of the full factorial design [16]. In this case, subdesigns containing only those experiments for which some factors or their combinations assume a fixed value are selected from the overall full design. For example, if in the experimental design matrix (Table 1) only those rows in which $x_2 = 1$ are selected and the experiments for which $x_2 = -1$ are deleted, the remaining truncated design will contain only four of the eight rows (numbered 2, 3, 6, 7). Therefore, such a design is called a semi-replica. The remaining design will still be orthogonal, and relations (9) can be used to estimate the values of the regression polynomial. Using such an orthogonal design halves the number of experiments processed without reducing the number of estimated regression polynomial coefficients. However, the fundamental design flaw of this method is the combination of the influence of certain factors and interactions. Indeed, let's say, for example, a semi-replica is selected for which $x_2 = 1$ for all experiments. Multiplying both sides of this equation by x_1 yields $x_1x_2 = x_2$. This means that in all experiments in the remaining design, the values of x_2 and x_1x_2 will be identical, and it is impossible to separate the influence of factor x_2 and the pairwise interaction x_1x_2 on the outcome variable. Successively multiplying the equation $x_2 = 1$ by x_1x_2, x_1x_3, and x_2x_3 yields the following equations: $x_2x_3 = x_3$, $x_1x_2 = x_1$, $x_1x_2x_3 = x_1x_3$. Thus, the influence of factors is superimposed on the influence of interactions. As the number of factors increases, the number of confounding influences of factors and their interactions rapidly increases, degrading the accuracy of the resulting regression polynomial.

Another drawback of the technology for processing measurement results using fractional replicas is the following. Each replica selects a subset of the entire set of experiments, depending on the specific factor (or their interaction) selected for subsequent selection of the design being processed. Moreover, the subset selected for processing may be unsuccessful due to the random nature of the distribution of

experiments in the observation space. The total number of ways in which a replica of granularity 2^{-K} can be selected is equal to $2^K C_{2^m}^K$. This number increases rapidly with an increase in the number of factors and the level of granularity. For example, if $m = 8, K = 3$, then it is equal to $2^3 \cdot C^3_{256} \cong 2 \cdot 10^7$. Clearly, enumerating all possible replicas to select the most representative one is not feasible in this case.

An alternative approach that radically improves the situation consists of developing a method for constructing and using an orthogonal maximally representative replica-like design [17].

Let's consider possible formulations of the criterion for the feasibility of selecting a truncated representative orthogonal subdesign from a full orthogonal design. The following notation is introduced: j – the hypercube vertex number, $j = 1, 2, ..., 2^m$; d_j – the number of points in the subspace adjacent to the j-th vertex of the hypercube;

$$\eta_j = \begin{cases} 1, \text{ if the } j\text{-th vertex is selected,} \\ M(\text{"large number"}), \text{ otherwise,} \end{cases}$$

$L_1 = \sum_{j=1}^{m} d_j \eta_j$ – the total number of points in the selected space;

$$L_2 = \min_j d_j \eta_j;$$

$$L_3 = \max_j d_j \eta_j - \min_j d_j \eta_j.$$

The formulations of the subdesign formation options are as follows.

Option one: the selected set L_1 contains the maximum number of points.

In this case, the optimal truncated subdesign corresponds to

$$L_1 = \sum_{j=1}^{2^m} d_j \eta_j \Rightarrow \max.$$

Option two: within the selected set of subspaces, the subspace with the minimum saturation in terms of the number of trials contains the maximum number of trials

$$L_2 = \min_j d_j \eta_j \Rightarrow \max.$$

Option three: among the selected subspaces, the difference in saturation level between the most and least saturated subspaces should be minimal. That is

$$L_3 = \left(\left(\max_j d_j \eta_j - \min_j d_j \eta_j \right) \sum_{j=1}^{2^m} \eta_j \right) \Rightarrow \min.$$

Obviously, other options for selecting a subdesign are also possible.

The meaning and computational procedure for implementing this method are most easily explained with a specific example. Let's conduct a six-factor passive experiment, which is transformed into an active one using artificial orthogonalization, and thus obtain a hypercube with the number of vertices equal to $2^6 = 64$. Now it is possible to formulate the problem of extracting from the obtained full-factorial experiment an orthogonal subdesign with the number of experiments corresponding to the fractional replica $1/4$. In this case, the number of experiments in such a design will be equal to $64/4 = 16$. The entire set of factors is divided into three subsets A_1, A_2, A_3 with a $p = m/3 = 2$ factor in each. Further, for each index i, the set of combinations of factors from the subset A_i creates the set I_i, $i = 1, 2, 3$. Since the number of elements in each of the subsets A_i is two, the number of elements in each of the sets Ii is four, and the number of elements in the composition of I_1, I_2, I_3, is $4^3 = 64$. Thus, each specific combination of elements $i_1 \in I_1, i_2 \in I_2, i_3 \in I_3$ determines the corresponding specific combination of factor values, the number of which is 64, and the corresponding row in the design.

Next, the indicator i_1, i_2, i_3 is introduced, which is set to 1 if the row i_1, i_2, i_3 is included in the design, and 0 otherwise.

Next, a system of Diophantine equations is formed:

$$\begin{aligned} &\sum_{i_1 \in I_1} x_{i_1 i_2 i_3} = 1, \quad i_2 = I_2, \ i_3 = I_3, \\ &\sum_{i_2 \in I_2} x_{i_1 i_2 i_3} = 1, \quad i_1 \in I_1, \ i_3 = I_3, \\ &\sum_{i_3 \in I_3} x_{i_1 i_2 i_3} = 1, \quad i_1 \in I_1, \ i_2 \in I_2. \end{aligned} \tag{13}$$

A set of elements from a matrix $x = \{x_{i_1 i_2 i_3}\}$ with fixed values of two indices is called a one-dimensional section. If indices i_2 and i_3 are fixed, then the set $\{x_{i i_2 i_3}\}$, $i = \{1,2,..,2^p\}$ is called a row; if i_1 and i_3 are fixed, then the set is called a column $\{x_{i_1 j i_3}\}$; if i_1 and i_2 are fixed, then the set is called a column $\{x_{i_1 i_2 k}\}$. A set of elements from $x = \{x_{i_1 i_2 i_3}\}$ with a fixed value of one element i_1, i_2, or i_3 is called a two-dimensional section, respectively – vertical, frontal, or horizontal.

The solution matrix of system of equations (13) has the property that in each of its one-dimensional sections only one element is equal to one, while the rest are equal to zero. Moreover, any solution to system (13) determines a certain symmetric orthogonal truncated experimental design [17]. This crucial circumstance determines the possibility of using the resulting truncated design of a full factorial experiment for independent estimation of all coefficients of the regression polynomial.

To solve system of equations (13), all matrix elements $\{x_{i_1 i_2 i_3}\}$ are numbered in order, with a specific matrix cross-section selected, for example, the horizontal cross-section. The result of this numbering is shown in Table 2.

Table 2

Horizontal cross-sections

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

Let's assume that after performing the orthogonalization procedure for a passive experiment, it is possible to obtain a distribution of the number of experiments across the regions (horizontal sections) of the factor space adjacent to the corresponding vertices of the resulting three-dimensional cube:

$$M_1 = \begin{matrix} 3 & 9 & 2 & 8 \\ 6 & 2 & 6 & 4 \\ 8 & 2 & 5 & 4 \\ 3 & 9 & 3 & 8 \end{matrix}; \quad M_2 = \begin{matrix} 3 & 9 & 2 & 8 \\ 9 & 2 & 8 & 8 \\ 1 & 3 & 7 & 9 \\ 6 & 6 & 8 & 4 \end{matrix};$$

$$M_3 = \begin{matrix} 7 & 9 & 8 & 6 \\ 8 & 7 & 2 & 7 \\ 3 & 9 & 7 & 8 \\ 8 & 6 & 9 & 5 \end{matrix}; \quad M_4 = \begin{matrix} 8 & 3 & 2 & 6 \\ 4 & 9 & 8 & 8 \\ 6 & 8 & 2 & 7 \\ 9 & 5 & 6 & 2 \end{matrix}. \tag{14}$$

According to (13), a single element must be selected in each row and column of each horizontal section. These elements, taken together, define a truncated orthogonal design. The task is to select the best, most representative design from a set of possible designs. The level of representativeness of a design can be assessed in various ways. The simplest way is to determine it by the sum of the number of experiments included in the selected subregions. A simple calculation determines this sum for each section. In this case, it is possible to obtain $n_1 = 82$,

ISSN-L 2664-9969; E-ISSN 2706-5448

$n_2 = 93$, $n_3 = 109$, $n_4 = 100$. Thus, the natural rational order of solving the problem in descending order of the number of experiments for a set of horizontal sections is: M_3, M_4, M_2, M_1. Accordingly, the problem for matrix M_3 is solved first. The number of experiments $C_{i_1 i_2 i_3}$ falling within the subdomain $i_1 i_2 i_3$ is introduced. Now, for matrix

$$M_3 = \begin{matrix} 7 & 9 & 8 & 6 \\ 8 & 7 & 2 & 7 \\ 3 & 9 & 7 & 8 \\ 8 & 6 & 9 & 5 \end{matrix}$$

the following problem is solved: find a set $\{x_{i_1 i_2 i_3}\}$, $i_3 = 1$, $i_1 = 1,2,3,4$, $i_2 = 1,3,4,4$, that maximizes

$$F = \sum_{i_1=1}^{4} \sum_{i_2=1}^{4} C_{i_1 i_2 3} x_{i_1 i_2 3}$$

and satisfies the constraints:

$$\sum_{i_1=1}^{4} x_{i_1 i_2 3} = 1, \quad i_2 = 1,2,3,4,$$

$$\sum_{i_2=1}^{4} x_{i_1 i_2 3} = 1, \quad i_1 = 1,2,3,4.$$

The formulated problem is the so-called "assignment problem", which is solved by a well-known algorithm [17]. For matrix M_3, this solution has the form

$$x_3 = \begin{matrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}.$$

The problems for matrices M_4, M_2, M_1 are solved next in turn. Now, since each column of the final matrix must contain only one 1 (constraints (13)), the columns in these matrices with the numbers ($i_1 = 1$, $i_2 = 3$), ($i_1 = 2$, $i_2 = 1$), ($i_1 = 3$, $i_2 = 1$), ($i_1 = 4$, $i_2 = 3$) must be prohibited from further selection. For this purpose, the corresponding elements in (14) are assigned a value equal to a large negative number in absolute value C. In this case, the corresponding matrices (14) are modified to the form:

$$M_2 = \begin{matrix} 8 & C & 2 & 6 \\ C & 9 & 8 & 8 \\ 6 & 8 & 2 & C \\ 9 & 5 & C & 2 \end{matrix}; \; M_4 = \begin{matrix} 3 & C & 2 & 8 \\ C & 2 & 8 & 8 \\ 1 & 3 & 7 & C \\ 6 & 6 & C & 4 \end{matrix}; \; M_1 = \begin{matrix} 7 & C & 2 & 8 \\ C & 2 & 6 & 4 \\ 8 & 2 & 5 & C \\ 3 & 9 & C & 8 \end{matrix}. \qquad (15)$$

Solving the assignment problems for matrices M_2, M_4, M_1 in a similar manner to the previous one, the corresponding solutions are obtained:

$$x_2 = \begin{matrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{matrix}; \; x_4 = \begin{matrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}; \; x_1 = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}. \qquad (16)$$

The resulting matrices x_1, x_2, x_3, x_4 uniquely define the truncated design by matching the positions of the selected elements of these matrices with the row numbers in the full factorial design (Table 2). Thus, the matrices x_1, x_2, x_3, x_4 form a truncated orthogonal subdesign containing the following rows of the full factorial design, numbered 4, 7, 10, 13, 18, 21, 27, 32, 35, 40, 41, 46, 49, 54, 60, 63. A fragment of the full factorial design is described in Table 3, and a fragment of the truncated orthogonal subdesign in Table 4. The right column of Table 5 lists the numbers of experiments falling within each region, taken from (14) in the right column.

Table 3

Full six-factor experimental design (fragment)

Number of rows	F_6	F_5	F_4	F_3	F_2	F_1	y
1	–	–	–	–	–	–	3
2	–	–	–	–	–	+	9
3	–	–	–	–	+	–	3
4	–	–	–	–	+	+	8
5	–	–	–	+	–	–	8
6	–	–	–	+	–	+	2
7	–	–	–	+	+	–	5
8	–	–	–	+	+	+	4
9	–	–	+	–	–	–	6
10	–	–	+	–	–	–	2
11	–	–	+	–	+	–	6
12	–	–	+	–	+	+	4
..	..	..	..	..	..	..	..
18	–	+	–	–	–	–	6
19	–	+	–	–	+	–	8
20	–	+	–	–	–	+	4
..	..	..	..	..	..	..	..
58	+	+	+	–	–	+	9
59	+	+	+	–	+	–	8
60	+	+	+	–	+	+	8
61	+	+	+	+	–	–	8
62	+	+	+	+	–	+	3
63	+	+	+	+	+	–	2
64	+	+	+	+	+	+	6

From this full factorial design, the resulting orthogonal subdesign is extracted.

Table 4

Truncated orthogonal experiment design

No.	F_6	F_5	F_4	F_3	F_2	F_1	y
4	–	–	–	–	+	+	y_4
7	–	–	–	+	+	–	y_7
10	–	–	+	–	–	+	y_{10}
13	–	–	+	+	–	–	y_{13}
18	–	+	–	–	–	+	y_{18}
21	–	+	–	+	–	–	y_{21}
27	–	+	+	–	+	–	y_{27}
32	–	+	+	+	+	+	y_{32}
35	+	–	–	–	+	–	y_{35}
40	+	–	–	+	+	+	y_{40}
41	+	–	+	–	–	–	y_{41}
46	+	–	+	+	–	+	y_{46}
49	+	+	–	–	–	–	y_{49}
54	+	+	–	+	–	+	y_{54}
60	+	+	+	–	+	+	y_{60}
63	+	+	+	+	+	–	y_{63}

Table 4 contains the response function values corresponding to the selected rows of the resulting truncated design, forming the vector y_{01}.

Below is a fragment of the factor interaction table, which includes all pairwise interactions of factor F_1 with other factors, as well as one of the triple interactions (Table 5).

Table 5

Pairwise interaction design of factor F_1 with other factors and one of the triple interactions

No.	F_1F_2	F_1F_3	F_1F_4	F_1F_5	F_1F_6	$F_1F_2F_3$
4	+	–	–	–	–	–
7	–	–	+	+	+	–
10	–	–	+	–	–	+
13	+	–	–	+	+	+
18	+	–	–	+	–	+
21	–	–	+	–	+	+
27	+	+	–	–	+	+
32	–	–	+	+	–	+
35	+	+	+	+	–	+
40	–	+	–	–	+	+
41	+	+	–	+	–	–
46	+	+	+	–	+	–
49	+	+	+	–	–	–
54	–	+	–	–	+	–
60	+	–	+	+	+	–
63	–	–	–	+	–	–

It is easy to see that the columns of this matrix are mutually orthogonal. This circumstance allows to use the significantly simpler relation (9) instead of the general relation (3) when calculating the regression polynomial coefficients.

As a result of solving Problem 1, a standard procedure for transforming the original passive experimental design into an orthogonal design was implemented. The most important advantage of this design is that it provides an extremely simple computational procedure for estimating the values of the regression polynomial coefficients.

As a result of solving Problem 2, a representative orthogonal subdesign for identifying the regression polynomial was obtained.

In interpreting the results, it should be noted that the proposed method allows to solve the problem in the unique situation where the number of experiments is smaller than the number of coefficients being estimated.

A distinctive feature of the proposed methods is that their structural, analytical, and software implementation are not significantly dependent on the problem dimension, transforming this factor from one that has a significant influence to one of second-order importance. Another important advantage of the proposed methods is that as the problem dimension increases, the number of options for selecting a truncated subdesign does not decrease, but rather increases, increasing the potential effectiveness of the result.

Thus, the stated problem of developing a correct method for processing passive experiment results with a small sample of initial data has been solved.

The practical significance of the obtained results lies in the development of a universal computational technology for processing passive experiment results with a limited amount of initial data. The proposed approach enables the transformation of unstructured experimental observations into an orthogonal factorial design, followed by the extraction of a representative truncated subdesign suitable for independent estimation of regression polynomial coefficients.

This creates the possibility of quantitatively assessing the significance of factors and their interactions without conducting expensive and difficult-to-implement active experiments. The method can be used in diagnosing the condition of technical objects, analyzing the operating modes of complex systems, and predicting equipment condition. An additional advantage is the reduced computational complexity of calculations due to the use of orthogonal structures, which is especially important for multivariate problems with a large number of factors.

A distinctive feature of the obtained results is the following: the proposed approach consists of a combination of two different methods of multivariate regression analysis.

The use of the proposed techniques allows to solve a relevant regression analysis problem (statistical assessment of the state of semi-Markov systems) under conditions of severe insufficiency of initial data.

The proposed technology overcomes the shortcomings of known approaches to solving similar problems, described in [1–6].

Limitations of the research: The proposed method is focused on cases where the studied dependence can be adequately approximated by a finite-order regression polynomial.

The effectiveness of constructing a truncated orthogonal subdesign also depends on the completeness of coverage of the factor space by the initial observations of the passive experiment. If the experimental data are concentrated only in a limited region of the factor space, the representativeness of the resulting subdesign decreases. The direction of further research is to extend the obtained results to the case where the initial data of a passive experiment are given by fuzzy numbers with known membership functions.

4. Conclusions

1. A technology for transforming the results of a passive experiment into an orthogonal design for a full factorial experiment has been developed. It is implemented as follows. The variable space is scaled to the interval [–1, 1]. The initial space of the initial experiment is transformed into an orthogonal hypercube with 2^m vertices, where m is the number of problem factors. A truncated orthogonal representative subdesign is then extracted from this hypercube, used to calculate the coefficients of the regression polynomial. The proposed technology is universal and can be implemented uniformly regardless of the problem dimension, which determines its significant qualitative advantage.
2. A method for selecting a representative truncated orthogonal subdesign from a full factorial experiment design has been developed. This method ensures the determination of the significance level of a selected group of factors for assessing the state of the system. A significant qualitative advantage of the proposed method is the linear dependence of the complexity of the technology implementation on the problem dimension.

Conflicts of interest

The authors declare that they have no conflict of interest regarding this research, including financial, personal, authorship or other, that could influence the research and its results presented in this article.

Financing

The research was conducted without financial support.

Data availability

Data will be provided upon reasonable request.

Use of artificial intelligence

In preparing the manuscript, the authors used the artificial intelligence (AI) tool ChatGPT (OpenAI, GPT-5.3).

AI was used to search for articles on the topic of interest for the literature review in Section 1.

ISSN-L 2664-9969; E-ISSN 2706-5448

The authors reviewed the cited AI sources for relevance to the topic discussed in this article and for their accuracy.

The AI tools did not influence the research results, their interpretation, or the conclusions reached, and all key points in the manuscript were independently derived by the authors.

Authors' contributions

Lev Raskin: Conceptualization, Investigation, Writing – original draft, Writing – review and editing, Supervision; ***Oksana Sira***: Methodology; ***Larysa Sukhomlyn***: Methodology; ***Vitalii Vlasenko***: Formal analysis, Writing – original draft; ***Ihor Pryshchepa***: Formal analysis, Writing – original draft.

References

1. Myśliwiec, P., Kubit, A. (2025). Integrated multiobjective optimization of RFSSW parameters for AA2024-T3 using ANOVA machine learning and NSGA II. *Scientific Reports, 15 (1).* https://doi.org/10.1038/s41598-025-21941-3
2. Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., Safaraliev, M. (2025). Mathematical Methods in Feature Selection: A Review. *Mathematics, 13 (6),* 996. https://doi.org/10.3390/math13060996
3. Herren, A., Hahn, P. R. (2022). *Statistical Aspects of SHAP: Functional ANOVA for Model Interpretation.* arXiv:2208.09970v2. https://doi.org/10.48550/arXiv.2208.09970
4. Puliyanda, A., Li, Z., Prasad, V. (2022). Real-time monitoring of reaction mechanisms from spectroscopic data using hidden semi-Markov models for mode identification. *Journal of Process Control, 117,* 188–205. https://doi.org/10.1016/j.jprocont.2022.07.011
5. Tian, X., Wei, G., Wang, J. (2022). Target Location Method Based on Compressed Sensing in Hidden Semi Markov Model. *Electronics, 11 (11),* 1715. https://doi.org/10.3390/electronics11111715
6. Liao, Y., Xiang, Y., Wang, M. (2020). *Health Assessment and Prognostics Based on Higher Order Hidden Semi-Markov Models.* arXiv:2002.05272. https://doi.org/10.48550/arXiv.2002.05272
7. Leemis, L. M. (2023). *Statistical Modeling: Regression, Survival Analysis, and Time Series Analysis.* Open Educational Resource. https://doi.org/10.21220/SQQ8-A372
8. Jarantow, S. W., Pisors, E. D., Chiu, M. L. (2023). Introduction to the Use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays. *Current Protocols, 3 (6).* https://doi.org/10.1002/cpz1.801
9. Pallavi, Joshi, S., Singh, D., Kaur, M., Lee, H.-N. (2022). Comprehensive Review of Orthogonal Regression and Its Applications in Different Domains. *Archives of Computational Methods in Engineering, 29 (6),* 4027–4047. https://doi.org/10.1007/s11831-022-09728-5
10. Gauss, C. F. (1823). *Theoria combinations observationum erroribns minimis obnoxial.* H. Dieterich. Available at: https://archive.org/details/bub_gb_ZQ8OAAAAQAAJ/mode/2up
11. Janković, A., Chaudhary, G., Goia, F. (2025). Optimization through classical design of experiments (DOE): An investigation on the performance of different factorial designs for multi-objective optimization of complex systems. *Journal of Building Engineering, 102,* 111931. https://doi.org/10.1016/j.jobe.2025.111931
12. Hastie, T., Tibshirani, R., Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science, 35 (4),* 579–592. https://doi.org/10.1214/19-sts733
13. Cheng, J., Sun, J., Yao, K., Xu, M., Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268,* 120652. https://doi.org/10.1016/j.saa.2021.120652
14. Kullback, S., Leibler, R. (1951). On information and sufficiency. *The annals of Mathematical Statistics, 1,* 79–86. Available at: https://www.jstor.org/stable/2236703
15. Deldossi, L., Tommasi, C. (2021). Optimal design subsampling from Big Datasets. *Journal of Quality Technology, 54 (1),* 93–101. https://doi.org/10.1080/00224065.2021.1889418
16. Nguyen, N.-K., Stylianou, S., Pham, T.-D., Phuong Vuong, M. (2023). Designs for Screening Experiments with Quantitative Factors. *Novel Aspects of Gas Chromatography and Chemometrics.* IntechOpen https://doi.org/10.5772/intechopen.106805
17. Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2 (1-2),* 83–97. https://doi.org/10.1002/nav.3800020109

Lev Raskin, *Doctor of Technical Sciences, Professor, Department of Software Engineering and Management Intelligent Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0002-9015-4016*

Oksana Sira, *Doctor of Technical Sciences, Professor, Department of Computer Mathematics and Data Analysis, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0002-4869-2371*

Larysa Sukhomlyn, *PhD, Associate Professor, Department of Management and Marketing, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine, ORCID: https://orcid.org/0000-0001-9511-5932*

✉***Vitalii Vlasenko,*** *PhD Student, Department of Software Engineering and Management Intelligent Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, e-mail: vitalik.vlasenko.000@gmail.com, ORCID: https://orcid.org/0000-0001-5427-0223*

Ihor Pryshchepa, *PhD Student, Department of Software Engineering and Management Intelligent Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, ORCID: https://orcid.org/0009-0009-0143-1519*

✉*Corresponding author*