Polynomial Theory of Complex Systems

A. G. IVAKHNENKO

Reprinted by permission from IEEETRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS
Vol. SMC-1, No. 4, October 1971, pp. 364-378

Reprinted by permission from IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS
Vol. SMC-1, No. 4, October 1971, pp. 364-378

Polynomial Theory of Complex Systems

A. G. IVAKHNENKO

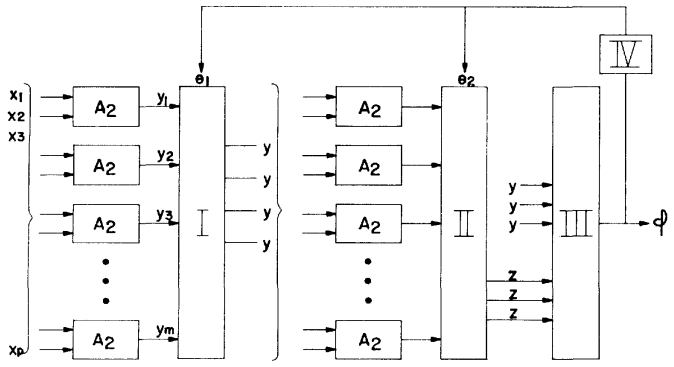
PAPER EDITOR'S PREFACE

The work of Academician A. G. Ivakhnenko, which appears regularly in *Avtomatika* (translated as *Soviet Automatic Control*), represents the approach of what might be called the "Kiev school" to the problems of modeling cybernetic systems. My colleague, R. L. Barron, first met Acad. Ivakhnenko in the Soviet Union in 1968 while attending technical conferences in that country and, at that time, invited him to submit a paper for this TRANSACTIONS to acquaint U.S. readers with his work. It is an impressive accomplishment that Acad. Ivakhnenko was able to write such a highly technical paper in a language that is not his native tongue. Parenthetical remarks have been inserted wherever the meaning might be unclear. These remarks are identified by the suffix "Ed." However, a diligent attempt was made to maintain the essence of Acad. Ivakhnenko's writing. The following comments are intended as an introduction to the paper.

Acad. Ivakhnenko models the input–output relationship of a complex system using a multilayered perceptron-type network structure. Each element in the network implements a nonlinear function of its inputs. The function implemented, termed an "algorithm" by him, is usually a second-order polynomial of the inputs. Since each element generally accepts two inputs, the function (algorithm) implemented by an element in one of the layers is

$$y = A_2(X) = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2$$

where the subscript in A_2 denotes a second-order transformation of the inputs. The following figure illustrates the structure of the overall input-output transformation.



Algorithm of group method of data handling (GMDH) using second-degree polynomials. I—first threshold self-selection; II—second threshold self-selection; III—selection from all solutions; IV—threshold optimization.

It can be seen that a number of "self-selection thresholds" are used. Their purpose is to filter out, at each layer, those elements which are least useful, "harmful" in Ivakhnenko's words, for predicting the correct output ϕ. Only those elements whose performance indices exceed the threshold at that layer are allowed to pass to the next layer. Therefore, the network representation is a feed-forward transformation, whereby each succeeding layer in the network increases by two the degree of the multinomial fit to the input properties x_i. Ivakhnenko has described the "selection hypothesis" used to select elements to be used in succeeding layers as the following.*

> . . . Our first conclusion is that composite control systems must be based on the use of signals which control the totality of elements of the system.
>
> The second conclusion . . . involves the extension to cybernetics of the principle of selection of plants and animals. The long history of the art of selection must be carefully studied by cyberneticists.
>
> . . . Let us examine the principle selection scheme. To obtain, say, plants (in the agricultural sense—Ed.) with certain properties, a large number of plants are sown which may have this property (the first heuristic being the selection of "elementary algorithms"); then the first crossing occurs (a generation of combinations of first complexity), and then the first harvest.
>
> From the harvest of the first generation (the first threshold self-selection), plants are chosen which better our requirements as compared to others; the seeds of these plants are again sown, and the plants are crossed (a generation of combinations of second complexity). From the second harvest one again selects certain seeds (the second threshold self-selection) which are sown, etc.
>
> The rules of the mass selection process are as follows.
>
> 1) For each generation there exists a certain optimal number of seeds being sown. Any change from this amount leads to slowing down and deterioration of selection.
>
> 2) The selection process cannot be completed in a single generation; several generations (at least three or four) are needed.
>
> If the selection process is too long (too many generations), the plants degenerate. The more complicated the selection problem, the more generations are needed.
>
> Now let us examine the perceptron algorithm. It fully duplicates all the main principles of selection. For example, a perceptron can be used for identification of extremal plants (in the control theory sense—Ed.). The complex surface of the extremal hump is approximated by polynomials. The signals applied to the perceptron input contain information about the surface of interest to us. The surface is usually described by a number of experimental points and simple functions of their coordinates ("elementary algorithms"). In accordance with the selection hypothesis, we first take simple polynomials of second degree that are easiest to inscribe in this surface. These simple combinations of data are subjected to the first threshold selection, and only some of the polynomials (namely, those which fit best into the sought surface) pass into the second layer, where they form more complex combinations (polynomials of fourth degree).
>
> From the second "generation," one again is singled out (by threshold selection) the most suitable, which can best be fitted into the sought surface. Here there is again an optimum number of combinations selected. Only these combinations pass into the third layer, where even more complex third "generation" combinations are formed (polynomials of sixth degree), etc. This constitutes the

Manuscript received May 28, 1970; revised June 22, 1971.
The author is with the Institute of Cybernetics, Ukrainian Academy of Sciences, Kiev, Ukraine, USSR.

* A. G. Ivakhnenko, Yu. V. Koppa, I. K. Tymchenko, and N. O. Ivakhnenko, "Group handling of data in identification of the static characteristic of a multi-extremal plant," *Sov. Automat. Contr.*, vol. 14, 1969, no. 2, pp. 30-37.

group method of data handling (GMDH—Ed.) for the problem under consideration.

In deterministic problems with "smooth" solution functions and a small number of measurements, the best solution is obtained at once after the first selection, in the first "generation" of combinations. The greater the data spread and the more complex and higher dimensional the functions, the more "generations" are needed for obtaining optimum results.

In complex problems, just as in selection of plants, at least three or four "generations" are needed for obtaining satisfactory results ("the perceptron must have at least four layers"). If the perceptron has too many layers, the combinations degenerate. This is indicated by an accuracy check (a correlation criterion or mean-square-error criterion) (on the independent testing data set—Ed.). To preclude degeneration, the best solution must be chosen based not on the results of the last layer, but on data from all layers of the perceptron. Hence all aspects of the selection process mentioned above occur in the perceptron.

The six coefficients of each element in the network are determined in the following way. Consider one element in the first layer. It implements the function $A_2(X)$ shown in the preceding. Assume that there are N input vectors in the training set, each composed of p property values, $X_n = (x_{n1}, \cdots, x_{np})$ $n = 1, \bullet\bullet\bullet, N$. Denote the nth desired output as ϕ_n. A set of six coefficients for this element (which has inputs x_{ni} and x_{nj}) must be found such that the mean-square error between the outputs of this element y_n, and the true outputs ϕ_n is minimized.

The coefficients are obtained from the "Gauss normal equations" as follows. First write the system of equations

$$\begin{aligned} \phi_1 &= a_0 + a_1 x_{1i} + a_2 x_{1j} + a_3 x_{1i}^2 + a_4 x_{1j}^2 + a_5 x_{1i} x_{1j} \\ &\vdots \\ \phi_N &= a_0 + a_1 x_{Ni} + a_2 x_{Nj} + a_3 x_{Ni}^2 + a_4 x_{Nj}^2 + a_5 x_{Ni} x_{Nj} \end{aligned}$$

in matrix form

$$\Phi = XA$$

where matrices Φ, X, and A are of order N x 1, N x 6, and 6 x 1, respectively. (The first element in each row of the X matrix is unity.) The normal equations are formed by premultiplying both sides by the transpose of X:

$$X^T\Phi = (X^TX)A.$$

Matrix X^TX is a 6 x 6 matrix, and the solution is found by inverting this matrix:

$$A = (X^TX)^{-1}X^T\Phi.$$

Matrix A contains the set of six coefficients which enables this element to approximate the true outputs with minimum mean-square error. This procedure is repeated for each element in the first layer, with the components in matrix X changing each time depending on the identity of the two inputs to the particular element. The same technique is used to find the six coefficients of each element in succeeding layers. After the coefficient values are computed based on the training data, the performance index of a given element (and, hence, of its two input properties) is determined by either correlating or computing the mean-square error between the output of the element and the true output for each vector in the independent testing set. Only those elements whose performance index exceeds the "self-selection threshold" for that layer are allowed to be used in the next layer. This use of the independent testing set to filter out poor properties is what Ivakhnenko calls "decision regularization."

The editor of this paper is joined by Mr. Barron and J. N. Warfield, Co-Editor of this TRANSACTIONS, in expressing gratitude to Acad. Ivakhnenko for the enjoyable correspondence which has taken place throughout the revision procedure.

ANTHONY N. MUCCIARDI
Adaptronics, Inc.
McLean, Va. 22101

***Abstract* (written by paper editor)—A complex multidimensional decision hypersurface can be approximated by a set of polynomials in the input signals (properties) which contain information about the hypersurface of interest. The hypersurface is usually described by a number of experimental (vector) points and simple functions of their coordinates. The approach taken in this paper to approximating the decision hypersurface, and hence the input–output relationship of a complex system, is to fit a high-degree multinomial to the input properties using a multilayered perceptronlike network structure. Thresholds are employed at each layer in the network to identify those polynomials which best fit into the desired hypersurface. Only the best combinations of the input properties are allowed to pass to succeeding layers, where more complex combinations are formed.**

Each element in each layer in the network implements a nonlinear function of two inputs. The coefficients of each element are determined by a regression technique which enables each element to approximate the true outputs with minimum mean-square error. The experimental data base is divided into a training and testing set. The training set is used to obtain the element coefficients, and the testing set is used to determine the utility of a given element in the network and to control overfitting of the experimental data. This latter feature is termed "decision regularization."

In contrast to the statistical decision theoretic approach which is "single layered," it is argued that the type of multilayered structure presented should be used to solve complex problems for four primary reasons: 1) a smaller training set of data is required; 2) the computational burden is reduced; 3) the procedure automatically filters out input properties which provide little information about the location and shape of the decision hypersurface; and 4) a multilayered structure is a computationally feasible way to implement multinomials of very high degree.

A network-implemented model of the British economy and results forecasted by the model are presented to demonstrate the utility of the polynomial theory.

I. INTRODUCTION

MODERN control theory, based on differential equations, is not an adequate tool for solving the problems of complex control systems. It is necessary to construct differential equations to trace the input-output paths, that is, to apply a deductive deterministic approach. But it is impossible to use this approach for complex systems because of the difficulty in finding these paths. The method of heuristic self-organization [1], [2] is more appropriate for the problem.

The starting point of the heuristic self-organization approach can be formulated in the following way. I know that I know nothing; let us generate and compare all possible input-output combinations. Self-organization uses mathematical combinatorial methods and a series of "threshold self-selections" on the different heuristic criteria [2].

"Examinations" of the threshold type are widely used in the selection of botanical plants. To obtain plants with certain desired characteristics, a number of generations ("seeds") of those plants are selected in which the properties (of interest—Ed.) are more pronounced. Our *hypothesis of selection* states that algorithms of selection are optimal for solving the interpolation problems of prediction, pattern recognition, identification, or optimal control of complex physical plants by any form of (commonly—Ed.) used basic functions. The only demand is as follows: the complexity of mathematical description relative to the inputs must increase with every new "generation" of formulas.

Using polynomials as the basic (plant describing—Ed.) functions, we obtain the polynomial theory; using Bayes' formulas, we obtain the probabilistic theory, etc. Each type of basic function gives us a different theory of complex systems.

II. POLYNOMIAL DESCRIPTIONS: THE BASIC MEANS OF THE NEW THEORY

The formulas of a predictive model have been previously used only for calculating the future values of the predicted variables. The polynomial theory proposes to use predictive polynomials as the basic means for the general investigation of complex dynamic systems [1].

The predictive polynomial is a regression equation which connects a future value of the output variable with past or running (current—Ed.) values of all input and output variables. Regression analysis allows us to evaluate the coefficients of the polynomial by the criterion of minimum mean-square error.

A more general mathematical theory to synthesize an optimizing decision algorithm using the information given in a few points of interpolation is not yet available [3]. So, the polynomial theory realizes the "hypothesis of selection": the polynomials are treated in the same manner as are seeds in agricultural selection. This enables us to obtain the polynomial description of a component or of the whole complex plant by observing their inputs and outputs during a comparatively short time.

The main problem solved by the polynomial theory is to find the polynomial description of optimum complexity, i.e., that which is adequate to describe the complexity of the plant. Only such descriptions can give high prediction accuracies.

There is an analogy between the polynomial description in the new theory and the operator transfer function in control theory; neither conception requires information about initial conditions. (Although information about the initial conditions is required to obtain a series solution to a finite difference or differential equation, it is possible to construct a polynomial description from a data base of input and output values without explicit initial conditions—Ed.)

The polynomial descriptions have some advantages for engineers. For example, there is no need to find solutions for the equations in finite difference form because the answers to all the interesting questions can be found from the polynomial description itself. So, the information about initial conditions and the solution of the equations are often not necessary.

No distinction is made between the statics and the dynamics of the plant in the polynomial theory. This division of the plant regimes is connected with the application of differential equations, and consequently, it is quite artificial.

There are at least three methods by which we can obtain the polynomial description of a component or of the whole control system:

1) by replacing the derivatives in the differential equations by finite differences;
2) by methods of indirect measurement of the disturbances by the so-called "differential fork;" these methods were taken from theory of invariance [5];[1]
3) by methods of active or passive experimentation.

The aforementioned regression analysis is a typical example of such methods. But many other methods of identification are known.

There is no reason to use identification methods for estimating the coefficients of the differential equations. It is much simpler to use them directly for the synthesis of polynomial descriptions. Less information is necessary for this operation. In particular, there is no need to know the type of differential equations.

The first two methods are used for deterministic problems. The experimental method is used also for stochastic problems where the differential equations are not known absolutely or are not applicable. Here, the polynomial theory has no rivals because only this theory allows us to obtain the *optimum complexity* mathematical model of the plant. Examples show that the most accurate description of extremely complex plants corresponds to polynomial descriptions of a rather high degree (e.g., to polynomials of up to the sixty-fourth degree). It is impossible to find nonlinear differential equations corresponding to such complex polynomials.

III. APPLICATION OF THE POLYNOMIAL THEORY TO DETERMINISTIC SYSTEMS

The advantages of the polynomial theory for deterministic systems are not obvious but nevertheless are essential. The theory enables us to find the optimization equation, thereby making structure transformations for nonlinear systems much simpler.

It is easy to deduce the polynomial descriptions for any differential equation by replacing the derivatives with finite differences. Using "lag differences" we obtain the *predicting equation*, and using "lead differences" we obtain the *control equation*. The control equation includes future values of the input variable among its arguments, but the predicting equation does not. The predicting equation is used to predict future values of the output variable for the conditions when the future values of any input are not known. The control equation can include these values because we consider them the manipulated variables used to control the plant.

The predicting equation for input x and output y is

$$y_F = f_1(y_{F-1}, y_{F-2}, y_{F-3}, \cdots, x_{F-1}, x_{F-2}, x_{F-3}, \cdots). \quad (1)$$

[1] The theory of invariance is a part of general control theory, considering conditions of complete compensation of disturbances by proper change of manipulated variables. An indirect measurement of variables by the "differential fork" is often used in this theory.

The control equation is

$$y_F = f_2(y_{F-1}, y_{F-2}, y_{F-3}, \cdots, x_F, x_{F-1}, x_{F-2}, x_{F-3}, \cdots). \quad (2)$$

The time instances are denoted as

F	future moment of time
$F-1$	instant moment
$F-2$	moment one step before
$F-3$	moment two steps before

and so on.

The optimization equation in its simplest form is

$$\frac{\partial y_F}{\partial x_F} = 0 \quad (3)$$

where the y_F are the future values of the index of extremum (or performance index), and x_F is the future value of the manipulated variable. In this case the control equation is the equation of the complete extremum system.

In complex systems the variable y_F is connected with the variable x_F by nonlinear polynomials which enable us to apply the technique of mathematical programming and gradient methods to solve the problem of optimal control. When the regression analysis or group method of data handling (GMDH) is used to obtain the equations, we have the "algorithm of optimal control with storage of information."

IV. RULES OF STRUCTURE TRANSFORMATION

The system control equation can be obtained from its component equations using the following rules [1].

1) Synthesize the polynomial description for the serial coupling of the components by the joint solution of the polynomial descriptions of all given components excluding the intermediate variables.

2) Synthesize the polynomial description for the parallel coupling of the components by the summation of the coefficients of the polynomial description members of each component having the same indices of time.

3) Synthesize the polynomial description for the component with feedback by subtracting the coefficients of the x_{F-j} members from the coefficients of the y_{F-j} members having the same indices of time.

Therefore, the structure transformation can be made as simple in the polynomial theory as in the common control theory.

Example

Suppose there are two components of first order:

$$(\tau_1 p + 1)z = \alpha y, \qquad (\tau_2 p + 1)y = \beta x, \qquad p = \frac{d}{dt}.$$

The polynomial descriptions (for six instants of time) are

$$z_F = c_1^5 z_{F-5} + \alpha y_{F-1} + \alpha y_{F-2} + \alpha c_1^2 y_{F-3} + \alpha c_1^3 y_{F-4} + \alpha c_1^5 y_{F-5}$$

and

$$y_F = c_2^5 y_{F-5} + \beta x_{F-1} + \beta c_2 x_{F-2} + \beta c_2^2 x_{F-3} + \beta c_2^3 x_{F-4} + \beta c_2^4 x_{F-5}$$

where

$$\alpha = \frac{h}{\tau_1}, \qquad \beta = \frac{h}{\tau_2}, \qquad c_1 = 1 - \frac{h}{\tau_1}, \qquad c_2 = 1 - \frac{h}{\tau_2}.$$

By series connection of the components we obtain

$$z_F = c_1^5 z_{F-5} + (\alpha\beta c_1^3 + \alpha\beta c_1^2 c_2 + \alpha\beta c_1 c_2^2 + \alpha\beta c_2^3) \cdot x_{F-5} + (\alpha\beta c_1^2 + \alpha\beta c_1 c_2 + \alpha\beta c_2^2) x_{F-4} + (\alpha\beta c_1 + \alpha\beta c_2) x_{F-3} + 2\beta x_{F-2}.$$

By parallel connection of the components we obtain

$$y_F = (c_1^5 + c_2^5) y_{F-5} + (\alpha + \beta) x_{F-1} + (\alpha c_1 + \beta c_2) \cdot x_{F-2} + (\alpha c_1^2 + \beta c_1^2) x_{F-3} + (\alpha c_1^3 + \beta c_2^3) \cdot x_{F-4} + (\alpha c_1^4 + \beta c_1^4) x_{F-5}.$$

And, finally, when the second component has a feedback, we obtain

$$y_F = (c_2^5 - c_2^4\beta) y_{F-5} + \beta x_{F-1} + c_2\beta x_{F-2} + c_2^2\beta c_{F-3} + c_2^3\beta x_{F-4} + c_2^4\beta x_{F-5}.$$

V. THE STABILITY CRITERION

For the analysis of stability the polynomial description of the system has to be transformed (by the so-called equivalent transformations [1]) to the form containing no intermediate variables. The left-hand side of this type of polynomial equation after linearization and substitution

$$y_F = z; \; y_{F-1} = z^2; \; \cdots; \; y_{F-n} = z^n \quad (4)$$

gives us the characteristic equation

$$a_n z^n + \cdots + a_2 z^2 + a_1 z + a_0 = 0. \quad (5)$$

For stability the roots of this equation have to lie in a circle of unit radius.

So, the well-known criteria for impulse systems stability is valid in the polynomial theory also. This is quite understandable because the analysis of stability, as stated by Lyapunov, applies to the level of linear equations where the polynomial theory seems to have no considerable advantages. Its advantages are connected with complex optimal nonlinear descriptions.

VI. THE CONDITIONS OF INVARIANCE

To obtain compensation in the polynomial description of the stabilization system, all the coefficients must be equal to zero. We can satisfy this condition only in the case where all the coefficients are the differences of two values, compensating one another. This case occurs in systems with feedforward links and in multifeedback systems with the "differential forks" [5] for intermediate measurement of the disturbances.

Example

The polynomial descriptions are obtained from the given differential equations and their transformations. The given differential equation is

$$\tau \frac{dy}{dt} + y = x. \tag{6}$$

Obtain the predicting equation using the lag finite difference:

$$\frac{dy}{dt} \cong \frac{\Delta y}{h} = \frac{1}{h}[y(t) - y(t - h)]. \tag{7}$$

Obtain the control equation using the lead finite difference:

$$\frac{dy}{dt} \cong \frac{\Delta y}{h} = \frac{1}{h}[y(t + h) - y(t)]. \tag{8}$$

The *recurrent transform* leads to increasing the interval of time taken into account: $T = Nh$, where N is the number of prehistory steps.

For the predicting equation we obtain, when $a = 1 - b$, $b = h/\tau$:

$$\begin{aligned}
N = 0: &\quad y_F = ay_{F-1} + bx_{F-1} \\
N = 1: &\quad y_F = a^2y_{F-2} + bx_{F-1} + abx_{F-2} \\
N = 2: &\quad y_F = a^3y_{F-3} + bx_{F-1} + abx_{F-2} + a^2bx_{F-3} \\
N = 3: &\quad y_F = a^4y_{F-4} + bx_{F-1} + abx_{F-2} + a^2bx_{F-3} \\
&\qquad + a^3bx_{F-4}
\end{aligned} \tag{9}$$

and so on.

For the control equation we obtain, when $a = \tau/(\tau + h)$, $b = h/(\tau + h)$:

$$\begin{aligned}
N = 0: &\quad y_F = ay_{F-1} + bx_F \\
N = 1: &\quad y_F = a^2y_{F-2} + bx_F + abx_{F-1} \\
N = 2: &\quad y_F = a^3y_{F-3} + bx_F + abx_{F-1} + a^2bx_{F-2} \\
N = 3: &\quad y_F = a^4y_{F-4} + bx_F + abx_{F-1} + a^2bx_{F-2} \\
&\qquad + a^3bx_{F-3}
\end{aligned} \tag{10}$$

etc.

The *equivalent transform* is the transformation of the argument set without any change in the time interval of the prehistory T, by the following formulas:

$$x_{F-1} = \frac{1}{b}y_F - \frac{a}{b}y_{F-1}$$

$$x_{F-2} = \frac{1}{b}y_{F-1} - \frac{a}{b}y_{F-2}$$

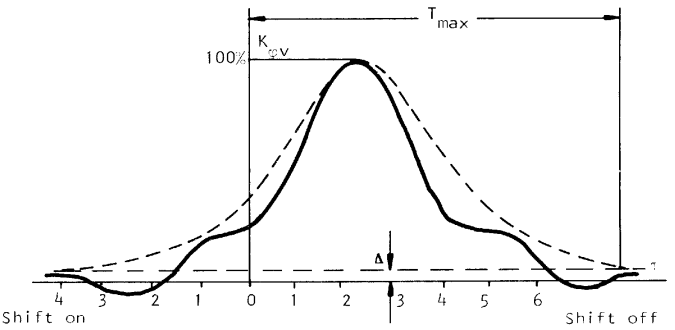


Fig. 1. Evaluation of time interval $T = Nh$. T—interval of prehistory; T_{max}—maximum value of T; $K_{\phi v}$—correlation function between output ϕ and input variable v; Δ—error; τ—time of shift; h—time shift incremental step; N—number of steps.

$$x_{F-3} = \frac{1}{b}y_{F-2} - \frac{a}{b}y_{F-3}$$

$$x_{F-4} = \frac{1}{b}y_{F-3} - \frac{a}{b}y_{F-4} \tag{11}$$

and so forth.

VII. APPLICATION OF THE POLYNOMIAL THEORY TO STOCHASTIC SYSTEMS

The interval of prehistory $T - Nh$, which we are to take into account, can be estimated from the correlation functions. In a deterministic system this interval is no shorter than the continuation of the dynamic transient processes in the given system. In a stochastic system the interval T is to be no shorter than the displacement time by which the correlation function is essentially different from zero (Fig. 1).

The error associated with the limited number of steps A used is exactly equal to the error associated with the indetermination of the termination of the transient process. Therefore, the accuracy of the polynomial theory is no less than the accuracy obtained based on the solution of the differential equations.

Example

The calculation of the correlation function is given in this example. The departures of variables ϕ_0 and v_1 from their mean values are presented in the following.

$i, j =$	1	2	3	4	⋯	$n-1$	n
$\phi =$	−0.01	0.02	0.01	−0.10	⋯	0.02	0.03
$v =$	0.06	0.05	0.02	−0.03	⋯	0.01	0.10

Indices of variables ϕ and v are $i = 1$ to $i = n$ and $j = 1$ to $j = n$, respectively. For the shift $\tau = 0$ the correlation coefficients

$$K_{\phi v}(0) = \frac{[(-0.01)(0.06) + (0.02)(0.05) + (0.01)(0.02) + \cdots + (0.03)(0.10)]}{\{[(-0.01)^2 + (0.02)^2 + \cdots + (0.03)^2][(0.06)^2 + (0.05)^2 + \cdots + (0.10)]\}^{1/2}}$$

$$K_{\phi v}(1) = \frac{[(-0.01)(0.05) + (0.02)(0.02) + (0.01)(-0.03) + \cdots + (0.02)(0.10)]}{\{[(-0.01)^2 + (0.02)^2 + \cdots + (0.02)^2][(0.05)^2 + (0.06)^2 + \cdots + (0.10)^2]\}^{1/2}}$$

and so on. The general formula

$$K_{\phi v}(\tau) = \frac{\sum_{i=1,j=1+\tau}^{i=n-\tau,j=n} \phi_i v_j}{\left\{ \sum_{i=1}^{i=n-\tau} \phi_i^2 \sum_{j=1+\tau}^{j=n} v_j^2 \right\}^{1/2}}$$

makes it possible to calculate the discrete correlation function for any shift $\tau = 1,2,3,\cdots,n$.

The GMDH is the best method for obtaining the polynomial description of a stochastic system from a small amount of experimental data. This algorithm follows.

VIII. THE MAIN RULE FOR CONSTRUCTING THE GMDH ALGORITHMS

The "complete" description of the plant

$$\phi = f_1(x_1,x_2,x_3,\cdots,x_n) \tag{12}$$

is to be constructed of several layers of "partial" descriptions:

$$y_1 = f_1(x_1,x_2),\ y_2 = f_1(x_3,x_4),\cdots,\ y_m = f_1(x_{n-1},x_n), \qquad m = c_n^2$$

$$z_1 = f_1(y_1,y_2),\ z_2 = f_1(y_3,y_4),\cdots,\ z_p = f_1(y_{m-1},y_m), \qquad p = c_m^2 \tag{13}$$

and so forth, where c_n^2 and c_m^2 are the number of pairwise combinations of n and m for the first and second layers, respectively.

The function f_1 is usually the same in all the equations. Eliminating the intermediate variables, we obtain the "analog" of the complete description. Comparing the analog and the actual complete description in its general form, we find the equations for the construction of the coefficients of the complete description.

In the polynomial theory the functions f_1 are the Kolmogorov–Gabor polynomials:

$$\phi = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_i a_j a_k x_i x_j x_k + \cdots.$$

IX. RULES FOR VARIABLE SELECTION

All experimental data are divided approximately in half into training and checking (testing—Ed.) sets. Coefficients of the first layer of partial descriptions are calculated by solving small systems of Gaussian normal simultaneous equations according to the minimum mean-square-error criterion. The left-hand sides of the equations are set equal to values of the output function at every point. After finding the values of the coefficients, we calculate the values of the intermediate variables. Then, using the data of the checking set, the mean-square error is determined for each variable. Only the most accurate variables, i.e., those giving the least error, are selected for subsequent use. These variables are retained in the training and checking sets, and the other variables are discarded. In the second layer of selection coefficients of partial descriptions of the layer are calculated, and the accuracy is checked again to select the most accurate intermediate variables of the layer: $z_1,z_2,\cdots,z_p$, etc.

The GMDH algorithm has a perceptron-like multilayer structure. Only the variables whose prediction accuracies in a given layer are larger than certain thresholds are optimized with respect to the next layer. The threshold values are optimized with respect to the criterion of minimum square output error also, which represents the third application of this criterion.

The accuracy of the GMDH is often extremely good due to the following considerations. The ordinary regression method of minimum mean-square error yields the optimum value only in the small domain where the number of the regression polynomial members is less than the number of points of interpolation. Optimization by the GMDH includes not only this domain but all other relationships between these two sets of numbers. It is possible to obtain the *optimal* complete descriptions of any complexity with respect to the mean-square error based on decisions on the separate testing data set. This much broader domain of optimization ensures us much higher accuracy.

X. THE RULE FOR ACCURACY CONTROL

Every intermediate variable is examined for its effect on the prediction accuracy. The main rule is to use the separate testing set for accuracy control. If the training data (which were used to estimate the coefficients) are used, incorrect solutions are obtained, since small changes in the training data will lead to large changes in the coefficient values (e.g., overfitting—Ed.). This is the reason for dividing the data into separate training (learning) and testing (checking) sets. The training set is used to calculate the coefficients of the partial descriptions, but the testing set is used to evaluate the quality of the partial descriptions. This is the basic method used for the so-called *regularization of the decisions*.

XI. THREE BASIC ADVANTAGES OF THE GMDH

The basic advantages of the GMDH are as follows.

1) There are insufficient data to estimate the coefficients of the Gaussian normal equations (by the method of least squares—Ed.) if the complete polynomial is used directly without the partial polynomials.

2) The matrices of the complete polynomial equations are always ill conditioned. But we can always select well-conditioned matrices among the diversity of the small partial equation matrices.

3) When the training set is small some arguments and intermediate variables are "harmful." That is, the accuracy will increase if we eliminate them. This important task is fulfilled by the thresholds on the prediction accuracy test after every layer of selection.

XII. EXAMPLE OF GMDH ALGORITHMS (DECISION RULES)

The GMDH can be realized by many algorithms which differ with respect to the basic functions f_1 used for the construction of the partial and complete descriptions. About

20 algorithms have been proposed up to the present. Some of them (the algorithm using probabilistic graphs, the algorithm using Bayes' formulas, and algorithms using second-order polynomials) are considered in [6].

Consider, as an example, a very simple algorithm of the GMDH for the case of four arguments. The four arguments are binary taking on the values -1 and $+1$. The complete polynomial has $2^4 = 16$ terms and is

$$\begin{aligned} z = {} & a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_1x_2 \\ & + a_6x_1x_3 + a_7x_1x_4 + a_8x_2x_3 + a_9x_2x_4 \\ & + a_{10}x_3x_4 + a_{11}x_1x_2x_3 + a_{12}x_1x_2x_4 \\ & + a_{13}x_1x_3x_4 + a_{14}x_2x_3x_4 + a_{15}x_1x_2x_3x_4. \end{aligned} \quad (14)$$

The partial polynomials for the pairwise combination of x_1,x_2 and x_3,x_4 are

$$\begin{aligned} y_1 &= b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 \\ y_2 &= c_0 + c_1x_3 + c_2x_4 + c_3x_3x_4 \\ z &= d_0 + d_1y_1 + d_2y_2 + d_3y_1y_2. \end{aligned} \quad (15)$$

The coefficients of the partial polynomials may be found by solving the Gaussian normal equations. Let us recall the basic rules of the minimum mean-square-error method. We obtain the system of conditional Gauss equations by writing every partial description a number of times equal to the number of points in the training set. For five points we shall obtain the system containing five equations. To obtain the A'th normal equation, every conditional equation is multiplied by the coefficient of b_k, and all the equations are added. This results in a system containing four normal equations, each with four unknown variables b_k, where $k = 0,1,2,3$.

When calculating the coefficients we assume $y_1 = y_2 = z$ and then determine the variables y_1 and y_2 for every interpolation point, which are then used in the third polynomial. The other combinations are x_1,x_4 and x_2,x_3 or x_1,x_3 and x_2,x_4. We can choose any combination which gives the best accuracy.

The prediction accuracy is to be checked on the separate testing set of data. The most accurate two variables from $y_1(x_1,x_2)$, $y_2(x_3,x_4)$, $y_3(x_1,x_4)$, $y_4(x_2,x_4)$, $y_5(x_1,x_3)$, and $y_6(x_2,x_4)$ are to be used as the arguments for z.

Using the first two combination of equations shown in (15), the coefficients of the complete polynomial can obviously be constructed by the following formulas:

$$\begin{aligned} a_0 &= d_0 + d_1b_0 + d_2c_0 + d_3b_0c_0 \\ a_1 &= d_1b_1 + d_3b_1c_0 \\ a_2 &= d_1b_2 + d_3b_2c_0 \\ a_3 &= d_2c_1 + d_3b_0c_1 \\ a_4 &= d_2c_2 + d_3b_0c_2 \\ a_5 &= d_1b_3 + d_3b_3c_0 \\ a_6 &= d_2c_3 + d_3b_0c_3 \\ a_7 &= d_3b_1c_1 \\ a_8 &= d_3b_1c_2 \\ a_9 &= d_3b_2c_1 \\ a_{10} &= d_3b_2c_2 \\ a_{11} &= d_3b_3c_1 \\ a_{12} &= d_3b_3c_2 \\ a_{13} &= d_3b_1c_3 \\ a_{14} &= d_3b_2c_3 \\ a_{15} &= d_3b_3c_3. \end{aligned} \quad (16)$$

It is easy to find similar formulas from the other two combinations of arguments if they prove to be more accurate than those shown in the preceding. Note that every term of the complete polynomial appears in (16). However, this does not mean that there are no additional limitations on the choice of coefficients. Using the complete polynomial does give more degrees of freedom when we attempt to minimize the mean-square error, but this is only one of several optimizations. Another algorithm of the GMDH (with second-order polynomials) follows.

XIII. THE POLYNOMIAL DESCRIPTION OF THE BRITISH ECONOMY

Let us obtain a polynomial description of the British economy. The set of characteristic variables (e.g., features, properties—Ed.) (or the characteristic vector) is to be established by experts.[2] They often can construct the credit and debit balances of energy, matter, or money for the given time unit. These balances are a great aid in selecting the characteristic vector. The polynomial descriptions constructed in this manner are called "after-balance models." This principle of model construction unites the deterministic balance with the self-organization approach. Let us consider an example of such an after-balance model.

The British scientists Parks and Pyatt [4] have established the characteristic vector for the British economy. There are 21 variables in this economic system from which 16 are interconnected by differential or algebraic equations. So, we can choose any 5 variables to be manipulated. We have chosen the following manipulated variables [1]:

$V_1 = T_H$—taxes on households; $V_{1(1968)} = 4559$;
$V_2 = T_c$—taxes on companies; $V_{2(1968)} = 1085$;
$V_3 = u$ —transfers from government (to unemployed people, for example); $V_{3(1968)} = 1520$;
$V_4 = K$ —total gross capital stock; $V_{4(1968)} = 146.7 \times 10^3$;
$V_5 = g$ —government expenditures; $V_{5(1968)} = 7702$.

As the performance indices, we choose the two following variables:

$$\Phi_1 = \frac{w + \Pi}{c_4K}$$

[2] The second way to find the characteristic vector is via a random search on the criterion of accuracy.

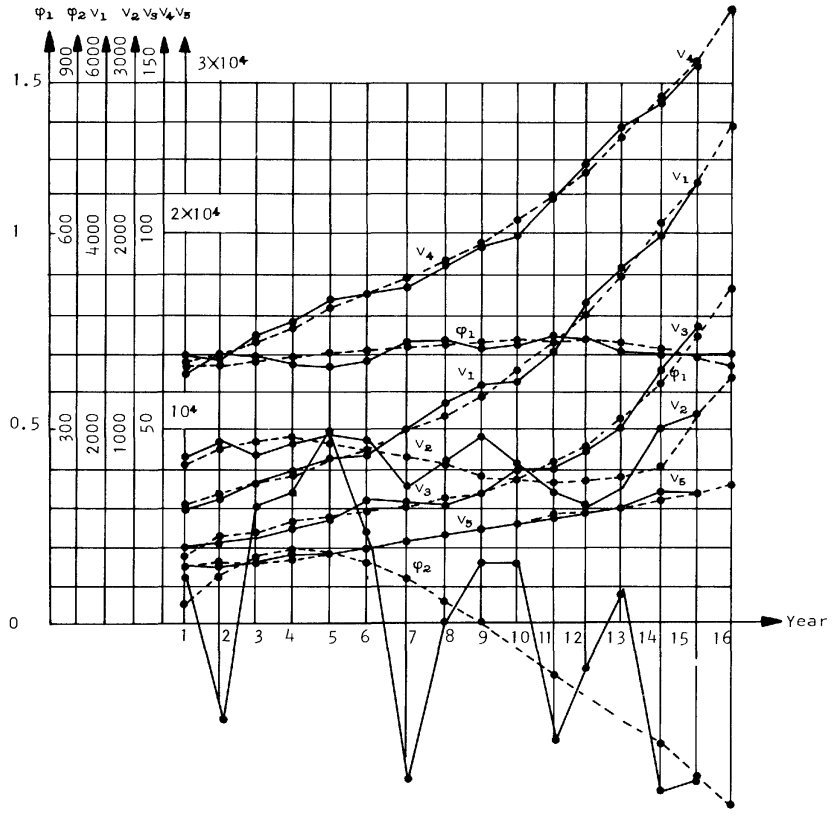


Fig. 2. Input data: variables and their trends.

where w is the income from employment, Π is the gross trading surplus, and $c_4 = \frac{1}{3}$ (for example, $\Phi_{1(1968)} = 0.71$), and

$$\Phi_2 = I_{Fg} + I_{Fc}$$

where I_{Fg} is the net investment abroad from government and I_{Fc} is the net investment abroad from companies (e.g., $\Phi_{2(1968)} = -265$). The economic situation is good if $\Phi_1 \to 1$ and $\Phi_2 \to 0$, and it is bad if $\Phi_1 \to 0$ and $\Phi_2 \to -\infty$. The input data are shown in Fig. 2 and Table I. Here we can see how all seven variables changed for the last 15 years (1954–1968). The problem is to find the polynomial description of the British economy for the future year 1969 and, therefore, to find the optimal values of the manipulated variables for this year. For every year thereafter the description changes according to the new data.

Solution of the Problem

1) Calculation of the Nonlinear Trends: The trends are calculated by ordinary regression analysis as third-degree polynomials. The results are shown in Fig. 2 by dotted lines. Instead of absolute value, we use the normalized deviations of all variables measured from their trends:

$$v_1 = \frac{V_1 - V_1^*}{V_1^*}$$

$$v_2 = \frac{V_2 - V_2^*}{V_2^*}$$

$$v_3 = \frac{V_3 - V_3^*}{V_3^*}$$

$$v_4 = \frac{V_4 - V_4^*}{V_4^*}$$

$$v_5 = \frac{V_5 - V_5^*}{V_5^*}$$

$$\phi_1 = \frac{\Phi_1 - \Phi_1^*}{\Phi_2^*}$$

$$\phi_2 = \frac{\Phi_2 - \Phi_2^*}{\Phi_2^*}.$$

We shall consider these deviations as mutual dependent stationary processes of time.

Example: We have 15 points shown in Fig. 2 and Table I. Choose the cubic polynomial

$$\Phi_1^* = a_0 + a_1 t + a_2 t^2 + a_3 t^3.$$

Writing it 15 times and inserting the data from Fig. 2 we obtain the system of conditional equations. Using the averaging rule we obtain the system of four normal Gaussian equations with four unknown variables:

$$9.309 = 15a_0 + 120a_1 + 1240a_2 + 14\,400a_3$$

$$81.164 = 120a_0 + 1240a_1 + 14\,400a_2 + 178\,312a_3$$

$$866.740 = 1240a_0 + 14\,400a_1 + 178\,312a_2 + 2\,299\,200a_3$$

$$10\,254.801 = 14\,400a_0 + 178\,312a_1 + 2\,299\,200a_2 + 30\,482\,920a_3.$$

TABLE I
INPUT DATA FOR 15 YEARS

	Variables	Years							
		1 (1954)	2 (1955)	3 (1956)	4 (1957)	5 (1958)	6 (1959)	7 (1960)	8 (1961)
Absolute values	$\Phi_1 = 1 - \delta$	0.689	0.698	0.697	0.692	0.674	0.697	0.795	0.736
	$\Phi_2 = I_{FG} + I_{FC}$	117	−155	208	233	344	143	−265	−4
	$V_1 = T$	1236	1330	1452	1602	1696	1776	1991	2249
	$V_2 = T_C$	893	972	894	972	1010	968	715	813
	$V_3 = U$	412	441	464	512	548	658	656	640
	$V_4 = K$	64.88×10^3	69.72×10^3	74.8×10^3	79.17×10^3	83.421×10^3	85.56×10^3	88.35×10^3	93.385×10^3
	$V_5 = G$	3185	3251	3515	3668	3750	4001	4248	4589
Trends	Φ_1^*	0.6851	0.6871	0.6909	0.6965	0.7033	0.7107	0.7181	0.7249
	Φ_2^*	43.972	93.9577	123.2492	134.3481	129.7768	112.0637	83.7372	47.3257
	V_1^*	1232.3352	1944.4569	1455.2951	1570.0491	1693.9152	1832.0917	1989.7769	2172.1671
	V_2^*	873.3798	944.5949	977.8422	980.4375	959.6966	922.935	877.4694	130.6147
	V_3^*	973.8002	450.7397	505.0590	544.1065	573.2306	597.7797	623.1022	654.5465
	V_4^*	65 593	70 033	74 138	78 044	81 892	85 819	89 966	94 471
	V_5^*	3171	3309	3459	3625	3812	4023	4264	4538
Normalized deviations from trends	$\phi_1 = \frac{\Phi_1 - \Phi_1^*}{\Phi_1^*}$	0.0048	0.0158	0.0088	−0.0064	−0.0416	−0.0192	0.0235	0.0153
	$\phi_2 = \frac{\Phi_2 - \Phi_2^*}{\Phi_2^*}$	1.6634	−2.6497	0.6876	0.7343	1.6507	0.2760	−4.1646	−1.0845
	$v_1 = \frac{V_1 - V_1^*}{V_1^*}$	0.0029	−0.0107	−0.0022	0.0203	0.0012	−0.0306	0.0006	0.0353
	$v_2 = \frac{V_2 - V_2^*}{V_2^*}$	0.0224	0.0290	−0.0857	−0.0086	0.0524	0.0488	−0.1851	−0.0212
	$v_3 = \frac{V_3 - V_3^*}{V_3^*}$	0.0963	−0.0216	−0.0812	−0.0590	−0.0440	0.1007	0.0527	−0.0222
	$v_4 = \frac{V_4 - V_4^*}{V_4^*}$	−0.0108	−0.0044	0.0089	0.0144	0.0186	−0.0030	−0.0179	−0.0115
	$v_5 = \frac{V_5 - V_5^*}{V_5^*}$	0.0042	−0.0176	0.0161	0.0118	−0.0161	−0.0054	−0.0036	0.0112

TABLE I *continued*

	Variables	Years							
		9 (1962)	10 (1963)	11 (1964)	12 (1965)	13 (1966)	14 (1967)	15 (1968)	16 (1969)
Absolute values	$\Phi_1 = 1 - \delta$	0.724	0.729	0.747	0.741	0.716	0.711	0.709	
	$\Phi_2 = I_{FG} + I_{FC}$	112	144	−381	−50	64	−283	−265	
	$V_1 = T$	2458	2510	2801	3344	3714	4051	4559	
	$V_2 = T_C$	981	857	709	659	725	1036	1085	
	$V_3 = U$	688	830	813	919	1030	1293	1520	
	$V_4 = K$	98.98×10^3	104.7×10^3	111.18×10^3	120.098×10^3	129.58×10^3	136.086×10^3	146.74×10^3	
	$V_5 = G$	4920	5184	5512	6043	6572	7246	7702	
Trends	Φ_1^*	0.7305	0.7343	0.7357	0.7341	0.7289	0.7193	0.7053	0.6857
	Φ_2^*	5.3576	−39.6387	−85.1348	−128.6023	−167.5125	−199.3375	−211.549	−231.6183
	V_1^*	2384.4616	2631.8579	2919.5532	3252.7459	3636.6336	4076.4141	4577.2852	5144.4447
	V_2^*	789.6871	762.0021	754.8758	775.6239	831.5622	930.0065	1078.2726	1283.6763
	V_3^*	697.4610	757.1941	839.0942	948.5097	1090.7896	1271.2805	1495.3326	1768.2937
	V_4^*	99 473	105 111	111 524	118 850	127 229	136 799	147 700	160 071
	V_5^*	4850	5205	5606	6058	6566	7133	7765	8464
Normalized deviations from trends	$\phi_1 = \frac{\Phi_1 - \Phi_1^*}{\Phi_1^*}$	−0.0088	−0.0072	0.0153	0.0093	−0.0149	−0.0118	0.0052	
	$\phi_2 = \frac{\Phi_2 - \Phi_2^*}{\Phi_2^*}$	19.9046	−3.8759	3.4752	−0.6112	−1.3820	0.4196	0.1961	
	$v_1 = \frac{V_1 - V_1^*}{V_2^*}$	0.0308	−0.0463	−0.0406	0.0280	0.0212	−0.0062	−0.0039	
	$v_2 = \frac{V_2 - V_2^*}{V_2^*}$	0.2422	0.1246	−0.0607	−0.1503	−0.1281	0.1139	0.0062	
	$v_3 = \frac{V_3 - V_3^*}{V_3^*}$	−0.0135	0.0961	−0.0310	−0.0311	−0.0557	0.0170	0.0164	
	$v_4 = \frac{V_4 - V_4^*}{V_4^*}$	−0.0049	−0.0039	−0.0030	0.0104	0.0184	−0.0052	−0.0065	
	$v_5 = \frac{V_5 - V_5^*}{V_5^*}$	0.0143	−0.0040	−0.0167	−0.0025	0.0009	0.0157	−0.0080	

TABLE II
TRANSFORMATION OF INPUT DATA (FIG. 1) TO OBTAIN THE SET OF INTERPOLATION POINTS

	$\mathcal{D}^2$	ϕ_1	$x_1 = V_{1F}$	$x_2 = V_{2F}$	$x_3 = V_{3F}$	$x_4 = V_{4F}$	$x_5 = V_{5F}$	$x_6 = V_{1(F-1)}$
16 (1969)								−0.0039
15 (1968)	12.100	0.0052	−0.0039	0.0062	0.0164	−0.0065	−0.0080	−0.0062
14 (1967)	15.080	−0.0118	−0.0062	0.1139	0.0170	−0.0052	0.0157	0.0212
13 (1966)	396.288	−0.0149	0.0212	−0.1281	−0.1503	0.0184	0.0009	0.0280
12 (1965)	1.220	0.0093	0.0280	−0.1503	−0.0311	0.0104	−0.0025	−0.0406
11 (1964)	17.456	0.0153	−0.0406	−0.0607	−0.0310	−0.0030	−0.0167	−0.0463
10 (1963)	0.108	−0.0072	−0.0463	0.1246	0.0961	−0.0039	−0.0040	0.0308
9 (1962)	2.826	−0.0088	0.0308	0.2422	−0.0135	−0.0049	0.0143	0.0353
8 (1961)	0.548	0.0153	0.0353	−0.0212	−0.0222	−0.0115	0.0112	0.0006
7 (1960)	0.535	0.0235	0.0006	−0.1851	0.0527	−0.0179	−0.0036	−0.0306
6 (1959)	7.0381	0.0192	−0.0306	0.0488	0.1007	−0.0030	−0.0054	0.0012
5 (1958)	2.786	−0.0416	0.0012	−0.0524	−0.0440	0.0186	−0.0161	0.0203
4 (1957)	0.013	−0.0064	0.0203	−0.0086	−0.059	0.0144	0.0118	−0.0022
3 (1956)	0.016	0.0088	−0.0022	−0.0857	−0.0812	0.0089	0.0161	−0.0170

TABLE II *continued*

	$x_7 = V_{3(F-1)}$	$x_8 = \phi_{1(F-2)}$	$x_9 = V_{2(F-2)}$	$x_{10} = V_{4(F-2)}$	$x_{11} = V_{1(F-3)}$	$x_{12} = \phi_{2(F-4)}$	$x_{13} = V_{2(F-1)}$	$x_{14} = V_{5(F-1)}$
16 (1969)	0.0164	−0.0118	0.1139	−0.0052	0.0212	−0.6112	−0.1503	−0.0026
15 (1968)	0.0170	−0.0149	−0.1281	0.0184	0.0280	3.4752	−0.0607	−0.0167
14 (1967)	−0.0557	0.0093	−0.1503	0.0104	−0.0406	−3.8759	0.1246	−0.0040
13 (1966)	−0.0311	0.0153	−0.0607	−0.0030	−0.0463	19.9048	0.2422	0.0143
12 (1965)	−0.0310	−0.0072	0.1246	−0.0039	0.0308	−1.0845	−0.0212	0.0112
11 (1964)	0.0961	−0.0088	0.2422	−0.0049	0.0353	−4.1646	−0.1851	−0.0036
10 (1963)	−0.0135	0.0153	−0,0212	−0.0115	0.0006	0.2760	0.0488	−0.0054
9 (1962)	−0.0222	0.0235	−0.1851	−0.179	−0.0306	1.6507	0.0524	−0.0161
8 (1961)	0.0527	0.0192	0.0488	−0.0030	0.0012	0.7343	−0.0086	0.0118
7 (1960)	0.1007	−0.0416	0.2524	0.0186	0.0203	0.6876	−0.0857	0.0161
6 (1959)	−0.0440	−0.0064	−0.0086	0.0144	−0.0022	−2.6491	0.0290	−0.0176
5 (1958)	−0.059	0.0088	−0.0857	0.0089	−0.0107	1.6634	0.0224	0.0042
4 (1957)	−0.0812	0.0158	0.0290	−0.0044	0.0029			
3 (1956)	−0.0216	0.0048	−0.0224	−0.010				

TABLE III
THE TRAINING SET

N	3	5	2	1	10	7	11	4	8	9
$\mathscr{D}^2$	396.288	17.456	15.080	12.100	7.0381	2.826	2.786	1.220	0.548	0.535
ϕ_1	-0.0145	0.0153	-0.0118	0.0052	-0.0192	-0.008	-0.0416	0.0093	0.0153	0.0235
x_1	0.0212	-0.0406	-0.0062	-0.0039	-0.0306	0.0308	0.0012	0.0280	0.0353	0.0006
x_2	-0.1281	-0.0607	0.1139	0.0062	0.0488	0.2422	0.0524	-0.7503	-0.0212	-0.1851
x_3	-0.056	-0.0310	0.0170	0.0164	0.1007	-0.0135	-0.0140	-0.0311	-0.0222	0.0527
x_4	0.0184	-0.0030	-0.0052	-0.0065	-0.0030	-0.0049	0.0186	-0.0104	-0.0115	-0.0179
x_5	0.0009	-0.0167	0.0157	-0.008	-0.0051	0.0143	-0.0161	-0.0025	0.0112	-0.0036
x_6	0.0280	-0.0465	0.0212	-0.0062	0.0012	0.0353	0.0203	-0.0406	0.0006	-0.0306
x_7	−0.0311	0.0461	-0.0557	0.0170	-0.0440	-0.0222	-0.059	-0.0310	0.0527	0.1007
x_8	0.0153	-0.0088	0.0093	-0.0147	-0.0064	0.0235	0.0088	-0.0072	-0.0192	-0.0416
x_9	-0.0607	0.2422	-0.1503	-0.1281	-0.0086	-0.1851	-0.0857	0.1246	0.0488	0.0524
x_{10}	-0.0030	-0.0049	0.0104	0.0184	0.0144	-0.0179	0.0089	-0.0039	-0.0030	0.0186
x_{11}	-0.0463	0.0353	-0.0406	0.0280	-0.0022	-0.0306	-0.0167	0.0308	0.0012	0.0203
x_{12}	19.9048	-4.1646	-3.876	3.4752	-2.6497	1.6507	1.6634	−1.0845	0.7343	0.6876
x_{13}	0.2422	-0.1851	0.1246	-0.0607	0.0240	0.0524	0.0224	-0.0212	-0.0086	-0.0857
x_{14}	0.0143	-0.0036	-0.0010	-0.0167	-0.0176	-0.0161	0.0042	0.0112	0.0118	0.0161
y_{23}	-0.0143	0.0078	-0.0156	-0.0091	-0.0098	-0.0127	-0.0116	0.0195	-0.0045	0.0225
z_{62}	-0.0152	0.0116	-0.0253	0.0063	-0.0126	-0.0161	-0.0173	0.0149	0.0017	0.0242
v_{45}	-0.0184	0.0114	−0.0118	0.0062	-0.0145	-0.0107	-0.0413	0.0133	0.0148	0.0232

TABLE IV
THE TESTING SET

N	6	13	12	$\delta^2 \times 10^4$
$\mathscr{D}^2$	0.108	0.016	0.013	
fa	−0.0072	0.0088	−0.0064	
x_1	−0.0463	−0.0022	0.0203	
x_2	0.1246	−0.0857	−0.0086	
x_3	0.0961	−0.0812	−0.0590	
x_4	−0.0039	0.0089	0.0144	
x_5	−0.0040	0.0161	0.0118	
x_6	0.0308	−0.0107	−0.0022	
x_7	−0.0135	−0.0216	−0.0812	
x_8	0.0153	0.0048	0.0158	
x_9	−0.0212	0.0224	0.0290	
x_{10}	−0.0115	−0.0108	−0.0044	
x_{11}	0.0006	—	0.0029	
x_{12}	0.2760		—	
x_{13}	0.0488	—	—	
x_{14}	−0.0054	—	—	
y_{23}	−0.0152	0.0054	−0.0053	0.770 16
z_{62}	−0.0092	0.01	−0.002	0.248 09
v_{45}	−0.0076	0.0104	−0.005	0.046 05

The solution is

$$a_0 = 0.6873, \quad a_1 = -0.0033$$

$$a_2 = 0.0018, \quad a_3 = -0.0001.$$

The trend of the process is

$$\Phi_1^* = 0.6873 - 0.0033t + 0.0018t^2 - 0.0001t^3.$$

The function for the trend is to be properly chosen. For example, it is clear from Fig. 2 that a polynomial trend is good for variable Φ_1, but a better trend for predicting variable Φ_2 would be a harmonic function.

2) Choice of the Length of the Prehistory Interval $T = Nh$: Let us show (for the sake of brevity only) how the control equation for the variable ϕ_1 (the index of industry occupation) was deduced. The reader can find the other equations (two for predicting ϕ_1 and ϕ_2 and two for controlling ϕ_1 and ϕ_2) in [1]. For the first attempt we choose the interval of prehistory equal to $T = \theta_1 = 4$ yr (the first self-sampling threshold). The control equation for ϕ_1 has the following form:

$$\begin{aligned}\phi_{1F} = f_1[&\phi_{1(F-1)}, \phi_{1(F-2)}, \phi_{1(F-3)}, \phi_{1(F-4)},\\ &\phi_{2(F-1)}, \phi_{2(F-2)}, \phi_{2(F-3)}, \phi_{2(F-4)},\\ &v_{1F}, v_{1(F-1)}, v_{1(F-2)}, v_{1(F-3)}, v_{1(F-4)},\\ &v_{2F}, v_{2(F-1)}, v_{2(F-2)}, v_{2(F-3)}, v_{2(F-4)},\\ &v_{3F}, v_{3(F-1)}, v_{3(F-2)}, v_{3(F-3)}, v_{3(F-4)},\\ &v_{4F}, v_{4(F-1)}, v_{4(F-2)}, v_{4(F-3)}, v_{4(F-4)},\\ &v_{5F}, v_{5(F-1)}, v_{5(F-2)}, v_{5(F-3)}, v_{5(F-4)}].\end{aligned}$$

3) Sampling of the More Useful Arguments: To shorten the calculations let us choose only those 14 arguments which have the strongest correlations with the output variable ϕ_1 (the second self-selection threshold on correlation criterion $q(\theta_2) = 14$). Denote the selected variables by

$$\begin{aligned} x_1 &= v_{1F} & x_8 &= \phi_{1(F-2)}\\ x_2 &= v_{2F} & x_9 &= v_{2(F-2)}\\ x_3 &= v_{3F} & x_{10} &= v_{4(F-2)}\\ x_4 &= v_{4F} & x_{11} &= v_{1(F-3)}\\ x_5 &= v_{5F} & x_{12} &= \phi_{2(F-4)}\\ x_6 &= v_{1(F-1)} & x_{13} &= v_{2(F-2)}\\ x_7 &= v_{3(F-1)} & x_{14} &= v_{5(F-4)}. \end{aligned}$$

In these denotations according to [1] we shall have the control equation of the form

$$\phi_{1F} = f_3(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}).$$

4) Construction of the Training and the Testing Sets: Having input data for only 15 yr (Fig. 2) and choosing the time interval of prehistory equal to 4 yr, we can construct only 13 points of interpolation. Ten of them, with the larger variation $\mathscr{D}^2$, we use for the training data set and three points (having less variation $\mathscr{D}^2$) for the testing set (Tables II–IV).

TABLE V
CHANGE OF MEAN-SQUARE ERROR BETWEEN LAYERS

y_8	0.000 201 8	z_3	0.000 120 3	v_1	0.000 036 1	w_{10}	0.000 008 6
y_{23}	0.000 077	z_{17}	0.000 029 5	v_{14}	0.000 233 3	w_{17}	0.000 085 9
y_{26}	0.002 873 4	z_{31}	0.000 416 8	v_{26}	0.000 065 0	w_2	0.000 039 1
y_{42}	0.000 164 9	z_{37}	0.000 052 7	v_{45}	0.000 004 6	w_{38}	0.000 004 9
y_{37}	0.000 214	z_{49}	0.000 049 2	v_{47}	0.000 007 7	w_{37}	0.000 014 0
y_{58}	0.000 254 1	z_{62}	0.000 024 8	v_{38}	0.000 026 7	w_{61}	0.000 003 8
y_{65}	0.000 465 5	z_{39}	0.000 126 5	v_{68}	0.000 056 2	w_{39}	0.000 023 8
y_{72}	0.000 161 9	z_{40}	0.000 080 9	v_{57}	0.000 034 5	w_{57}	0.000 007 5
y_{80}	0.000 143 7	z_{58}	0.000 110 0	v_8	0.000 068 1	w_{79}	0.000 017 9
y_{83}	0.000 117 7	z_{21}	0.000 125 8	v_{42}	0.000 019 5	w_{83}	0.000 007 9
y_{82}	0.000 217 5	z_{22}	0.000 079 6	v_{52}	0.000 008 1	w_{87}	0.000 009 1
y_{79}	0.000 145 4	z_{23}	0.000 083 9	v_{53}	0.000 020 8	w_{44}	0.000 005 4
y_{24}	0.000 171 8	z_{87}	0.000 111 6	v_{54}	0.000 010 5	w_{12}	0.000 013 8
y_{85}	0.000 338 9	z_{76}	0.000 168 9	v_{13}	0.000 104 6	w_{91}	0.000 041 0

Numeration of Indices

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		1	2	3	4	5	6	7	8	9	10	11	12	13
2			14	15	16	17	18	19	20	21	22	23	24	25
3				26	27	28	29	30	31	32	33	34	35	36
4					37	38	39	40	41	42	43	44	45	46
5						47	48	49	50	51	52	53	54	55
6							56	57	58	59	60	61	62	63
7								64	65	66	67	68	69	70
8									71	72	73	74	75	76
9										77	78	79	80	81
10											82	83	84	85
11												86	87	88
12													89	90
13														91

TABLE VI
CONTROL EQUATION FOR OUTPUT VARIABLE ϕ_1

$$\phi_1 = v_{45} = 0.003\,323 + 0.646\,507z_{87} + 0.526\,509z_{37} - 16.453\,265z_{87}z_{37} + 0.865\,475z_{87}^2 + 0.180\,493z_{37}^2$$

$$z_{87} = 0.008\,973 - 0.156\,455y_{82} + 1.251\,193y_{24} - 39.357\,736y_{82}y_{24} + 8.323\,526y_{82}^2 - 17.029\,865y_{24}^2$$

$$z_{37} = -0.007\,554 + 1.215\,627y_{37} - 0.093\,074y_{42} - 134.128\,761y_{37}y_{42} + 62.837\,984y_{37}^2 + 82.155\,422y_{42}^2$$

$$y_{82} = -0.105\,098 - 0.279\,247x_{10} + 0.502\,991x_{11} + 6.177\,137x_{10}x_{11} + 14.975\,779x_{10}^2 + 8.386\,482x_{11}^2$$

$$y_{24} = -0.009\,200 - 0.075\,002x_2 - 0.075\,860x_{13} + 0.283\,867x_2x_{13} + 0.276\,377x_2^2 + 0.126\,828x_{13}^2$$

$$y_{37} = -0.002\,446 - 0.900\,065x_4 - 0.041\,400x_5 + 78.041\,982x_4x_5 + 19.375\,319x_4^2 - 1.912\,654x_5^2$$

$$y_{42} = 0.004\,638\,79 - 0.945\,760x_4 - 0.728\,913x_{10} - 109.743\,571x_4x_{10} - 8.106\,316x_4^2 - 53.126\,352x_{10}^2$$

Example: Calculate the variation for point 3 (Table III):

$$\mathscr{D}^2 = \frac{1}{n}(x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2)$$

$$= \frac{1}{14}[(0.0212)^2 + (-0.1281)^2 + (-0.0557)^2 + \cdots + (0.0143)^2]$$

$$= 396.288.$$

5) First Layer of Selection via the Mean-Square-Error Criterion: There are 91 ways to order the 14 arguments, 2 at a time. For every combination we can write the partial regression equation of second order

$$y_a = a_{0a} + a_{1a}x_b + a_{2a}x_c + a_{3a}x_b^2 + a_{4a}x_c^2 + a_{5a}x_bx_c$$

where $a = 1,2,\cdots,91$ and b and c are indices for all 91 combinations. So we obtain 91 systems of conditional equations and therefore 91 systems of the normal Gaussian equations with matrices of the order 6×6. We find the values of the coefficients by the solution of the normal equation systems constructed from the training set data. Then we calculate the accuracy of every variable y_a using the testing set data only. From all the variables we choose the 14 most accurate (the third self-selection threshold $q(\theta_3) = 14$).

6) Remaining Layers of Selection via the Mean-Square-Error Criterion: Fourteen intermediate variables y_a chosen from the first layer give us the 91 combinations of 2, again in the next (the third) layer

$$z_a = b_{0a} + b_{1a}y_b + b_{2a}y_c + b_{3a}y_b^2 + b_{4a}y_c^2 + b_{5a}y_by_c$$

where $a = 1,2,\cdots,91$ and b and c are indices for all 91 combinations. We repeat the calculation of the coefficients and the estimation of accuracy for variables z_a. Then we choose the 14 z_a variables for the next (the fourth) layer where

$$v_a = c_{0a} + c_{1a}z_b + c_{2a}z_c + c_{3a}z_b^2 + c_{4a}z_c^2 + c_{5a}z_bz_c$$

where $a = 1,2,\cdots,91$ and b and c are indices of the 91 combinations. So we proceed from one layer to the next layer, controlling the mean-square error on the testing set.

We denote the variables of the different layers by the letters:

Layer: 0 1 2 3 4 5 6 7

Letter: x y z v w p q s.

The change in the prediction accuracy of the variables is shown in Tables IV and V. The first minimum of error is reached in the q layer (the sixth layer).

7) Results of the Calculations: The control equation for the output variable ϕ_1 can be represented by the polynomials shown in Table VI. The accuracy can be evaluated by the figures

$$\delta^2 = 0.46 \times 10^{-5} \quad \text{or} \quad \Delta = \frac{\sum_{i=1}^{3} (\phi_i - q_{9i})^3}{\sum_{i=1}^{3} \phi_i^3} \cdot 100$$

$$= \frac{0.46 \times 10^{-5}}{0.17 \times 10^{-3}} \times 10^2$$

$$= 2.71 \text{ percent.}$$

Taking into account that this error is calculated for the separate testing set, it is already extremely low.

Weakness of Control Theory: Tracing the change in the error rate for each layer of the selection procedure, we can explain why the theory based on differential equations is not adequate for solving complex problems. What can control theory propose for the description of the dynamics of the British economy? It can yield linear differential equations of first or fifth order [4]. Such descriptions are equivalent to the linear polynomial descriptions, which are less accurate than the first layer of the quadratic selection. They provide very low correlation to experimental data (the error is more than $\Delta = 100$ percent). In the sixth layer of the selection process, where the degree of the complete polynomial is equal to $2^6 = 64$, we obtain the error $\Delta = 2.71$ percent. Control theory cannot yield such an accurate description of the dynamics of the British economy.

We can insert in the control equation, valid for the year 1969, the following figures from Table I:

$$\Phi_1 = \phi_1 \Phi_1^* + \Phi_1^*, \quad \text{where } \Phi_1^* = 0.6857 \text{ (trend value)}$$

$$x_6 = -0.0039 \qquad x_9 = 0.1139 \qquad x_{12} = -0.6112$$

$$x_7 = 0.0164 \qquad x_{10} = -0.0652 \qquad x_{13} = -0.1503$$

$$x_8 = -0.0118 \qquad x_{11} = 0.0212 \qquad x_{14} = -0.0026.$$

So, we obtain the control equation of the form

$$\Phi_1 = f(a, v_1, v_2, v_3, v_4, v_5)$$

where a is a digital parameter.

The *optimization equations* (for 1969) shall be

$$\frac{\partial \Phi_1}{\partial v_1} = 0, \qquad \frac{\partial \Phi_1}{\partial v_2} = 0, \qquad \frac{\partial \Phi_1}{\partial v_3} = 0$$

$$\frac{\partial \Phi_1}{\partial v_4} = 0, \qquad \frac{\partial \Phi_1}{\partial v_5} = 0.$$

These equations enable us to find the optimal values of the manipulated variables which maximize the index Φ_1.

8) Optimization of the Thresholds: We have chosen the following values of the thresholds: $\theta_1 = 4$ yr, $q(\theta_2) = 14$ properties, and $q(\theta_3) = 14$ variables. We have divided the set of the interpolation points in the training and the testing sets in the proportion 10 : 3. Then we find the trends of all variables as curves of third degree only ($m = 3$). All these figures can be optimized by the calculation and comparison of several variants to increase the accuracy (if it is necessary). The other indexes of the economic situation can be predicted and controlled very accurately in the manner shown in the preceding.

9) Prediction for 1969: The values of the 1969 manipulated variables are known: $x_1 = 0.011\,965\,39$, $x_2 = -0.124\,389\,8$, $x_3 = -0.045\,972\,96$, $x_4 = -0.044\,417\,76$, $x_5 = -0.040\,879\,0$. We can insert these figures into the control equation and calculate[3]

$$\phi_1 = 0.0177.$$

The value of the trend is $\Phi^* = 0.6857$. So we predict $\Phi_1 = \phi_1 \Phi^* + \Phi^* = 0.6857(1 + 0.0177) = 0.6978$. The true value for 1969 is $\Phi_1 = 0.691$.

XIV. THE CONSTRUCTION OF VERY ACCURATE TRENDS BY THE ALGORITHM OF GMDH WITH LINEAR POLYNOMIALS

Let us approximate the mean line of a given random process $\phi(t)$by the curve

$$\phi = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \cdots + a_m t^m$$

where m is rather high (for example, $m = 128$). The best value of m is to be found by optimization using the mean-square-error criterion on the testing set. Denoting $x_i = t^i$, we have

$$\phi = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_m x_m.$$

This is the complete description. The partial descriptions of the first layer are

$$y_1 = a_0' + a_1' x_1 + a_2' x_2, \; y_2 = a_0'' + a_1'' x_1 + a_2'' x_3, \; \cdots, \; y_s = a_0^{(s)} + a_1^{(s)} x_{m-1} + a_2^{(s)} x_m.$$

The partial descriptions of the second layer are

$$z_1 = b_0' + b_1' y_1 + b_2' y_2, \; z_2 = b_0'' + b_1'' y_1 + b_2'' y_3, \; \cdots, \; z_p = b_0^{(p)} + b_1^{(p)} y_{s-1} + b_2^{(p)} y_s$$

and so on. The minimum number of interpolation points in the training set is equal to the number of unknown coefficients of one partial polynomial plus one, whereas it is sufficient to know at least five points of interpolation (four in the training set and one in the testing set) to evaluate all the coefficients in the partial and, therefore, in the complete polynomial. The more points that are used, the more confident we can be in the results.

[3] This figure can be obtained only when the coefficients are calculated with at least an accuracy of 15 digits after the decimal point. The coefficients in the equations given in the preceding have 6 digits after the decimal point only for brevity in this paper.

XV. LONG-TERM PREDICTIONS AND STABILITY OF CONTROL

The best way to predict the future values of the variable for two or three years ahead is to repeat the prediction year after year, two or three times. For a more long-term prediction the best way is to use the last two or three real points of interpolation and one, two, or three predicted points[4] to construct very accurate trends with high degree m. This very accurate trend can be used for long-term prediction and therefore for the stability investigation of the control processes in the system.

XVI. CONCLUDING REMARKS

The GMDH can be considered as the combination of regression analysis and the methods of decision regularization. Let us denote the number of interpolation points by N. Regression analysis gives us the possibility of finding the unique polynomial which has any number of terms from $n = 1$ to $n = N - \backslash$ according to the criterion of minimum mean-square error. When the number of terms is $n - N$, we obtain the unique polynomial, with zero error, which fits all the points of interpolation exactly. When the number of terms is more than $n = N$, we can find an unlimited number of polynomials having zero errors, but none of them is applicable—they are too sensitive to changes in the input data.

The unique and stable (regular) decision (when the number of terms and the degree of polynomial are high) corresponds to the minimum mean-square error which is estimated on the separate testing set. As the complexity of the plant increases, so does the degree of the number of terms of the optimal polynomial. When the number of terms is more than $n = N$, the decision can be reached by the GMDH only.

Pessimistic conclusions have been drawn concerning the quality of solution found by regression analysis. It was stated in [6] that the coefficients of the regression equations cannot be considered as the true weighting coefficients for every particular argument. The second disadvantage is the necessity of very large experimental data sets for a regression equation of high degree. The GMDH shows that both of these statements are incorrect. It gives us the unique and regular regression equation in which every coefficient is the true weight of the argument associated with it. Large data sets are not necessary for the regression equations of any complexity which is optimal. Consequently, the GMDH is the basis for the new polynomial theory of complex dynamic systems. The consequences of this new theory should be considerable.

The polynomial theory of complex dynamic systems will bring about a complete revolution of the art of prediction, pattern recognition, identification, optimalizing control with information storage, and to the other problems of engineering cybernetics. For example, there will be no need to wait for a half century to prove which economic system is best. It will be sufficient to identify the polynomial descriptions of two systems and to compare their potential possibilities. V. I. Lenin has pointed out that the main performance index of the economic system is the potential productivity of labor (i.e., not the productivity measured in the given instant of time, but the productivity which can be reached in the system based on its long-term trends). It will be possible to identify the polynomial descriptions for the productivity of labor in capitalist and in socialist economic systems to prove the potential advantages of the socialist system without waiting for many years. Such predicting, having been made very objectively, will have strong ideological influence, first to the minds of scientists, then to other people.

Many other problems of modern complex life can be solved by the use of the polynomial theory. The spectrum of these problems includes the models of the brain, controllers for complex industrial processes, the problems of clean environment (particularly the model for the formation of clean water), and economic and social problems.

[4] The number of points used is to be defined by the minimum error criterion, again calculated on the separate testing set.

REFERENCES

[1] A. G. Ivakhnenko, Yu. V. Koppa, and V. S. Min, "Polynomial and logical theory of dynamic systems (pt. I)," *Sov. Automat. Contr.*, vol. 3, 1970, no. 3, pp. 1–13.
——, "Polynomial and logical theory of dynamic systems (pt. II)," *ibid.*, vol. 3, 1970, no. 4, pp. 11–30.
[2] A. G. Ivakhnenko, "Heuristic self-organization in problems of engineering cybernetics," *Avtomatika*, vol. 6, Mar. 1970; also *Mes. Steuern Regeln*, 1970, nos. 4 and 6.
[3] I. Babushka and S. L. Sobolev, "Optimization of numerical methods," *Aplikace Matematiky*, vol. 10, 1965, no. 2, pp. 96–129.
——, "Optimization of numerical methods," *ibid.*, vol. 13, 1968, no. 1, pp. 27–38.
[4] P. Parks and Y. Pyatt, "Controlling the economy—an exercise in collaboration," *Electron. Power*, vol. 15, Aug. 1969, pp. 270–274.
[5] L. M. Boychuk, "Polynomial descriptions of nonlinear first-order plants by the 'differential fork' method," *Sov. Automat. Contr.*, vol. 3, 1970, no. 4, pp. 1–8.
[6] V. C. Nalimov and N. A. Chernova, *Statistical Methods of Planning the Extremum Experiments*. Moscow, 1965.