Chapter 6 from the book
"Inductive Learning Algorithms for Complex System Modeling",
Madala H.R. and Ivakhnenko A.G., 1994, ISBN: 0-8493-4438-7, CRC Press;

Chapter 6
Applications

Inductive learning algorithms, called Group Method of Data Handling (GMDH) were developed using the principles of self-organization modeling. Self-organization of modeling is the process of finding the optimal structure of mathematical description of a complex object by sorting many variants according to a certain ensemble of external criteria. Unlike the traditional modeling approaches which are deductive in character, the inductive learning algorithms are based on the sorting of models according to the external criteria agreed on by experts. The inductive approach does not eliminate the experts or take them away from the computer, but rather assigns them a special position. Experts indicate the selection criteria of a very general form and interpret the chosen models given by the criteria as the best. They can influence the result of modeling by formulating new criteria. The overall approach of the expert becomes that of an objective referee in resolving scientific controversies. The improvement of man-machine (ergatic) systems is based on the gradual reduction of the human involvement in the process. The automated systems become more mechanized; i.e., they are not only automated but are fully automatic as well. The human element often involves errors and undesired decisions. One of the examples is the process of specifying the objectives, or determining the set of criteria. Future development might lead to sequential decision-making algorithms that include an automatic setting of selection criteria, their sequential determination, freedom-of-choice, and so on—almost without having to involve experts to solve important problems. This means that the involvement of human element will be reduced.

At the present state of development, experts specify external criteria and consider results without interfering with the optimization processing. In case of disagreement, the experts can utilize higher levels of criteria (noise stability criteria), so that controversies are quickly resolved. The problems of nonlinear identification and of long-range forecasting of complex processes are solved by the computer.

Two-level prediction schemes of learning allow one to leave the choice of the model to the computer, significantly increasing the lead time of forecasting. The objective character of the modeling speeds up the understanding of the object, allowing us to avoid false subjective selection.

The following features of inductive approach allow us to improve the quality of self-organization modeling and to give the procedure an objective character [32].

1. The minimum-bias criterion agrees that models obtained with the use of two different sets of data be identical. Such a model leads to a physical model that is isomorphic to the mechanism of the object under consideration. In self-organization modeling, a

physical model is used only to determine the composition of the set of output variables subject to forecasting. Knowing this, we proceed to self-organization of nonphysical models.

Nonphysical models differ from physical models in the composition of the arguments used and in the external criteria. The step-by-step prediction criterion assumes great importance. Nonphysical models can be constructed in various mathematical languages that differ in the degree of prediction accuracy which is measurable by the correlation time [44]. The nonphysical models can use two- or three-dimensional time readout. All this is important because the limit of informativeness of a forecast is determined by the degree of blurredness of the modeling language.

Nonphysical models are used in the self-organization modeling of two-level schematic predictions—for example, annual and seasonal predictions for which the balance-of-predictions criterion allows the increase in the lead time of detailed (seasonal) predictions. The usage of an auxiliary criterion like prediction criterion is necessary to curtail the volume of sorting to the reasonable number of annual and seasonal models and to ensure the uniqueness of the choice of predictions according to the prediction-balance criterion. The balance criterion serves as the choice of both short-term as well as long-term forecasts.

2. The class of the equations and the form of support functions are selected by sorting many variants of models according to the selection criteria. For example, a system of finite-difference equations are formed for use in self-organization modeling of ecological systems. In traditional approaches, the physical equations are considered and their discrete analogues with variable coefficients are formed to represent the object, but ultimately it ends up with poor predictive characteristics.
3. The third feature of the inductive approach is the selection of the set of output and input variables, and of the "leading" variable among them. The objective system analysis allows us to obtain the least biased system of equations according to the system criterion of minimum bias.

   The inductive algorithm for OSA allows us to sort out all possible systems of equations consisting of one, two, three, and so on equations and to select the most unbiased model of system of equations. This determines the structure of the object and the set of output variables. The best output variable which forecasts better than all others is called as the "leading" variable.

   The variables that are interested may not enter into the set of output variables during OSA and have to be predicted as a supplement in terms of the functions of the output variables according to the two-level prediction scheme in time.
4. The self-organization modeling allows us to obtain models with optimal complexity even in case of incomplete information; i.e., without having the data of many important arguments. This is an antithesis of the idea of increasing the information basis up to some universal measure. The inductive learning approach, however, demonstrates the success of such modeling; for example, for predicting winter wheat harvest more than 50 arguments are required to have—the use of fertilizer, method of tillage, periods of irrigation, and so on. All available arguments are very important. Nevertheless, in self-organization modeling, only two to three arguments participate for sufficient accuracy of forecasts (this is further explained in the given examples). Although the connections among the arguments made during the self-organization modeling remain unknown to us, these may be taken into account for an accurate forecasts.
5. The self-organization modeling is possible with noisy data. The inductive learning algorithms obey the laws valid in the communication theory—particularly, Shannon's

second-limit theorem for transmission of noisy signals.

The inductive algorithms allow us to restore the physical model of the object under study even in the case of noisy data exceeding three to four times the regular signal [63]. The methodologies used in the communication theory and the pattern recognition theory allow us to raise the noise immunity of the algorithms. This means that the accurate models can be obtained with an incomplete information basis and noisy data in the same way as the accurate signal is restored under the noisy conditions and distortions of various kinds.

The overall modeling is object-oriented because all problems are solved according to the agreed selection criteria. The final results of modeling may not coincide with the ideas of the modeler about the object being modeled.

1 FIELD OF APPLICATION

The inductive method is an empirical method and is intended for self-organization of mathematical models based on measured data. The object of the modeling is identification and prediction of the object. The usual methods of regularization (for example, regression analysis) are mathematically elegant but inexpedient external additions.

In self-organization theory, an entire series of more apt criteria are proposed—regularity, minimum bias, prediction and others oriented towards satisfaction of the practical users of the models. These criteria are applied sequentially one after another, to eliminate the difficulties of normalization (choice of the coefficients) of the criteria. Multicriterial choice of the model is one of the foundations of noise immunity of the inductive algorithms.

Gödel's incompleteness theory is fulfilled by the problem being concerned with the choice of the ensemble of external complements, and its composition and sequence of application at different levels is solved by sifting a number of variants. The upper level criteria is based on:

1. criterion for noise stability: $\Theta_1 = \frac{\text{noise}}{\text{signal}} \rightarrow \max$,
2. criterion for maximizing the lead time of predictions: $\Theta_2 = \frac{\text{lead time}}{\text{observation time}} \rightarrow \max$, and
3. criterion for minimizing the amount of computational time: Θ_3 = computer operational time $\rightarrow \min$.

The ensemble of criteria is to better ensure the required value of the pair of criteria Θ_1 and Θ_2. The objectivity is based on the choice of the set of criteria.

The field of application of the inductive learning algorithms is shown in Figure 6.1. This is widely spread on the plane with the coordinating values of Θ_1 and Θ_2, and $\Theta_3 \geq 3$ hours of operational time (usually on minicomputers).

The modeling of optical systems requires application of single-level models with a usual time reference, accurate initial measured data, and the complete information basis. Two-dimensional time readout (in terms of seasons and years) in case of noisy data is required in predicting certain agricultural productions, river flows, and so on. Two-level forecasting with two-dimensional time readout with incomplete information basis is used for econometric models, modeling of climatic changes, and ecological systems. Some of the practical examples corresponding to these are given in the preceding chapters. This chapter is extended further for more specific examples in the areas of weather modeling, ecosystem studies, economical systems modeling, agricultural system studies, and solar activity.

We hope that the reader will get an overall idea of how to approach to this type of model-

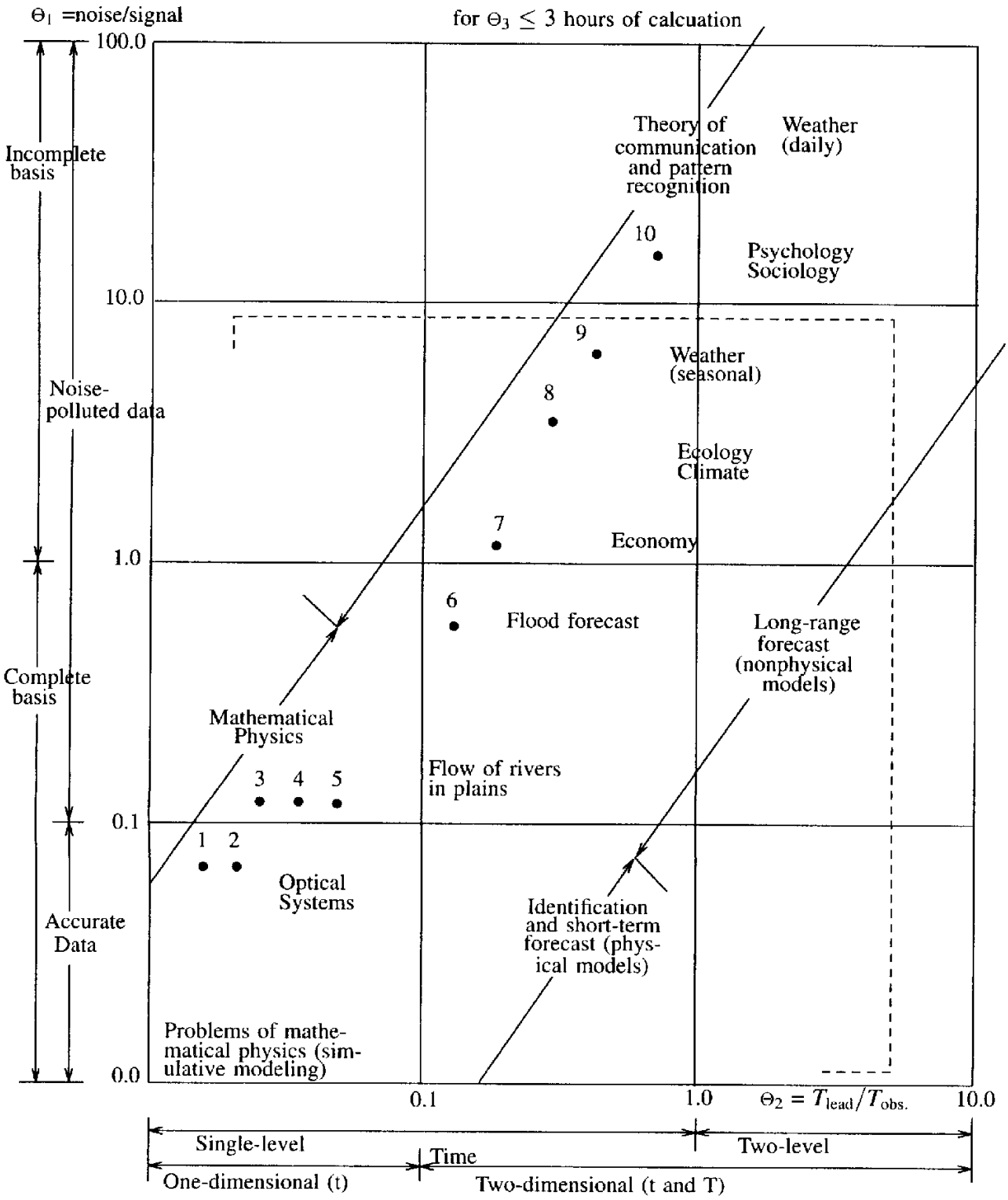

Figure 6.1. Various fields of application using inductive algorithms

ing. However, the reader must keep in mind that more complex modeling problems with very high noise levels require development of new algorithms using special mathematical formulas, new external criteria, and criteria with canonical forms. The field of self-organization modeling is open to receiving and providing further technological achievements.

2 WEATHER MODELING

Self-organization modeling requires the presence of two components: a generator of variety of models (combinatorial or multilayer) and some sensibly chosen ensemble of external criteria to evaluate these models. A relatively short data sample is needed to estimate the parameters of the models and to compute the values of the criteria. A general description of the equations comprising the system to be formed may be known to the modeler. So, a composite approach is frequently the optimum approach, whereby a general description specified *a priori* as a reference by the human author and remaining analysis is done by computer sorting in accordance to various criteria.

2.1 Prediction balance with time- and space-averaging

The problem of identification of complex objects based on the emperical data is treated as an ill-posed problem, in as much as a unique solution always requires the use of some external information or an external supplement according to Gödel. This means that it is in principle impossible to obtain a unique model in optimal complexity without regularization. In addition to the basic criteria like regularity and minimum-bias, the following are some criteria are convenient to use for cylindrical polar coordinates of meteorological field.

Interpolation balance criterion

The difference equation pattern has only six nodes (Figures 6.2 and 6.3). During the training process the patterns are moved vertically upwards in the time axis at one-day steps. Each position of the pattern yields one conditional equation. All models are trained according to the inductive algorithm. The expression is

$$q_{ij}^t = \frac{1}{4}(q_{i+1j}^t + q_{i-1j}^t + q_{ij+1}^t + q_{ij-1}^t), \tag{6.1}$$

which, from the theory of difference equations, can be used to construct the criterion of interpolation balance. If N is the number of pattern positions in the interpolation region, then the criterion is written as

$$b^2 = \sum_{i=1}^{N}[q_{ij}^t - \frac{1}{4}(q_{i+1j}^t + q_{i-1j}^t + q_{ij+1}^t + q_{ij-1}^t)]^2. \tag{6.2}$$

Balance criteria

Balance-of-variables and balance-of-predictions criteria are used when several variables are predicted simultaneousy. The former requires that some relationship that exists between the variables at a given time should also exist in the future, whereas the latter should be used in an ensemble with the predicting models.

As given before, the prediction balance criterion with time averaging of variables (averaging over a season and over a year) which is used mainly for cyclic processes can be

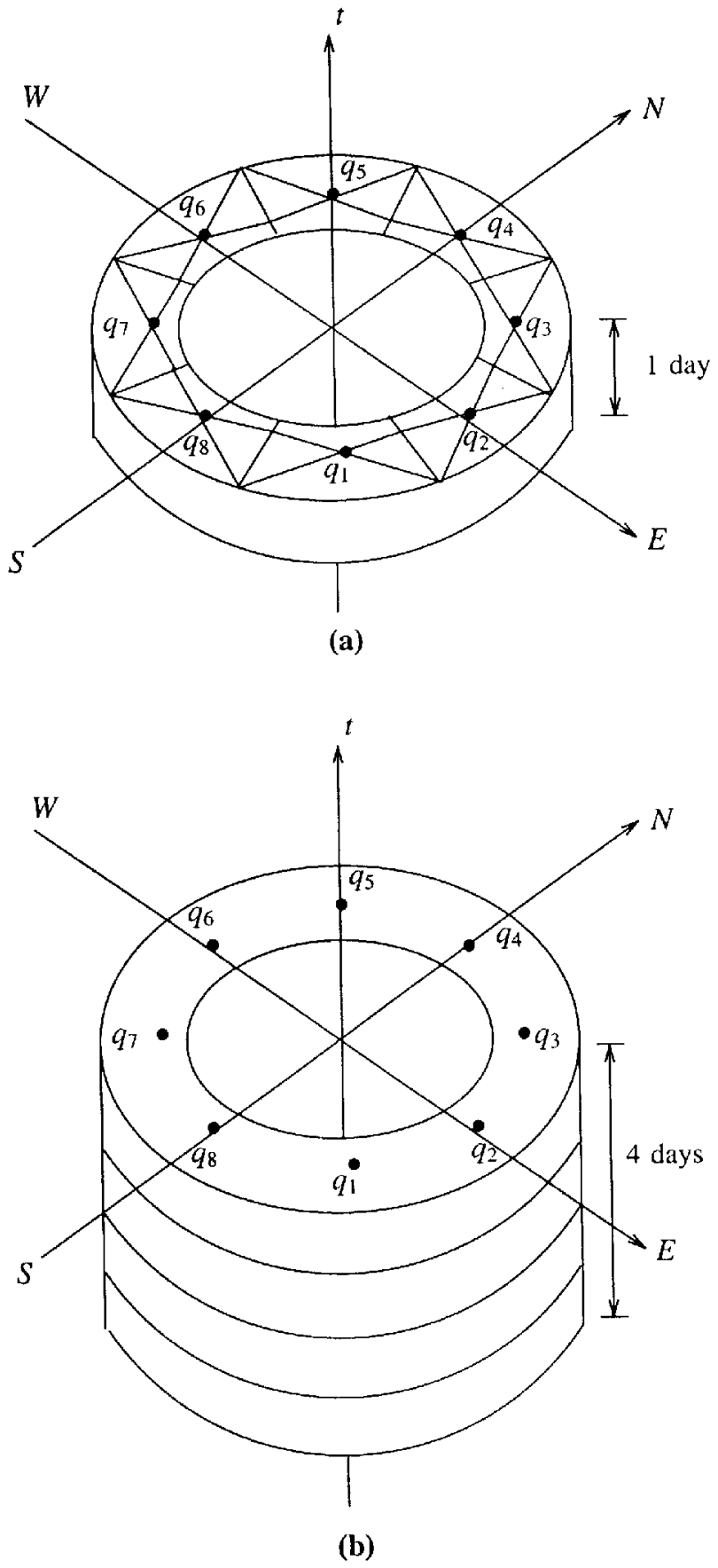

Figure 6.2. Application of interpolation balance criteria with averaging a) in space (average of eight predictions is balanced to the average over entire area of the ring), and b) in time and space (average of $8 \times 4 = 32$ predictions is balanced to the prediction of the average over entire volume of the tube

written as

$$b^2 = [\hat{q}_{year} - \frac{1}{4}(\hat{q}_w + \hat{q}_{sp} + \hat{q}_{su} + \hat{q}_f)]^2,$$
$$B^2 = \sum_{i \in N} b_i^2 \rightarrow \min, \tag{6.3}$$

where $\hat{q}_w, \hat{q}_{sp}, \hat{q}_{su}$, and $\hat{q}_f$ are predicted average seasonal values of the winter, spring, summer and fall variables correspondingly and $\hat{q}_{year}$ is the predicted average annual value. N is the number of years covered by the prediction.

Prediction balance criterion with space averaging of variables

The prediction balance criterion can be applied in a similar way with the averaging over the area of a ring, which can be written as

$$b^2 = [\frac{1}{8}(\hat{q}_1 + \hat{q}_2 + \cdots + \hat{q}_8) - \hat{q}_0]^2,$$
$$B^2 = \sum_{i \in N} b_i^2 \rightarrow \min, \tag{6.4}$$

where $\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_8$ are predictions obtained for eight patterns which form a ring around the axis of a cylinder (Figure 6.2), $\hat{q}_0$ is a prediction of the variable averaged over the entire area of the ring, and N is the number of steps at which the prediction is checked on the time axis.

Prediction balance criterion with time and space averaging

Here time intervals are, for example, days and 4-day period and space intervals are one pattern and the area of the ring; this can be written as

$$b^2 = [\frac{1}{8}\{\frac{1}{4}(\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4)_1 + \frac{1}{4}(\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4)_2 + \cdots$$
$$\cdots + \frac{1}{4}(\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4)_8\} - \hat{q}_0]^2,$$

$$B^2 = \sum_{i \in N} b_i^2 \rightarrow \min, \tag{6.5}$$

where $\hat{q}_0$ is the predicted value of the variable averaged over the entire volume of the pipe consisting of four rings.

Normalized combined criterion

$$c6^2 = (\frac{\eta_{bs} - \eta_{bs_{\min}}}{\eta_{bs_{\max}} - \eta_{bs_{\min}}})^2 + (\frac{\Delta(C) - \Delta(C)_{\min}}{\Delta(C)_{\max} - \Delta(C)_{\min}})^2 + (\frac{b - b_{\min}}{b_{\max} - b_{\min}})^2 \rightarrow \min \tag{6.6}$$

Alternatively, normalization can be avoided by using the criteria $\eta_{bs}^2 \rightarrow \min$, $\Delta^2(C) \rightarrow \min$ and $b^2 \rightarrow \min$ in a sequence one after the other.

To ensure the needed freedom of choice is the goal of every sequential decision making procedure; first F_1 models are selected from total of F_0 models using the first criterion, from this F_2 are selected by the second criterion, and finally the third criterion is used to

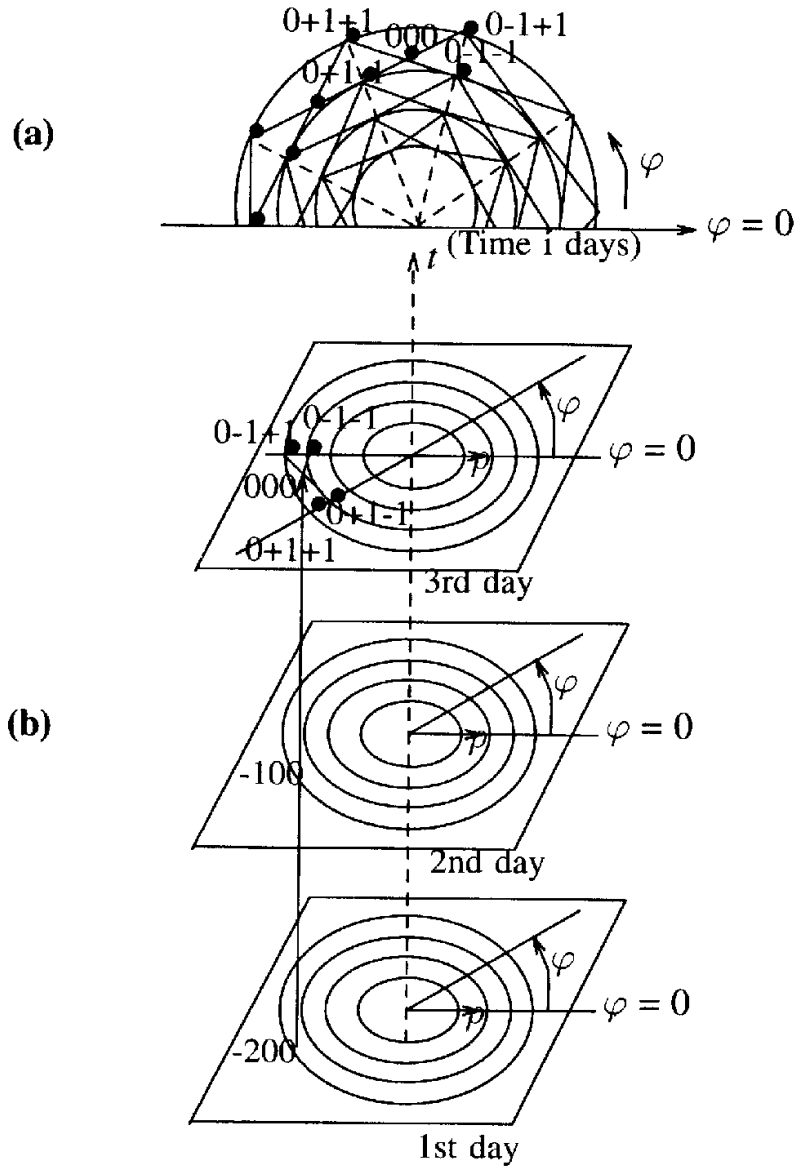

Figure 6.3. Positioning of implicit patterns on three daily charts in cylindrical polar coordinates ρ and φ; a) plan, and b) axonometry

select a single model of optimal complexity $F_0 > F_1 > F_2 > F_3 \geq 1$. The sequence of application is

$$F_0 \rightarrow \eta^2_{bs}(F_1) \rightarrow \Delta^2(C)(F_2) \rightarrow b^2(1). \tag{6.7}$$

To optimize the freedom-of-choice, F_1 and F_2 are chosen to select several identical models on the basis of η^2_{bs} and $\Delta^2(C)$. A single optimal model is selected from this group using the balance criterion. The number of models tested by the balance criterion can be increased depending on the computer capacity—usually $F_3 \leq 8$.

Sequential application of the criteria does not require the normalization of their values and also there is no need of introducing criterion weighting coeficients.

2.2 Finite difference schemes

Self-organization modeling requires one to indicate the list of variables containing a large access, an appropriate emperical data sample, and a reference function. The computer selects

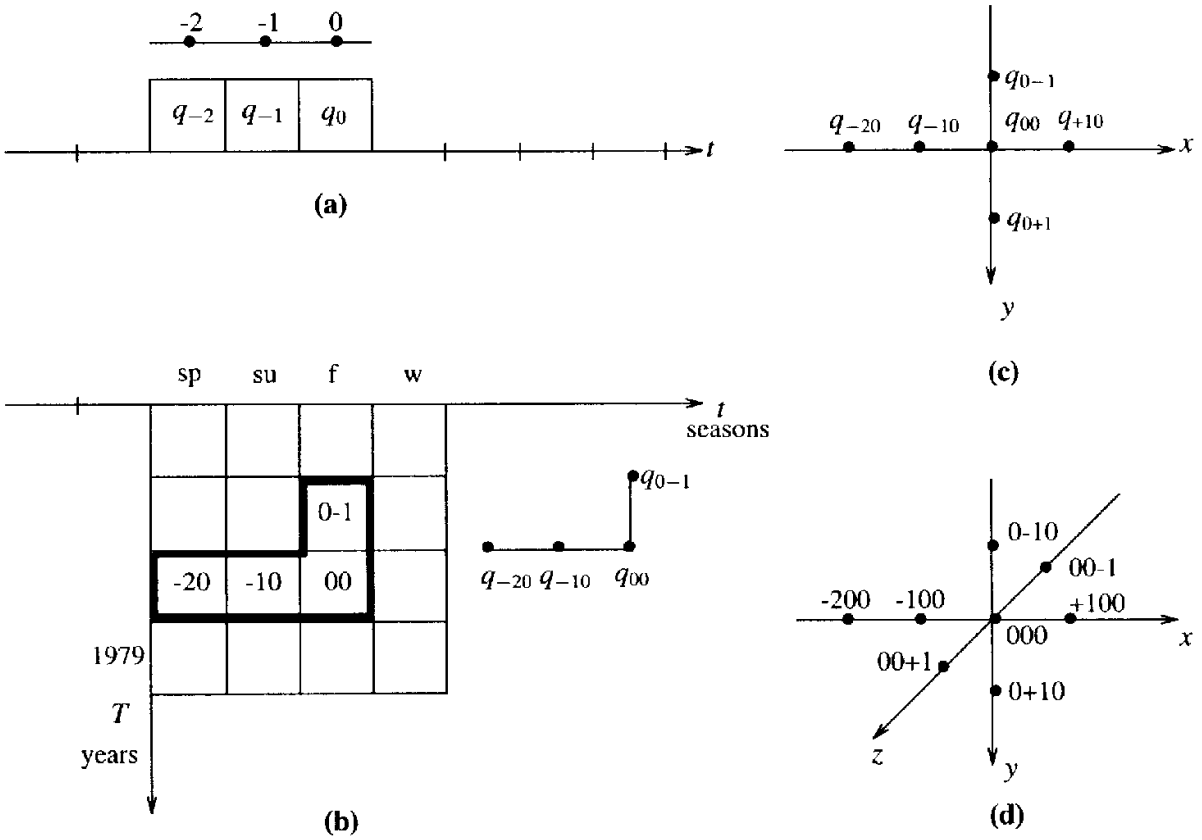

Figure 6.4. Explicit elementary patterns: (a) for the point problem $q(t)$, (b) for the two-dimensional time scale $q(t, T)$, (c) for the problem $q(x, y)$, and (d) for the three-dimensional problem $q(x, y, z)$

the ensemble of most effective arguments, finds the optimum structure of each equation, and estimates the coefficients. The depth of minimum of the selection criterion is the measure of the modeling.

Fields of meteorological parameters (temperature, pressure, humidity, etc.) are usually described by the linear partial differential equations. The linear finite-difference equations, which are discrete analogues of the linear differential equations are suggested as reference functions; these equations can be expanded by introducing nonlinear terms into them.

The following equation is an example of a point problem:

$$\frac{d^2q}{dt^2} + a_1\frac{dq}{dt} + a_0q = f(t). \tag{6.8}$$

By using the forward finite-differences of $\Delta q = (q^t - q^{t-1})$ and $\Delta^2 q = (q^t - q^{t-1}) - (q^{t-1} - q^{t-2})$, this equation has the following finite-difference form:

$$(q^t - 2q^{t-1} + q^{t-2}) + a_1(q^t - q^{t-1}) + a_0q^t = f(t) \tag{6.9}$$

The elementary pattern corresponding to this equation represents the arguments q^{t-1} and q^{t-2} which have an effect on the output q^t.

$$q^t = f_1(t) + f_2(q^{t-1}, q^{t-2}), \tag{6.10}$$

where $f_1(t)$ is the so-called source function or trend function.

To construct the system of conditional equations, the data collected by moving the pattern step-by-step along the time axis (Figure 6.4a) are arranged in the table form with the columns of t, q^t, q^{t-1}, and q^{t-2}. Each position of the pattern corresponds to one row of the data table which is divided into different parts. The prediction is obtained through step-by-step integration by moving the pattern into the region of future time.

Two-dimensional time-readout

A significant improvement in accuracy can be achieved in modeling of cyclic processes by using the bivariate time scale; for example, months and years or seasons and years, hours and days, etc. The pattern shown in Figure 6.4b indicates that the output $q_{t,T}$ is influenced by the arguments $q_{t-1,T}, q_{t-2,T}$, and $q_{t,T-1}$. This pattern can include not only the output variable but also the auxiliary variable like x as

$$q_{t,T} = f_1(t,T) + f_2(q_{t-1,T}, q_{t-2,T}, q_{t,T-1}, q_{t,T-2}, \cdots, x_{t,T}, x_{t-1,T}, x_{t-2,T}, x_{t,T-1}, \cdots). \tag{6.11}$$

The patterns related to the (x,y) and (x,t) planes can be identical to the patterns with bivariate time scale (t,T) (Figure 6.4c).

Similarly, the patterns related to the planes $(x,y,t),(x,y,z)$, or (x,y,z,t) (Figure 6.4d) are represented for spatial problems having number of delay arguments along each axis.

Implicit patterns in cylindrical coordinates

In self-organization modeling, the physical fields are identified on the basis of different reference functions; algebraic, harmonic, or finite-difference equations. Usually the finite-difference equations are preferred because they have additional advantages. They are linear in coefficients and nonlinear in variable parameters. But sometimes they may create some serious problems like providing unstable predictions. In such cases addtional measures are taken in achieving the convergence of step-by-step predictions such as

- decreasing the sampling interval of variables;
- simplifying the patterns and functions (not considering the nonlinear terms);
- changing from "explicit" to "implicit" patterns; and so on.

Let us see the concept of *implicit patterns*; they are realized on a closed curve and are moved simultaneously by one step. This yields a simultaneous system of equations. The most promising are the finite-difference models realized by implicit patterns that are constructed in cylindrical-polar coordinates (t,ρ,ϕ) (Figure 6.3). The differential equation of diffusion in this plane contains a linear sum of derivatives of not higher than the second order

$$a_0\frac{\partial^2 y}{\partial t^2} + a_1\frac{\partial^2 y}{\partial \rho^2} + a_2\frac{\partial^2 y}{\partial \phi^2} + a_3\frac{\partial y}{\partial t} + a_4\frac{\partial y}{\partial \rho} + a_5\frac{\partial y}{\partial \phi} = f(t,\rho,\phi), \tag{6.12}$$

where $a_0, a_1, a_2, \cdots, a_5$ are the constants and $f(t,\rho,\phi)$ is the source function.

The following finite-difference analogue can be replaced considering one additive trend equation.

$$y_{ij}^t = f_1(t,\rho,\phi) + f_2(y_{ij}^{t-1}, y_{ij}^{t-2}, y_{i+1j+1}^t, y_{i-1j+1}^t, y_{i-1j-1}^t, y_{i+1j-1}^t, z_{ij}^t, z_{ij}^{t-1}, z_{ij}^{t-2}, z_{i+1j+1}^t, z_{i-1j+1}^t, z_{i-1j-1}^t, z_{i+1j-1}^t), \tag{6.13}$$

where $f_1(t,\rho,\phi)$ is the trend function, y is the predicted variable, z are the variables which are correlated with the predicted variable, and i and j are the indices of ρ and ϕ coordinates.

The table of data sample is prepared for each position of the pattern along the closed circle that is achieved by placing the patterns along circular layers of the cylinder in such a way that the adjacent patterns have two common points. The patterns are trained using

the part of the data in one region and are used for step-by-step predictions in the other region. The meteorological factors are predicted for the first layer near the central axis of the cylinder making it possible to determine the weather in the second layer, and then the third layer, etc. for the entire region included in the cylinder.

For example, Figure 6.3 shows the placement of eight patterns on one of the three layers of the cylinder. The eight pattern equations can be written for the first layer as

$$a_0 + a_1 q_{ij}^{t-2^k} + a_2 q_{ij}^{t-1^k} + a_3 q_{i+1j+1}^{k} + a_4 q_{i+1j-1}^{k} + a_5 q_{i-1j+1}^{k} + a_6 q_{i-1j-1}^{k} = 0 \qquad (6.14)$$

where $k = 1, 2, \cdots, 8$ and the coefficients $a_0, a_1, \cdots, a_6$ are estimated using one of the inductive learning algorithms. While predicting, the values of all $q_{ij}^{t-2^k}$ and $q_{ij}^{t-1^k}$ are known as these were positioned in the past and the values of q_{i+1j-1}^{k} and q_{i-1j-1}^{k} are determined with a separate prediction because these are located near the cylinder axis. Pattern coupling equations $q_{i+1j-1}^{k} = q_{i+1j+1}^{k-1}$ or $q_{i-1j+1}^{k} = q_{i-1j-1}^{k+1}$ hold for q_{i+1j+1}^{k} and q_{i-1j+1}^{k}. Thus, 16 necessary and sufficient equations are available for determination of 16 variables (eight q_{i+1j+1}^{k} and eight q_{i-1j+1}^{k}).

The equations are obtained for the second and third layers of the cylinder; the only difference is that the values of q_{i+1j-1}^{k} and q_{i-1j-1}^{k}) are determined from the data of the previous layer, not by a separate prediction. This means that a separate prediction is required only in the region located near the cylinder axis.

The self-organization theory offers the additional means for improving the convergence of step-by-step prediction:

1. By using the finite-difference equations with variable coefficients; these have simpler patterns than the equations with constant coefficients.
2. By using the step-by-step prediction criterion (i^2) in the ensemble of the external criteria; this criterion selects from the set of possible models those that have adequate convergence.

2.3 Two fundamental inductive algorithms

Combinatorial algorithm

Since the original differential equation is idealized, the nonlinearities in the system are taken into account by expanding the finite-difference analogue (reference function) with higher ordered terms. For example, for a point problem we can have

$$\begin{aligned} q^t = f_1(t) + (a_0 + a_1 q^{t-1} + a_2 q^{t-2} + a_3 q^{t-3} + a_4 q^{t-1} q^{t-2} + a_5 q^{t-1} q^{t-3} \\ + a_6 q^{t-2} q^{t-3} + a_7 q^{t-1^2} + a_8 q^{t-2^2} + a_9 q^{t-3^2}), \end{aligned} \qquad (6.15)$$

where $f_1(t)$ is the trend or coarse model. This is usually obtained in advance by applying the least squares method through the experimental data or by using an inductive algorithm with the minimum-bias criterion as an external criterion. The purpose of this model is to decrease the number of arguments of the difference part, lumping together some of them into a separate term. In weather forecasting this model can correspond to a climatic forecast averaged over a long time. The remained difference part of the model refines this forecast.

The combinatorial inductive algorithm enables us to evaluate the "structure of functions" obtainable from this equation on the basis of minimum-bias and regularity criteria. There are 10 terms in the difference part of the above equation and the total number of equations to be tested is 2^{10}.

Multilayer algorithm

If the difference part of the above polynomial reference function contains more than 20 terms (varies according to the computer capacity), the multilayer algorithm with linearized terms is applied. The linearized version of the reference function is

$$q^l = f_l + (a_0 + a_1 w_1 + a_2 w_2 + \cdots + a_m w_m), \tag{6.16}$$

where $m \geq 20$ and w represent the terms q.

The multilayer algorithm determines the complete linear polynomial as a superposition of partial polynomials with two-variable of type ($\hat{q}_k = a_{0k} + a_{1k} w_i + a_{2k} w_j$, where k indicates the unit number; $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, m$; and $i \neq j$). This algorithm realizes the method of incomplete induction by omitting some partial polynomials during sorting and is never tested against the criteria. This is done in a multilayer feedforward network structure. The result of self-organization modeling is the output model with an optimal complexity. The computer selects the structure of the model, its nonlinearity, and the content of its arguments. During the processing the ineffective terms of the reference function are discarded.

2.4 Problem of long-range forecasting

The method of analogues is one of the interesting methods considered in the literature for reliable long-term weather forecasting. The idea of analogues is based on finding the interval whose meteorological characteristics are identical to those presently observed in the measured data and the future of this interval which is measured in the past is the best forecast at the present time. Although the idea is so simple, attempts to apply this idea always produce results that are not very convincing because with a large number of observed variables it is not possible to find the exact analogues in the pre-history data. The self-organization method based on the inductive approach can be interpreted as an improved method of group analogues in which the analogues of the present state of the atmosphere are selected by using special criteria to produce the most probable forecast.

The problem encountered is how to estimate, at least approximately, the achievable prediction time. The maximum achievable prediction time T_p of a one-step forecast is determined by the coherence time τ_c of the autocorrelation function $A_q(\tau)$. The maximum allowed prediction time of a multiple step-by-step forecast is equal to the coherence time multiplied by the number of steps ($T_{p_{\max}} = \tau_c . n$). This means that prediction error increases with each integration step that imposes a definite limit on the step-by-step forecast. This leads to seeking of the maximum capabilities of multiple step-by-step prediction, assuming that they are determined by the coherence time in the same way as they are for one-step prediction. However, studies of autocorrelation functions of meteorological parameters to determine maximum prediction time have not yet been completed. The expected results should be similar to the studies of autocorrelation functions for other complex systems. It turns out that averaging of variables in time increases the coherence time. One has to remember that it also depends on the physical properties of the process being predicted as well as on the quality and characteristics of the mathematical apparatus. This corresponds to the extreme variations from predicting "purely" deterministic objects like motions of planets to "purely" random objects like games of "lotto." The actual physical problems are always located in between these two cases.

The autocorrelation function of a process contains some information on its predictability (the degree of determinancy or randomness). According to studies it is evident that by increasing the averaging interval of variables in time or space shift the process from the region of unpredictability into the region of long-term calculability—i.e., centennial averages

and global averages of pressure, temperature, or humidity can be predicted a thousand years in advance. At the same time the predictions based on the daily avearges cannot be valid for more than 15 days. Nevertheless, it is possible to overcome the predictability limit by the following suggested possibilities that are applicable in predicting some weather variables like temperature, pressure, etc. at the surface layer.

2.5 Improving the limit of predictability

Here we discuss the possibilities of increasing the time of weather forecasts in the self-organization method.

1. The first potential contribution of the self-organization method in improving the predictability is in the mathematical apparatus and objective synthesis of the system of equations; the structures of the equations are selected by the minimum-bias criterion.

   The proposed equations are to be valid not only in the training region of the data sample, but also in the testing region. This is precisely what is done by the minimum-bias criterion. This means that the computer sorts out the number of equations for each specified output variable separately and finds out a system of equations that is invariant in time. This is the objectively evaluated system of equations based on the empirical data.
2. The second potential contribution of the self-organization theory in improving predictability is the composite use of different averaging of variables in time and space with the help of the prediction balance criteria. The balance criterion provides a reference point in the future, allows one to perturb the divergence of solutions which is a typical property of the hydrodynamic equations, and from there increases the entropy of predictions.

   In predicting the average daily values, the reference point in the future can be the forecast of average monthly values. Similarly, in predicting the variable of average over an area, the reference point of a forecast at a specified point on the surface of the earth can be the forecast of a sum of variables at several points. This means that when the space averaging is used, the forecast ability of variables averaged over large areas becomes higher. The forecast validity time based on short averages is pulled up toward the validity time on a large regions.
3. The third possibility provided by the self-organization theory is a significant widening of the complete set of arguments (input variables). It is followed by selecting the most effective of these arguments. It is reasonable to include the connected patterns that realize the long-range effect inserted with appropriate delays in time in the use of moving average sums of variables, which are analogous to the integral terms of hydrodynamic equations. If the proposed arguments turn out to be ineffective, these will not be included in the ensemble of predictor arguments.

Example 1. Self-organization modeling of air pressure and temperature at a point located on the cylindrical axes [58].

The possibilities mentioned for improving the limit of predictability in the self-organization method, which uses implicit patterns in cylindrical coordinates, allow one to compute future values of the meteorological parameters. These parameters are already predicted one-step (day) ahead at nodes of all patterns located on the cyldrical axes. The problem requires the one-step prediction of the parameters at one point on the surface of the earth. This can be obtained by using the multilayer algorithm which is demonstrated in this example.

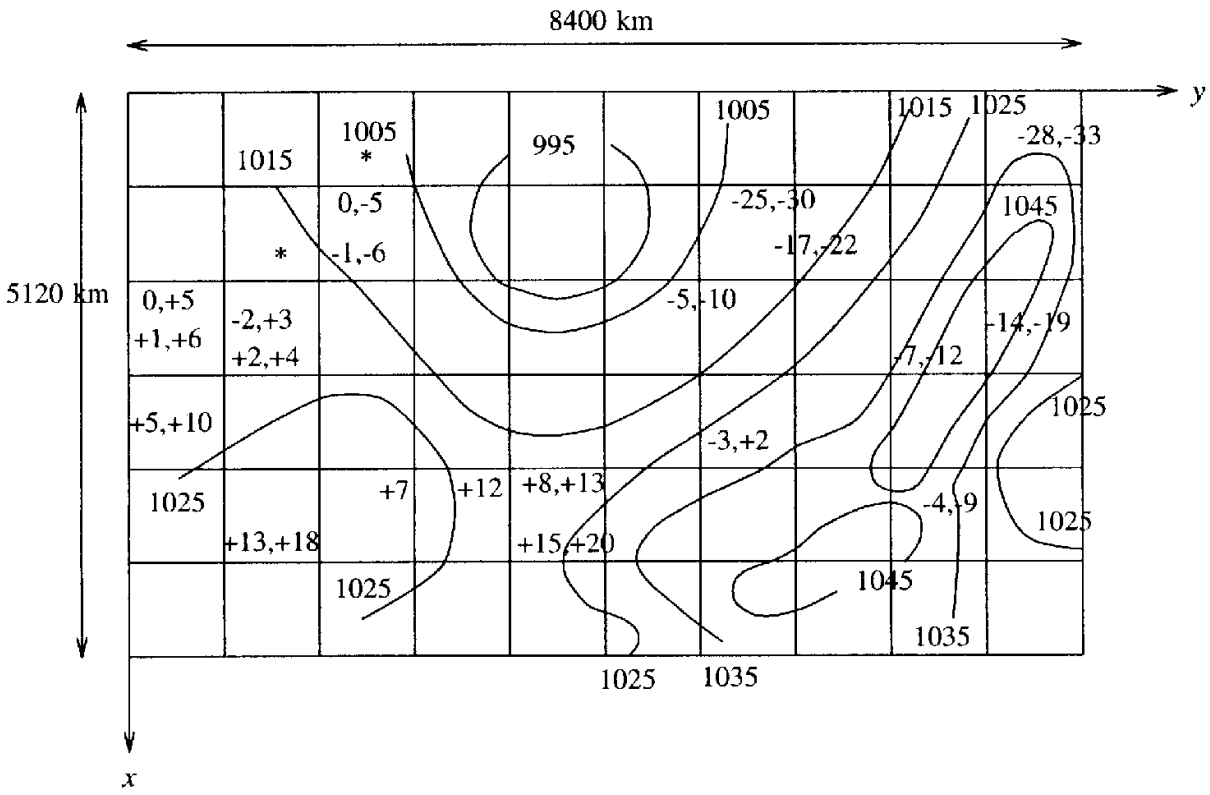

Figure 6.5. A sample weather chart

The example shows how one can find the optimum structure of equations which express the variation of temperature and pressure fields as a function of time. To determine the models used for predicting pressure (P) and temperature (T), the meteorological data charts (Figure 6.5) are divided into elementary cubes and average daily values of the meteorological parameters are prepared in each cube. These charts are stacked on time axis as a parallelopiped-space of given observations as shown in Figure 6.6. This consists of $6 \times 10 \times 10$ unit cubes with the units of measure 5120/6 km along the x-axis, 8400/10 km along the y-axis, and 1 day along the time axis.

Six shapes of even patterns which are tested in the example are given in Figure 6.7 and the corresponding finite-difference models in otimal complexity with the estimates of their accuracy are listed in the Table 6.1. (Here $\bar{P}$ and $\bar{T}$ are the average values of the arguments for the four nodes.) The equations of the complete pattern have the following general forms:

$$\begin{aligned} P_{ij}^{t+1} &= f_1(t,x,y) + f_2(P_{ij}^{t}, P_{ij}^{t-1}, P_{ij}^{t-2}, \cdots, T_{ij}^{t}, T_{ij}^{t-1}, T_{ij}^{t-2}, \cdots) \\ T_{ij}^{t+1} &= f_3(t,x,y) + f_4(T_{ij}^{t}, T_{ij}^{t-1}, T_{ij}^{t-2}, \cdots, P_{ij}^{t}, P_{ij}^{t-1}, P_{ij}^{t-2}, \cdots). \end{aligned} \tag{6.17}$$

The data are collected according to each pattern from the parallelopiped-space of the observed data and arranged in the tables in relation to the output and input variables; the data are normalized. The data are divided into sequences of A, B, and C, where $W = A \cup B = 80\%$, and $C = 20\%$. The data sequence C is used as a separate sequence for checking the models. The system of equations corresponding to each pattern are selected on the basis of the combined criterion ($c3^2 = \eta_{bs}^2 + \Delta^2(C)$). All equations are compared with each other on the basis of the relative error on the sequence C. The results show that the system of equations corresponding to the fifth pattern (Figure 6.7e) is the most accurate. The relative error is measured as

$$\delta_P = \sqrt{[\frac{1}{N_C} \sum_t \sum_x \sum y(P - \hat{P})^2]}/(P_{\max} - P_{\min}) \rightarrow \min$$

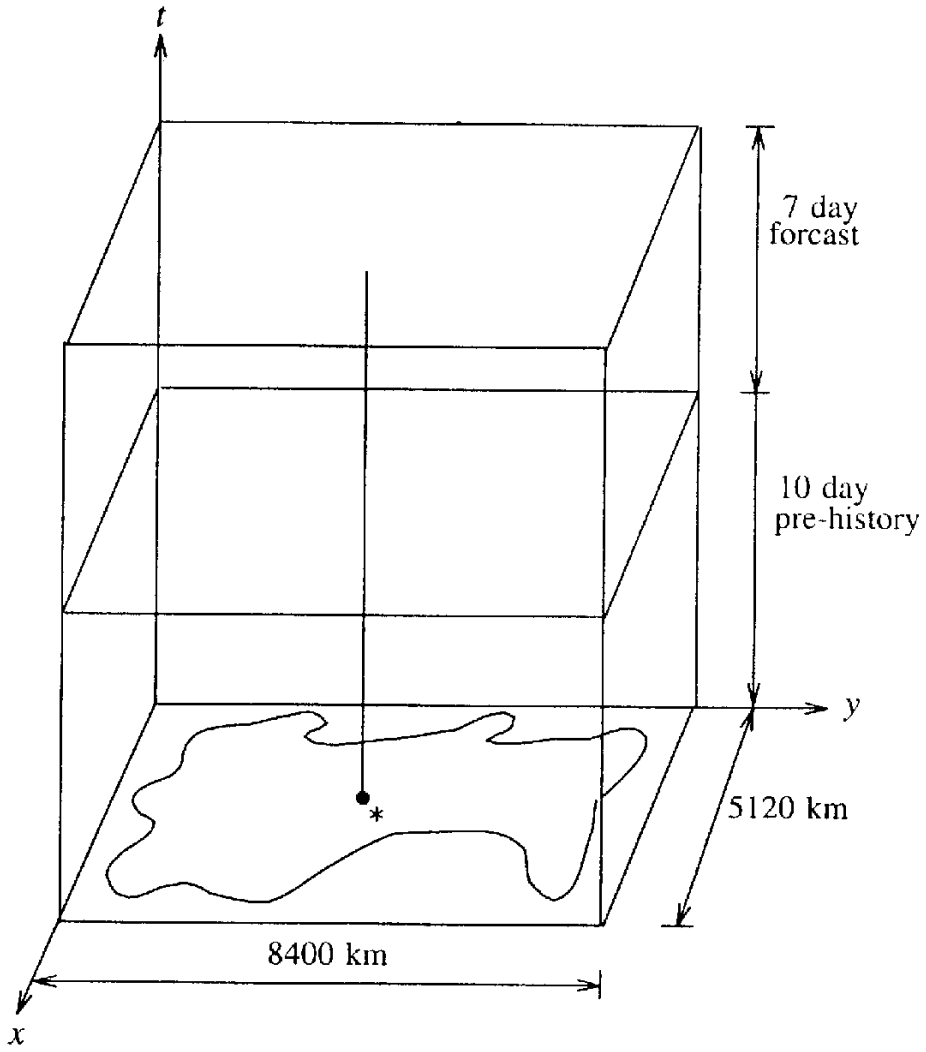

Figure 6.6. Parallelopiped experimental data for ten days (training and testing data) and of predictions for seven days; '*' indicates the point of interest, a location on the earth

$$\delta_T = \sqrt{[\frac{1}{N_C} \sum_t \sum_x \sum y(T - \hat{T})^2]/(T_{\max} - T_{\min})} \rightarrow \min, \tag{6.18}$$

where P and T are the actual values, $\hat{P}$ and $\hat{T}$ are the predicted values, $P_{\max}$ and $T_{\max}$ are the maximum values, and $P_{\min}$ and $T_{\min}$ are the minimum values of the pressure and temperature, correspondingly; N_C is the number of data points in the sequence C.

For $N_C = 7$ (, i.e., for a seven-day forecast,) the relative prediction errors on pressure and temperature are $\delta_P = 0.3709$ and $\delta_T = 0.3043$, correspondingly. Figure 6.8 shows the curves of predicted temperature and pressure for the axial point shown in the parallelopiped-space of data. The system of equations corresponding to the pattern V (Table 6.1) is used for this purpose. The system is adaptive—i.e., the forecast is updated everyday as new data are received.

The example given above is conducted on an experimental basis. In this example, neither prediction balance nor an objective synthesis of system of equations are used. This means that the possibilities of improving the accuracy of the forecasts have not been exhausted.

Table 6.1. Optimal models for different patterns

No.	System of equations	η_{bs}	$\Delta(C)$	$c3$
I	$P_{ij}^{t+1} = -1.0083 - 0.1099t + 0.3920T_{ij}^{t} + 0.8809\bar{P}_{ij}^{t}$ $+0.0047P_{ij}^{t}T_{ij}^{t} + 0.0121\bar{P}_{ij}^{t}\bar{P}_{ij}^{t} + 0.000177\bar{P}_{ij}^{t}\bar{T}_{ij}^{t}$	0.021138	0.276223	0.277031
	$T_{ij}^{t+1} = -1.7269 + 0.1411t + 0.9834\bar{T}_{ij}^{t}$	0.201362	0.290929	0.353817
II	$P_{ij}^{t+1} = 6.0649 + 0.2366\bar{P}_{ij}^{t} - 0.00348T_{ij}^{t-1} + 0.0059P_{ij}^{t}P_{ij}^{t-1}$ $+0.0047P_{ij}^{t}\bar{P}_{ij}^{t} + 0.01059P_{ij}^{t-1}T_{ij}^{t} + 0.00503\bar{P}_{ij}^{t}\bar{P}_{ij}^{t} + 0.00405\bar{P}_{ij}^{t}\bar{T}_{ij}^{t}$	0.022273	0.289461	0.290317
	$T_{ij}^{t+1} = -0.6725 + 0.9838\bar{T}_{ij}^{t} + 0.00172P_{ij}^{t-1}T_{ij}^{t-1}$	0.176932	0.352216	0.394159
III	$P_{ij}^{t+1} = 4.0505 - 0.00734P_{ij}^{t-2} + 0.4682\bar{P}_{ij}^{t} + 0.00745P_{ij}^{t}P_{ij}^{t-1}$ $+0.0124P_{ij}^{t-2}P_{ij}^{t-2} + 0.00566P_{ij}^{t-2}T_{ij}^{t-1} - 0.00265P_{ij}^{t-2}\bar{T}_{ij}^{t}$	0.121926	0.362449	0.382407
	$T_{ij}^{t+1} = 1.6342 + 0.4908T_{ij}^{t-2} + 0.5019\bar{T}_{ij}^{t} - 0.0000256P_{ij}^{t-1}P_{ij}^{t-2}$ $-0.00311P_{ij}^{t-2}P_{ij}^{t-2} - 0.00309T_{ij}^{t-2}T_{ij}^{t-2}$	0.079602	0.233970	0.247141
IV	$P_{ij}^{t+1} = -3.6239 + 0.6961t + 0.8436P_{ij}^{t} + 0.2970P_{i+1j}^{t}$ $+0.2335P_{ij-1}^{t} + 0.00815P_{i+1j}^{t}P_{i+1j}^{t} - 0.000075P_{i+1j}^{t}T_{ij-1}^{t}$ $+0.00238P_{ij-1}^{t}P_{i-1j}^{t} + 0.0112P_{ij-1}^{t}T_{ij-1}^{t}$	0.057442	0.352762	0.357409
	$T_{ij}^{t+1} = 1.28603 + 0.4706T_{i+1j}^{t} + 0.3842T_{ij-1}^{t}$	0.186287	0.197606	0.271572
V	$P_{ij}^{t+1} = 5.06 + 0.3648P_{ij-1}^{t} + 0.00787T_{ij-1}^{t} - 0.00538P_{ij}^{t}T_{i-1j}^{t}$ $+0.01848P_{i+1j}^{t}P_{ij-1}^{t} - 0.005722T_{ij}^{t}T_{i-1j}^{t} + 0.0000252P_{ij}^{t-1}P_{ij}^{t}$	0.080550	0.235598	0.248988
	$T_{ij}^{t+1} = -0.8886 + 0.2417T_{ij}^{t} + 0.2834T_{i+1j}^{t} + 0.2712T_{ij-1}^{t}$ $+0.00961P_{i+1j}^{t}T_{ij}^{t-1}$	0.038617	0.235150	0.238300
VI	$P_{ij}^{t+1} = 5.056 + 0.572P_{ij+1}^{t} - 0.00234P_{ij}^{t-2}T_{ij}^{t-1}$ $+0.0108P_{ij}^{t-1}T_{i+1j}^{t} + 0.006951P_{ij}^{t-1}T_{ij}^{t}$	0.190290	0.198452	0.279023
	$T_{ij}^{t+1} = 3.9901 + 0.642T_{i+1j}^{t} + 0.2215T_{ij}^{t-1}$ $+0.003696P_{i+1j}^{t}T_{ij}^{t-1} - 0.000228T_{ij}^{t}T_{ij}^{t-2}$	0.046587	0.298915	0.302524

2.6 Alternate approaches to weather modeling

Here are some other suggestions on how to use the inductive approach in solving the weather forecasting [60].

The self-organization method is of a heuristic nature. Its main idea is to generate a large variety of variables and functions connecting them, and to choose the best structure in optimal complexity according to an external criterion. The following proposals are made for better predictions:

- an autonomous system of homogeneous difference equations (for short-range predictions) is proposed to describe the change in instantaneous as well as the averaged values of the variables and to include any source function and external disturbances;
- the use of two-level predictions on the basis of several balances (for example, year-season, and year-month);
- a two-level algorithm (for medium-range predictions) to use with several balance criteria;
- the use of correlational models for predicting weather in movable coordinates;
- the use of ecological variables in a combined system of weather-climate equations to increase prediction accuracy and prediction time; and

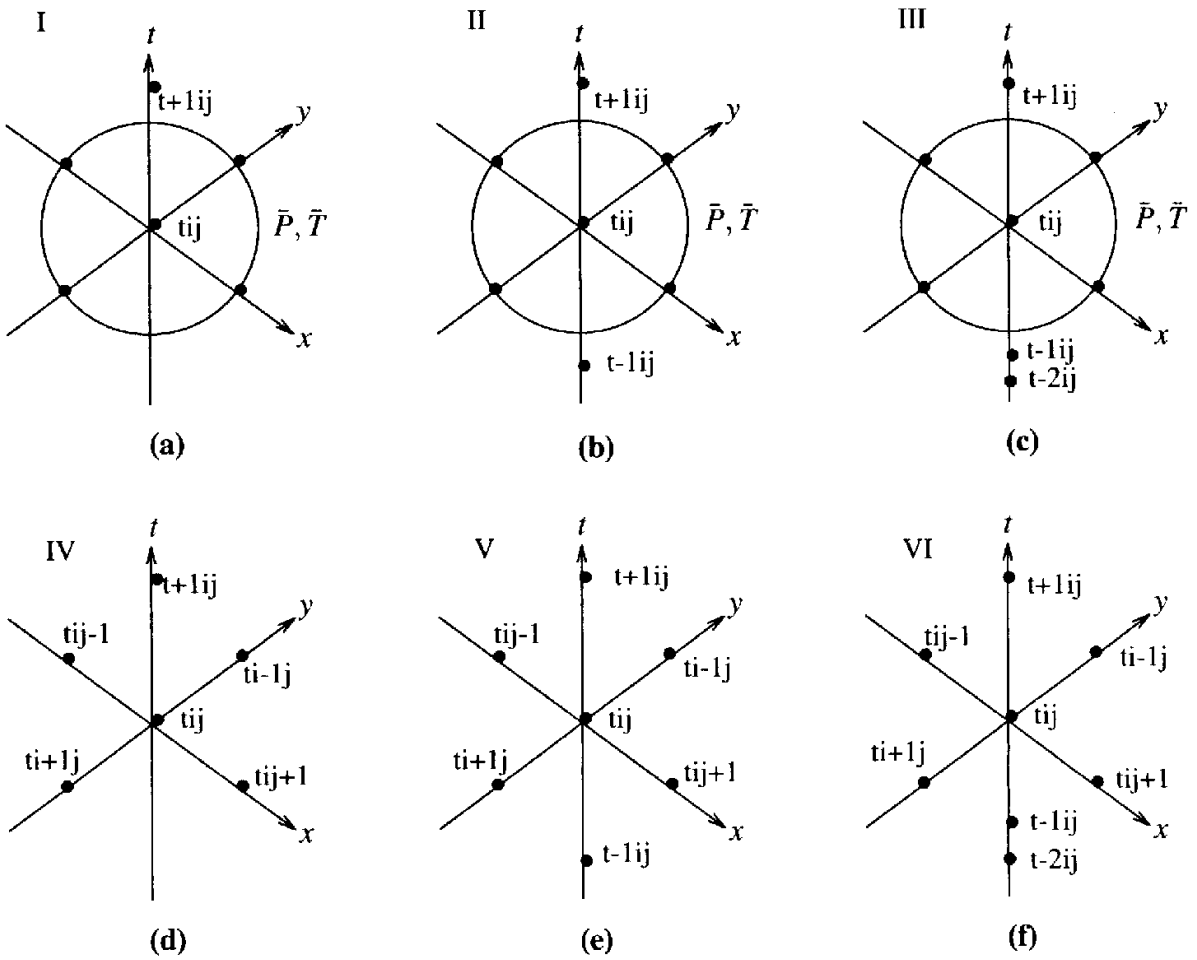

Figure 6.7. Different pattern schemes considered

- the use of a method of group analogues (for long-range predictions) that is based on objective clustering of the weather into number of clusters (not specified in advance) with averaging of the predictions of the variables one by one; the set of significant variables can be set up objectively by the objective system analysis algorithm.

The meteorological variables that determine the weather (air pressure, temperature, wind force, humidity, etc.) oscillate continuously around a mean climatic value (or a trend) in a random manner. The trend is usually known and can be predicted rather accurately. This means that the problem of weather forecasting is reduced to predicting the random deviations of the variables from the trend. These deviations are called the "remainder." One should note that all of the variables referred below correspond to such meteorological variables.

The suggested approaches are described below.

Weather modeling in fixed coordinates

The first approach suggests including the variables of external disturbing influences with the averaged, delayed, and higher-ordered arguments into the reference functions under consideration. The candidate variables usually used are in the original equations; that is, $q_1 = u$, the projection of the wind velocity onto the north-south axis, $q_2 = v$, the same onto the east-west axis, $q_3 = \rho$, the air density, and $q_4 = h$, the air humidity are introduced along with the variables $q_5 = P$, the pressure, $q_6 = T$, the temperature, and t, x, and y are

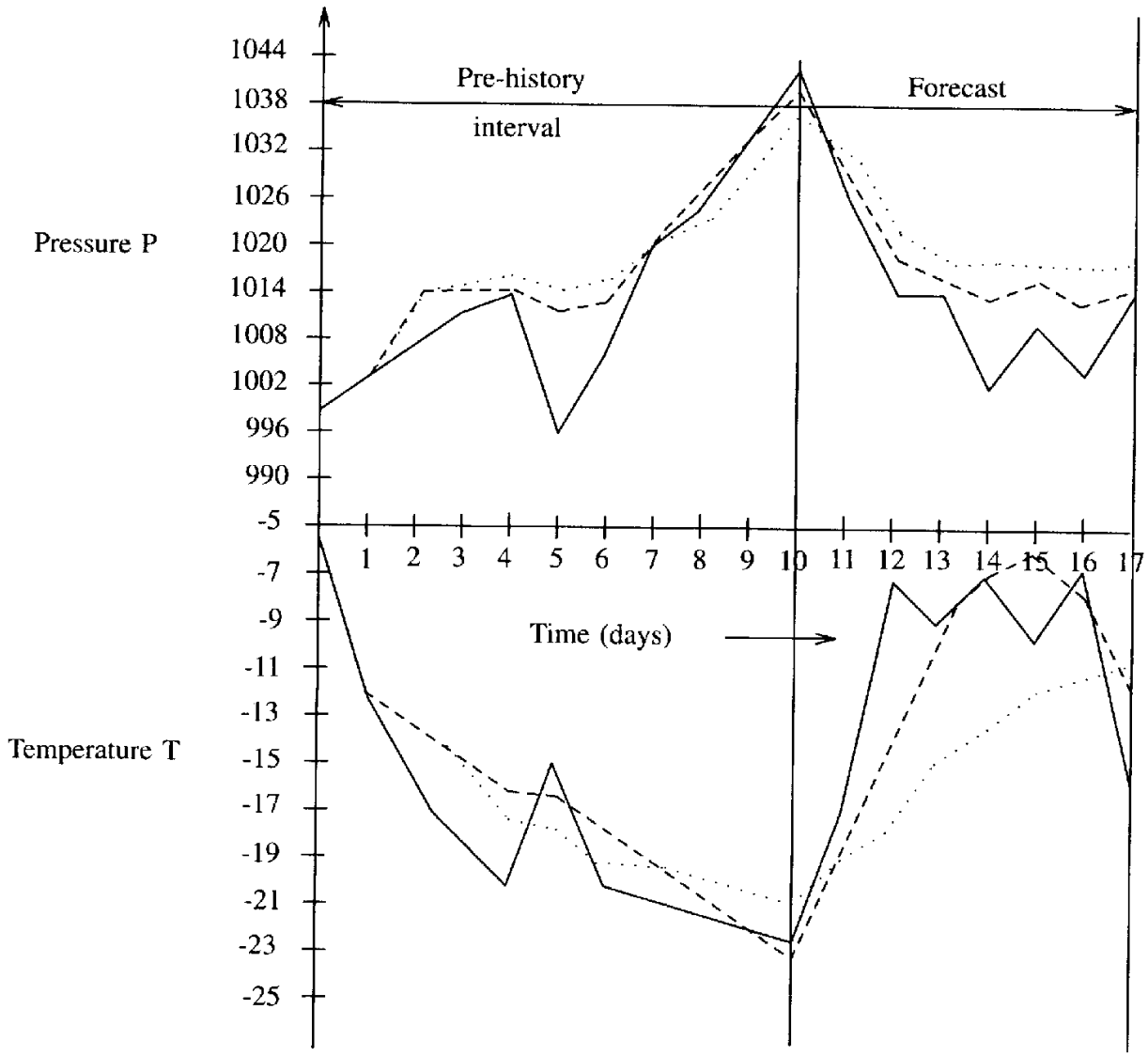

Figure 6.8. Performance of the system of equations trained based on the optimal pattern structure for pressure and temperature for pressure and temperature at the point of interest; the full line indicates the actual measured data, the broken line indicates the pattern training based on the measured data from the point of interest, and the dotted line indicates the pattern training based on the whole parallelopiped of data

the time-space coordinates. The variables of external influences can be the parameters of distant points of the geosphere; for example, the temperature or pressure of air at the centers of the Atlantic ocean (minimum and maximum values), etc.

Using the implicit form of the pattern (Figure 6.9), the following equations can be formed for each of the variables listed.

(a) autoregression equations ($k = 1, 2, \cdots, 6$):

$$q^t_{k_{ij}} = f_1(q^{t-1}_{k_{ij}}, q^{t-2}_{k_{ij}}, q^{t-3}_{k_{ij}}, q^t_{k_{i-1j}}, q^t_{k_{i-2j}}, q^t_{k_{ij-1}}, q^t_{k_{ij-2}}); \tag{6.19}$$

(b) multivariate equations ($k = 1, 2, \cdots, 6;\ l = 1, 2, \cdots, 6;\ k \neq l$):

$$\begin{aligned} q^t_{k_{ij}} = {} & f_1(q^{t-1}_{k_{ij}}, q^{t-2}_{k_{ij}}, q^{t-3}_{k_{ij}}, q^t_{k_{i-1j}}, q^t_{k_{i-2j}}, q^t_{k_{ij-1}}, q^t_{k_{ij-2}}) \\ & + f_2(q^t_{l_{ij}}, q^{t-1}_{l_{ij}}, q^{t-2}_{l_{ij}}, q^{t-3}_{l_{ij}}, q^t_{l_{i-1j}}, q^t_{l_{i-2j}}, q^t_{l_{ij-1}}, q^t_{l_{ij-2}}); \end{aligned} \tag{6.20}$$

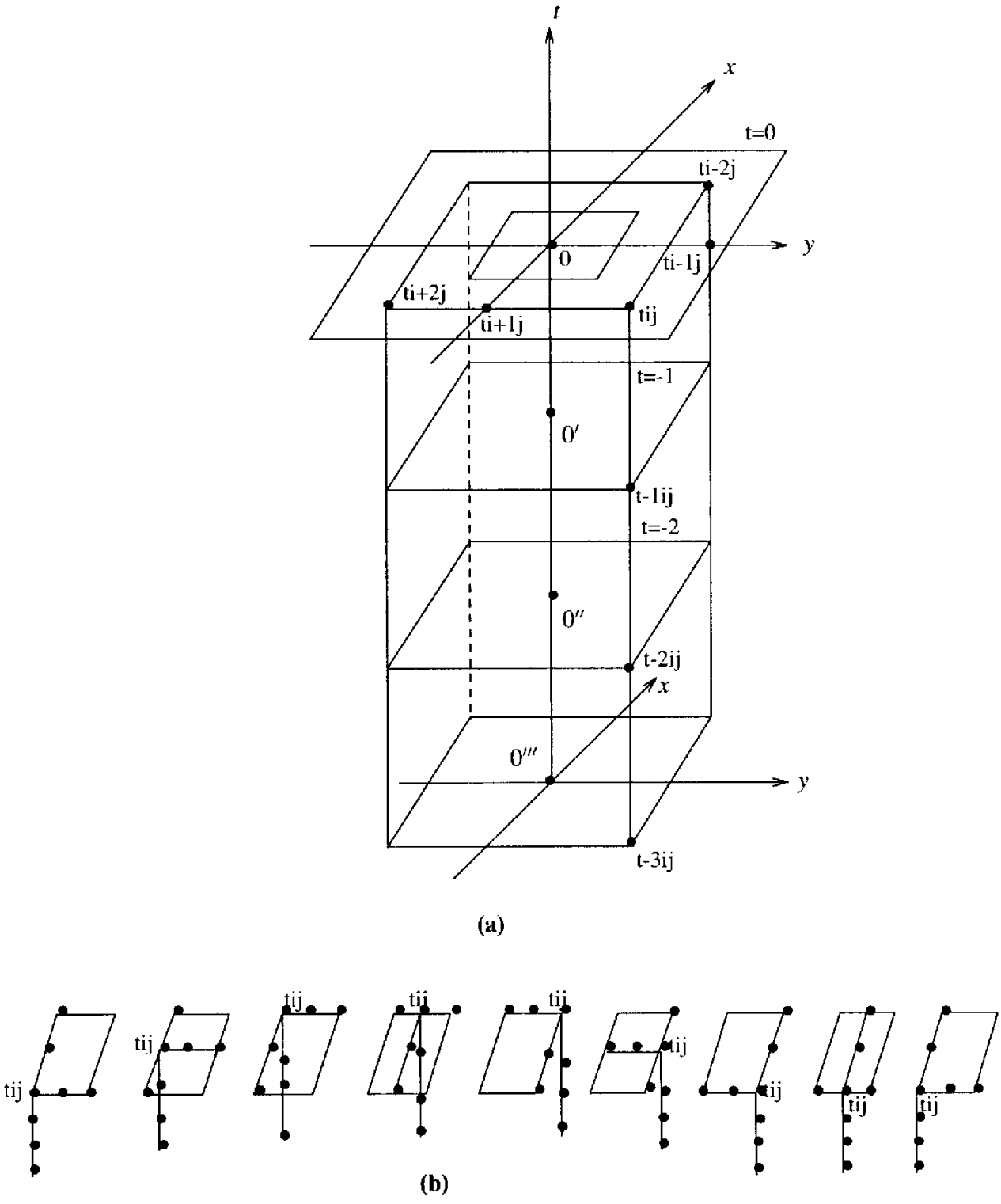

Figure 6.9. (a) implicit pattern and (b) patterns formation for nine simultaneous equations

(c) equations with source functions ($k = 1, 2, \cdots, 6$):

$$q^t_{k_{ij}} = f_1 + f_2 + f_3(t, x, y); \tag{6.21}$$

(d) equations considering the external influences ($k = 1, 2, \cdots, 6$):

$$q^t_{k_{ij}} = f_1 + f_2 + f_3 + f_4(u_1, u_2, \cdots). \tag{6.22}$$

Two-level predictions

In one-level weather predictions, either instantaneous values of the variables or the variables averaged over the same interval of time are used. In two-level predictions, these calculations are made for two time intervals (for example, season and year). The balance criterion enables one to choose a pair of seasonal and yearly predictions. For regularization purposes one uses the balance criteria set up for other time intervals or space regions. The prediction time in the averaging of the variables tends to the prediction time of the variables averaged on a long time interval or a large space region. We recall that balance criteria can be either temporal or spatial.

The descritization step of the data in time is chosen for both the components in relation with the extrema of the corresponding correlation functions.

Two-step algorithm

The two-step algorithm is used for solving the system of difference equations and to obtain the model in optimal complexity.

At the first step of the algorithm, the multilayer algorithm is used with the regularity criterion to obtain only the effective candidate variables and their estimates.

At the second step of the algorithm, the reference function includes the higher ordered arguments (product terms to the order of three) for the effective variables selected in the first step. This uses the multiplicative-additive and non-linear functions. Depending on the number of arguments, either combinatorial or multilayer algorithm is applied with the minimum-bias or regularity criterion to obtain the optimal structure.

With the "implicit" patterns (Figure 6.9), each of them is trained up along the time axis by estimating the coefficients. The output variables are found by solving the simultaneous system of nine equations for all patterns located at the corners of the square. The stability of the step-by-step prediction is increased with this procedure. For each candidate variable the implicit patterns are used. This means that there are nine equations for each variable, and for the external influences and source function, explicit patterns are used to obtain a single equation. Ultimately, an autonomous system of finite-difference equations are obtained in which the outputs are the averaged system variables and external disturbances. The solution of such a system serves as analogues corresponding to find the free motion of some closed system and, hence, it does not need any special orthogonalization such as the use of Chebyshev series. Predictions are obtained by step-by-step integration of the system of equations. Here it is assumed that the external disturbances change in the future as they have in the past.

Weather forecasting in movable coordinates

In meteorological problems, the equations of motion of a cyclone or anticyclone are uncoupled into an equation of motion of the center or a system of equations describing the motion around the center. The center is defined as the point corresponding to the two-dimensional

correlation function formulation. The most accurate prediction is given by a method using the weather satellites' data. The modeling is done by using synoptics and that is why it is subjective and why it depends on the experience of the modeler. For example, in determining the position of the center of a cyclone (about 20 to 40 km/hr) for a prediction time of 24 hrs, the average error in determining the speed of displacement is $\delta_s = 300km/hr$ and the average error in determining the direction of motion is $\delta_\phi = 12°$. This leads to an average error of $\delta_s = 200km$ for the position of the center. This is because the pictures taken from the satellites do not show the wind at high levels when there is no cloudiness. This leads to unexpected motions of the formations [81].

In place of subjective forecasting, one has to use mathematical modeling, particularly the inductive approach which is based on objective reasoning.

Atmospheric formations like cyclones exist for only a short time. This means that only few points of observations are available within that short interval of time. Under such conditions, the inductive algorithms work very efficiently to study the situation. For example, the model that can be obtained from the above two-step algorithm consists of a system of two finite-difference equations (one equation for each coordinate). This can be integrated for step-by-step predictions of several steps ahead.

Prediction of the change in the atmospheric formations

It is convenient to locate the coordinate origin at the center of an atmospheric formation such as the center of a cyclone. This means that the problem is predicting change in the shape of the formation around the coordinate origin. The two-step algorithm is based on the use of finite-difference equations with the implicit form of the patterns. The pressure at a definite point in the x, y- plane which moves together with the center of the cyclone over the surface of the earth, can be predicted using its delayed values and the pressures at the neighboring points located at the corners of the square (Figure 6.9).

$$\begin{aligned} P_{ij}^t &= a_0 + a_1 P_{ij}^{t-1} + a_2 P_{ij}^{t-2} + a_3 P_{ij}^{t-3} + \cdots + b_1 P_{i-1j}^t + b_2 P_{i+1j}^t \\ &= f_1(P_{ij}^{t-1}, P_{ij}^{t-2}, P_{ij}^{t-3}, \cdots, P_{i-1j}^t, P_{i+1j}^t), \end{aligned} \tag{6.23}$$

where the time axis t is located at the center of the square as shown.

Use of correlational models

The isobars shown on the meteorological charts can be considered as random functions in the space coordinates and time. They can also be represented as a two-dimensional spectrum and a two- dimensional correlation function of the surface.

We assume that the self-organizing correlation function (its numerator) is stable; i.e., it holds the same characteristic in the prediction region as in the interpolation region. This condition enables us to obtain an optimal nonphysical nonlinear model according to the combined criterion of minimum-bias plus prediction.

The advantage of the correlation models is that they can become multifactor models in a simple manner and that they take into account several meteorological variables and their delayed arguments.

Use of graphs with binary transformation [55]

The problem of predicting the shape of a cyclone (in mobile coordinates with its center) can be simplified by considering the shape of isobar curve as a representative of binary

discretization of the pressure (for example, for $P = 1000$ mm, the pressure in the region close to the center (Figure 6.9) is taken equal to $P = -1$ and outside the isobar it is $P = +1$). Such a binary approach decreases the requirements on the number of observations and increases the prediction accuracy and its time.

The algorithm consists of three parts: (i) training on the graphs of prehistory, (ii) adaptation of the graphs as per the predictions (ensuring the stability of step-by-step predictions), and (iii) selection of the best graph according to the criteria.

The field is partitioned into small squares in which the pressure is either $+1$ or -1. Four adjacent squares provide a single input for the graph. The training is conducted by calculating the number of transitions of the output variables to $+1$ and -1 (Figure 6.10). The adaptation is done by changing the graphs with each step in the prediction such that the number of transitions on the entire prediction interval is equal to their number in the observation interval; accordingly, the number of transitions in the graph is decreased by one at each step.

The graphs can be considered with the preceding states (the last value or the last two or three values of the pressure in time). All the graphs are used for predictions and the best one is chosen according to the combined criterion.

Use of ecological variables

Usually, prediction of climate involves the prediction of variables measured with a large siding time averaging or moving average interval. Many ecological variables result from averaging different influences activating in the process—first, meteorological variables; hence, there is considerable correlation between the ecological and the meteorological variables. One can refer to the work of Lebow et al. [80] for such practical examples.

The possibility of using ecological variables for predicting weather is debated. The objective system analysis algorithm often yields a set of significant variables that include both ecological and meteorological variables. For example, a model obtained for the ecosystem of lake Baykal includes the variables (yearly average values): q_1—the water transparency, q_3—the biomass of the plant life in the water, q_4—the biomass of zooplankton, and u_1—the water temperature at the surface layer. These are effective output variables in studying the ecosystem. Corresponding models are obtained for each of these variables. They are used for step-by-step predictions to the year 2000 to study the changes in the system. This means that one meteorological variable like water temperature can be predicted by using the ecological data.

The equation obtained for the temperature in the system is not a physical law, but is merely a tool for predicting the temperature. For extrapolation or prediction of a variable we can use the selected model, whether it be a finite-difference equation or an algebraic equation treated in the same way. For example, Figure 6.11 shows the autocorrelation function of temperature and the cross-correlation function of temperature and biomass of zooplankton. It shows the presence of considerably higher frequencies in the temperature than in the process of interaction between the temperature and the biomass of zooplankton. The correlation time of the remainder of the first process is less than that of the second. Hence, the prediction interval of the variable temperature when using an ecological variable is greater because the limiting attainable prediction interval is proportional to the correlational interval. This means that the ecological variables help in the process of predicting meteorological variables by increasing accuracy and the prediction interval.

In this example, the difference equations used for the ecological variables in predicting the temperature are treated as approximations of their variation. Similarly, the difference equation used for temperature in predicting the ecological variables should be treated as an approximation of its variation with time.

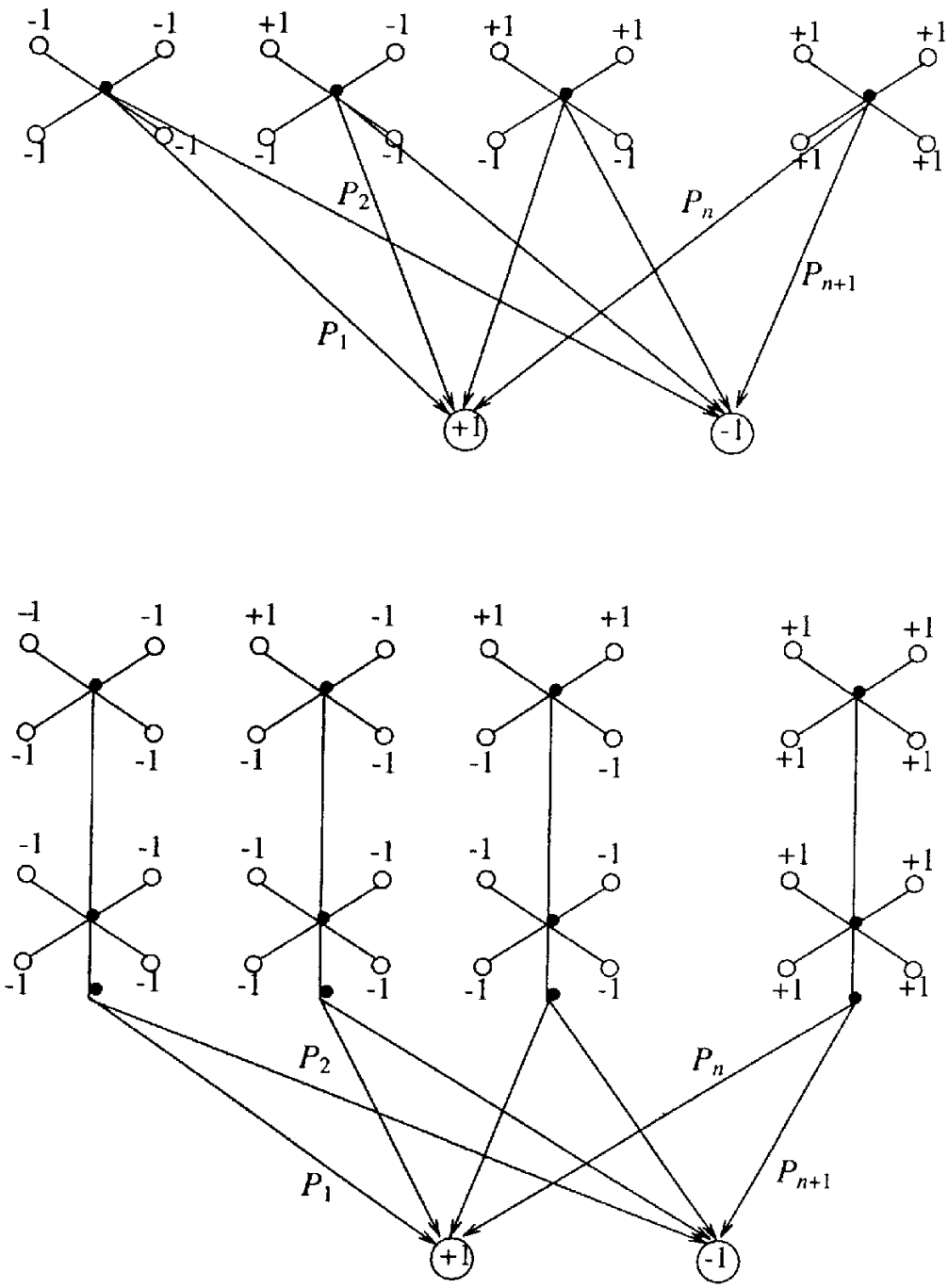

Figure 6.10. Graphs of the number of transitions: upper graph with allowance only for the last state of the field and the lower graph with allowance for the last several states of the field

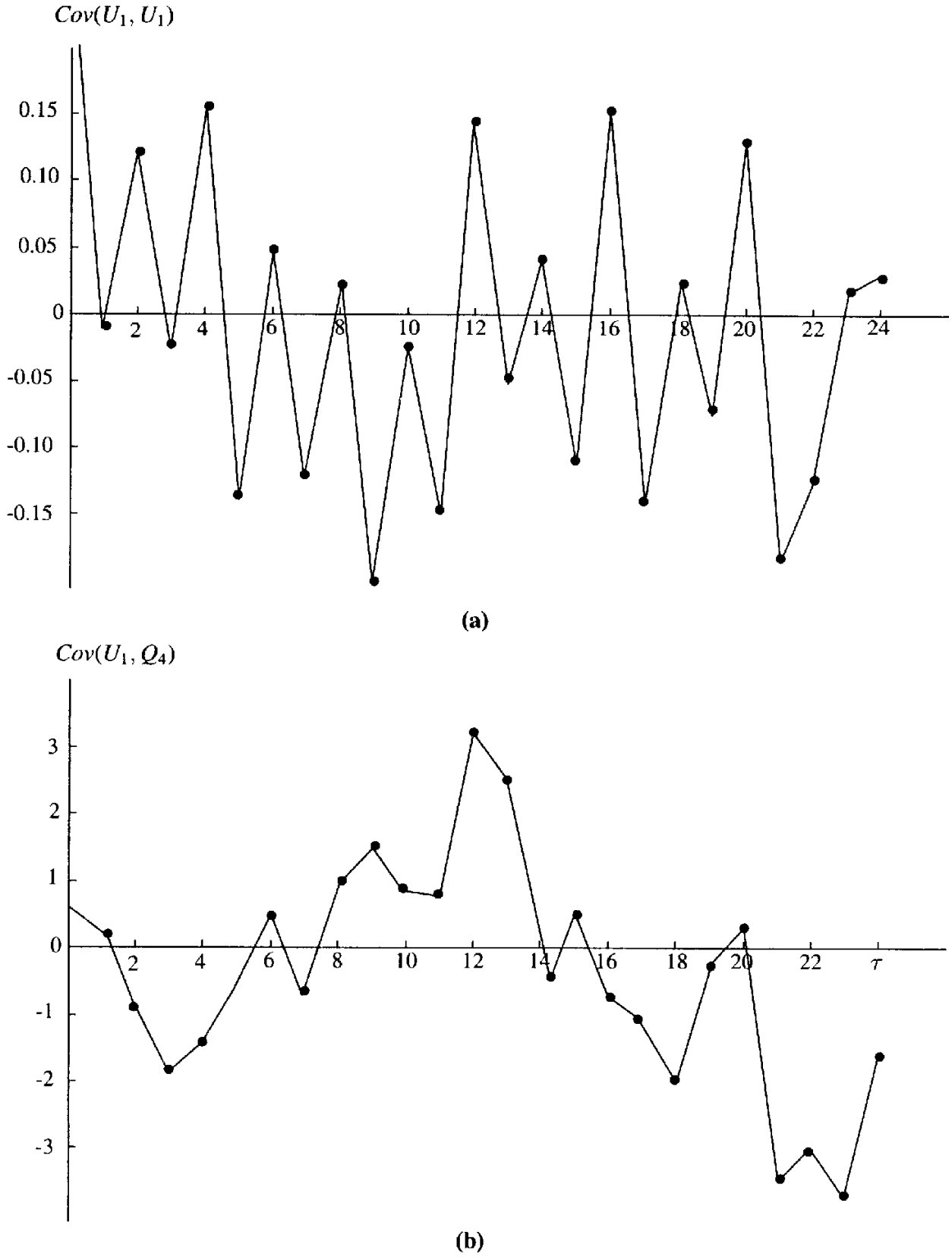

Figure 6.11. (a) autocorrelation function for the temperature U_1 and (b) cross-correlation function of temperature U_1 and biomass Q_4

Use of objective clustering

Let us suppose that the computer has selected four out of six variables to use in the system of autonomous equations listed for characterizing the weather. Any point in the four-dimensional coordinate space represents the weather at a given time. Similar points constitute a cluster of "good" or "bad," "dry" or "rainy" weather. The number of clusters can be found by self-organization clustering with an algorithm of objective clustering of multidimensional space into an unknown number of clusters. A more accurate and detailed representation of the input information helps to find more weather clusters. Here the threshold values for the clusters are not necessarily specified.

Method of group analogues

According to the method of group analogues, when using prehistory data one has to find several situations of the weather (in some interval of time) that is very similar to the present weather conditions under consideration. The duration of a situation is equal to the correlation interval of the anamoly. Then, prediction of the weather can be obtained by averaging all similar situations. This is the foundation for contemporary long-range prediction.

The selection of an analogue for a given synoptic situation or pattern is a common operation in synoptic practice, particularly in long-range weather predictions. To avoid the subjective choice of thresholds for distinguishing situations, one can entrust this operation to the objective clustering algorithm. Here the more extensive prehistory data the computer can examine, the closer to each other will be the situations associated with a single cluster. However, there is a limit to the length of the data sample. Beyond that limit the number of clusters does not increase. Thus, the inductive approach can help in comparing many candidate cases and in finding the optimal number of clusters which can help predict the near future. Usually, the computer chooses several noncontradictory clusterings differing from each other in the number of clusters and in the set of variables. It is expedient to choose from these a single clustering which corresponds to the longest correlation interval. This can be done by further regularization as described in the previous chapter.

Long-range weather prediction

There are at least five different types of atmospheric circulations known in the northern hemisphere. If one type of circulation exists more often than the others, that is called the prevailing circulation. The change from one type of circulation to another is a purely random process and is not subject to prediction (like the result of flipping a coin—"head" or "tail"). However, the change of prevailing circulation from one type to another does lend itself to prediction like any averaged variable, because the correlation function of the process of change of prevailing circulation must be rather broad. Each cluster of the non-contradictory clustering enables us to obtain a long-range weather prediction.

3 ECOLOGICAL SYSTEM STUDIES

Here the performance of the multilevel algorithm along with the objective systems analysis described in the second chapter will be demonstrated in the study of ecological systems.

Let us briefly discuss various stages of this algorithm.

The *first stage* is to divide the set of variables into three subsets: the output variables, the input variables, and variables which have no substantial effect on the subsets. This is the first level of the multilevel algorithm (also called as objective system analysis) used in detecting the relationships among the variables.

The next *two stages* belong to the two-level analysis of the algorithm; the purpose of the *first level* is to divide the set of predictions of the average annual values of the variables (those not discarded during the first stage) into "good," "satisfactory," and "unsatisfactory" predictions and to select the best predictions (one for each variable). The purpose of the *second level* is to predict the average seasonal values of the output variables on the basis of a series of sets of seasonal models.

3.1 Example—ecosystem modeling

Example 2. Self-organization modeling in the Lake Baykal ecological system.

Lake Baykal was thought to be exposed to so-called anthropogenic perturbations because of industrial waste, tourism, etc. The views of scientists on this are diversified. Finding an objective method for predicting the condition of the lake is desirable. The inductive learning methods based on the principle of self-organization are good candidates for exploring the objective characteristics of the system.

The list of possible variables is given by biologists. The seasonal and annual values of the following parameters for a 23-year period are used in this example;

q_1 – the transparency of the water in meters,
q_2 – the biomass of the phytoplankton in mg/m^3,
q_3 – the biomass of the small plants(*Melosira*)in mg/m^3,
q_4 – the biomass of the zooplankton in mg/m^3,
q_5 – the biomass of the *epischura* in g/m^2,
u_1 – the surface water temperature in$^\circ C$,
u_2 – the water level in meters, and
u_3 – the number of hours of sunlight.

Small letters denote the seasonal values and big letters denote the mean annual values.

Here the problem identifies a point physical model that represents the ecosystem. This is solved by using the multilevel iterative algorithm which has the levels of the objective system analysis to identify the characteristic variables of the system and the two-level scheme to select two nonphysical models (annual and seasonal) for long-range quantitative predictions.

Objective system analysis

This level is used to synthesize a model in the form of systems of from one to five equations. The primary variables are used to form the polynomials in the form of finite-difference equations.

First layer. The finite-difference models of the form given below are used for each of the five variables ($q_1 - q_5$).

$$\begin{aligned} q_i^t = {} & a_0 + a_1 u_1^t + a_2 u_1^{t-1} + a_3 u_1^{t-2} + a_4 u_1^{t-3} + a_5 u_1^{t-4} \\ & + a_6 u_2^t + a_7 u_2^{t-1} + a_8 u_2^{t-2} + a_9 u_2^{t-3} + a_{10} u_2^{t-4} \\ & + a_{11} u_3^t + a_{12} u_3^{t-1} + a_{13} u_3^{t-2} + a_{14} u_3^{t-3} + a_{15} u_3^{t-4} \\ & + a_{16} q_i^{t-1} + a_{17} q_i^{t-2} + a_{18} q_i^{t-3} + a_{19} q_i^{t-4}. \end{aligned} \tag{6.24}$$

Second layer. At the second layer, there are $C_5^2(=10)$ systems of two equations of the form as given below:

$$\begin{aligned} q_i^t = {} & a_0 + a_1 u_1^t + a_2 u_1^{t-1} + a_3 u_1^{t-2} + a_4 u_1^{t-3} + a_5 u_1^{t-4} \\ & + a_6 u_2^t + a_7 u_2^{t-1} + a_8 u_2^{t-2} + a_9 u_2^{t-3} + a_{10} u_2^{t-4} \\ & + a_{11} u_3^t + a_{12} u_3^{t-1} + a_{13} u_3^{t-2} + a_{14} u_3^{t-3} + a_{15} u_3^{t-4} \\ & + a_{16} q_i^{t-1} + a_{17} q_i^{t-2} + a_{18} q_i^{t-3} + a_{19} q_i^{t-4} \\ & + a_{20} q_j^t + a_{21} q_j^{t-1} + a_{22} q_j^{t-2} + a_{23} q_j^{t-3} + a_{24} q_j^{t-4}. \end{aligned} \tag{6.25}$$

In the *third and fourth layers*, $C_5^3(=10)$ systems of three equations and $C_5^4(=5)$ systems of four equations are used correspondingly. In the *fifth layer*, a single system of five equations which contain 40 terms is used;

$$\begin{aligned} q_i^t = {} & a_0 + a_1 u_1^t + a_2 u_1^{t-1} + a_3 u_1^{t-2} + a_4 u_1^{t-3} + a_5 u_1^{t-4} \\ & + a_6 u_2^t + a_7 u_2^{t-1} + a_8 u_2^{t-2} + a_9 u_2^{t-3} + a_{10} u_2^{t-4} \\ & + a_{11} u_3^t + a_{12} u_3^{t-1} + a_{13} u_3^{t-2} + a_{14} u_3^{t-3} + a_{15} u_3^{t-4} \\ & + a_{16} q_i^{t-1} + a_{17} q_i^{t-2} + a_{18} q_i^{t-3} + a_{19} q_i^{t-4} \\ & + a_{20} q_j^t + a_{21} q_j^{t-1} + a_{22} q_j^{t-2} + a_{23} q_j^{t-3} + a_{24} q_j^{t-4} + \\ & \cdots \\ & + a_{35} q_l^t + a_{36} q_l^{t-1} + a_{37} q_l^{t-2} + a_{38} q_l^{t-3} + a_{39} q_l^{t-4}. \end{aligned} \tag{6.26}$$

The computational volume can be reduced at the higher levels by discarding the terms that are not effective at the preceding layers.

The F best models are selected at each layer by using the system criterion of the minimum bias. In this selection, a system with at least one equation having $\eta_{bs} \geq 0.01$ is eliminated from the sorting. The optimal model is selected according to the step-by-step integrated prediction accuracy of the prediction criterion on the total points N.

From the above analysis, the following system of equations containing the variables q_3 and q_4 is obtained as the optimal one with the limit of $\eta_{bs_s} \leq 0.005$.

$$\begin{aligned} q_3^t = {} & 22.4042 - 10.0977 u_1^{t-3} + 1.8842 u_2^{t-3} \\ & - 2.4647 u_2^{t-3} - 0.1673 q_4^{t-1} \\ q_4^t = {} & 58.9093 + 0.0233 u_1^{t-2} - 0.1382 u_2^{t-1} \\ & - 0.0425 u_2^{t-2} - 0.0325 u_2^{t-4} - 0.0132 u_3^{t-1} \\ & - 0.0386 u_3^{t-3} - 0.0062 q_1^{t-4} + 0.0006 q_4^{t-4}. \end{aligned} \tag{6.27}$$

The characteristic vector of the system includes the variable q_1 along with the variables q_3 and q_4.

Two-stage scheme

There are different ways for solving the two-stage scheme of long-range quantitative predictions of this problem [48], [50]. The latter work is proven to be the best heuristic approach for this scheme. We present here both approaches for giving an idea of using different heuristics.

First approach: Examining the above system of equations, one can easily establish that the vector of output variables consists of q_3 and q_4, and the vector of inputs is q_1, u_1, u_2 and u_3. The variables q_2 and q_5 are excluded from further consideration.

The two-stage scheme comprises of identifying the average annual models and the system of seasonal models and the optimal system of equations to be selected from the both using the prediction balance criterion.

In the *first stage*, for identifying the average annual models for the variables Q_3, Q_4, Q_1, U_1, U_2 and U_3, the harmonical and the single layered combinatorial algorithms are used. The harmonical algorithm is recommended when a large number of data points are available, while the combinatorial algorithm is recommended for a small number of input variables. The finite-difference scheme with the delayed arguments considered in the combinatorial algorithm is

$$Q_{i_{(T)}} = a_0 + a_1 Q_{i_{(T-1)}} + a_2 Q_{i_{(T-2)}} + \cdots + a_m Q_{i_{(T-m)}}, \tag{6.28}$$

where T denotes the number of the year.

Harmonic models have shown better performance than the finite-difference models obtained from the combinatorial algorithm for the variables Q_1, Q_4, and U_1 with five, ten, and eight harmonic components in the trends, correspondingly. The predictive models obtained for the variables Q_3, U_2, and U_3 are insufficiently accurate; variables U_2 and U_3 are excluded from the future consideration. Along the four variables Q_3, Q_4, Q_1, and U_1 considered, the variable Q_4 is called the leading variable because of its better annual predictions.

In the *second stage*, the combinatorial algorithm is used to obtain the seasonal models for the variables q_3 and q_4 having the following type reference functions:

$$\begin{aligned} x_{1_{(t,T)}} = {} & a_0 + a_1 x_{1_{(t-1,T)}} + a_2 x_{1_{(t-2,T)}} + a_3 x_{1_{(t-3,T)}} + a_4 x_{1_{(t,T-1)}} \\ & + a_5 x_{2_{(t,T)}} + a_6 x_{2_{(t-1,T)}} + a_7 x_{2_{(t-2,T)}} + a_8 x_{2_{(t-3,T)}} + a_9 x_{2_{(t,T-1)}} \\ & + a_{10} X_{2_{(T)}} + a_{11} X_{2_{(T-1)}} + a_{12} X_{3_{(T)}} \\ & + a_{13} X_{3_{(T-1)}} + a_{14} U_{1_{(T)}} + a_{15} U_{1_{(T-1)}}, \end{aligned} \tag{6.29}$$

where t and T denote the season and year; $x_1 = q_3$, $x_2 = q_4$; $X_{2_{(T)}} = Q_{4_{(T)}}$, $X_{3_{(T)}} = Q_{1_{(T)}}$ and $U_{1_{(T)}}$ are the average annual values at the year T.

Similarly, the reference function for the variable x_2 is considered. The reference functions can be expanded further with the trend equation of two-dimensional time read-out and with the variables of $x_3 = q_1$ and u_1 and with their delayed values according to the data points and allotted computer time. The equations containing the variables $x_{1_{(t,T)}}$ and $x_{2_{(t,T)}}$ means that during the step-by-step predictions both equations must be integrated jointly—the system of two equations use their estimated values.

Here, for each season five best models are selected for the leading variable x_2 and one model for the variable x_1 according to the criteria minimum-bias and prediction. Using the prediction balance, the optimal system of equations is selected for each season. This is done by using the seasonal models of the output variables one after the other in step-by-step predictions.

The balance-of-predictions criterion is used for selecting the system of seasonal equations for variable x_2 on the total data sample.

$$b_i = \frac{1}{4}(x_{2_{(w)}} + x_{2_{(sp)}} + x_{2_{(su)}} + x_{2_{(f)}})_i - X_{i2_{(yr)}}$$

$$c_i = \frac{1}{4}(x_{2_{(w)}} + x_{2_{(sp)}} + x_{2_{(su)}} + x_{2_{(f)}})_i + X_{i2_{(yr)}}$$

$$B = \sum_{i=1}^{N} b_i^2 \Big/ \sum_{i=1}^{N} c_i^2 \to \min, \tag{6.30}$$

where $x_{2(w)}, x_{2(sp)}, x_{2(su)}$, and $x_{2_{(f)}}$ are the predicted values of the variable x_2 for winter, spring, summer, and autumn, correspondingly; and $X_{i2_{(yr)}}$ is the predicted average annual value of the variable X_2. There are a total of 5^4 (= 625) formations of system of equations evaluated for their predictions. The optimal system which is found to be better has the value of the criterion $B_{x_2min} = 0.06$.

The balance of prediction criterion for both the variables x_1 and x_2 is evaluated using the criterion as a system criterion

$$B^* = \sqrt{(B_{x_1}^2 + B_{x_2}^2)}, \tag{6.31}$$

where B_{x_1} is computed the same way as the variable x_2, but only on the interpolation interval.

Second approach: In the *first stage*, an algorithm similar to the objective systems analysis is used for sorting the systems of equations only for those variables Q_1, Q_3, and Q_4 that appear in the characteristic vector. The trend component is included into the equations. The complexity of the models is increased by replacing the addition of polynomials in the right side with multiplication. This means that it indicates switching to nonlinear equations or equations with variable coefficients.

Three layers are necessary for comparison because of the three variables. The multilayer algorithm is used to select the best models at each layer because of the large number of input variables.

The first layer consists of three equations with $9 \times 5 = 45$ terms in the right side of each equation.

$$\begin{aligned} Q_{i_{(T)}} = {} & (a_0 + a_1 T + a_2 T^2 + a_3 U1_{(T)} + a_4 U1_{(T-1)} \\ & a_5 U2_{(T)} + a_6 U2_{(T-1)} + a_7 U3_{(T)} + a_8 U3_{(T-1)}) \\ & .(1 + a_9 Q_{i_{(T-1)}} + a_{10} Q_{i_{(T-2)}} + a_{11} Q_{i_{(T-3)}} + a_{12} Q_{i_{(T-4)}}). \end{aligned} \tag{6.32}$$

The second layer consists of three systems of two equations containing the $8 \times 3 \times 3 = 72$ terms in the left side of each equation;

$$\begin{aligned} Q_{i_{(T)}} = {} & (a_0 + a_1 T + a_2 U1_{(T)} + a_3 U1_{(T-1)} \\ & a_4 U2_{(T)} + a_5 U2_{(T-1)} + a_6 U3_{(T)} + a_7 U3_{(T-1)}) \\ & .(1 + a_8 Q_{i_{(T-1)}} + a_9 Q_{i_{(T-2)}}) \\ & .(a_{10} Q_{j_{(T)}} + a_{11} Q_{j_{(T-1)}} + a_{12} Q_{j_{(T-2)}}). \end{aligned} \tag{6.33}$$

The third layer is a single system of three equations, each of which contains $8 \times 3 \times 3 \times 3 = 216$ terms on its right side;

$$\begin{aligned} Q_{i_{(T)}} = {} & (a_0 + a_1 T + a_2 U1_{(T)} + a_3 U1_{(T-1)} \\ & a_4 U2_{(T)} + a_5 U2_{(T-1)} + a_6 U3_{(T)} + a_7 U3_{(T-1)}) \\ & .(1 + a_8 Q_{i_{(T-1)}} + a_9 Q_{i_{(T-2)}}) \\ & .(a_{10} Q_{j_{(T)}} + a_{11} Q_{j_{(T-1)}} + a_{12} Q_{j_{(T-2)}}) \\ & .(a_{13} Q_{l_{(T)}} + a_{14} Q_{l_{(T-1)}} + a_{15} Q_{l_{(T-2)}}). \end{aligned} \tag{6.34}$$

If the total number of data points permits, one can introduce the moving average terms as inputs into the equations. Here another difference from the objective systems analysis is that the optimal model is chosen using the system criterion of prediction based on its step-by-step prediction accuracy. If the prediction error of any equation of a system is above 5%, then that system is eliminated from further sorting.

The optimal system obtained has two equations for the variables $Q_{3_{(T)}}$ and $Q_{4_{(T)}}$.

The predictions of the annual values of the variables Q_3 and Q_4 are obtained by step-by-step integration of the optimal system of equations obtained above. Stability of the integration is ensured by the explicit form of the pattern used; this means that the delayed arguments identical to the output variable used in the equations. In doing the step-by-step predictions, the future values of the external influences U_1, U_2, and U_3 are needed. Harmonical algorithm is used successfully to obtain the optimal harmonical trends of these variables. The predictions using the harmonical trends which have the prediction errors ≤ 0.05, are used for step-by-step integration of the system of equations. If the harmonical trends of the external influences are poorly predicted (> 0.05), then they are treated along with the variables Q_3 and Q_4 and the order of the systems of equations is increased with those external influences.

In this example U_1 has achieved better predictions with the harmonical model with the norm (≤ 0.05). The other two disturbances U_2 and U_3 could not achieve the norm. The system of difference equations in the output variables are $Q_{3_{(T)}}, Q_{4_{(T)}}, U_{2_{(T)}}$, and $U_{3_{(T)}}$. Harmonic predictions of U_1 and step-by- step predictions of U_2 and U_3 helped in predicting the Q_3 and Q_4 to the year 2000. The variable Q_4 is chosen as the leading variable because of its better performance than the other two output variables and the best F =10 models for the leading variable are selected by using the prediction criterion.

When we are sure that the statistical characteristics of the predictions are stable, we can use the statistical criteria of stability of moments or the stability of correlations to select the best F =3 models out of the 10 selected models of the leading variable.

At the *second stage* the nonphysical seasonal predictive models are identified. Here the two-dimensional time-read out is used in constructing the models and the arguments of the models include both seasonal and annual values. The objective systems analysis is used with a general form of the additive type descriptions of the three output variables q_1, q_3, and q_4. The two-dimensional time trend is considered in the descriptions.

In the *first layer*, there are three first-order systems, one for each output variable with 18 terms on the right side of the quation,

$$\begin{aligned} q_{i_{(t,T)}} = {} & a_0 + a_1 t + a_2 T + a_3 t^2 + a_4 T^2 + a_5 u_{1_{(t,T)}} + a_6 U_{1_{(T)}} \\ & + a_7 u_{2_{(t,T)}} + a_8 U_{2_{(T)}} + a_9 u_{3_{(t,T)}} + a_{10} U_{3_{(T)}} \\ & + a_{11} Q_{i_{(T)}} + a_{12} q_{i_{(t-1,T)}} + a_{13} q_{i_{(t-2,T)}} + a_{14} Q_{i_{(T-1)}} \\ & a_{15} q_{i_{(t,T-1)}} + a_{16} q_{i_{(t,T-2)}} + a_{17} Q_{i_{(T-2)}}, \end{aligned} \tag{6.35}$$

where $i = 1$, 3, and 4; and t and T denote the season and year, correspondingly.

In the *second layer*, there are three systems of second-order equations with 26 terms on the right side of each equation as shown below:

$$\begin{aligned} q_{i_{(t,T)}} = {} & a_0 + a_1 t + a_2 T + a_3 t^2 + a_4 T^2 + a_5 u_{1_{(t,T)}} + a_6 U_{1_{(T)}} \\ & + a_7 u_{2_{(t,T)}} + a_8 U_{2_{(T)}} + a_9 u_{3_{(t,T)}} + a_{10} U_{3_{(T)}} \\ & + a_{11} Q_{i_{(T)}} + a_{12} q_{i_{(t-1,T)}} + a_{13} q_{i_{(t-2,T)}} + a_{14} Q_{i_{(T-1)}} \\ & a_{15} q_{i_{(t,T-1)}} + a_{16} q_{i_{(t,T-2)}} + a_{17} Q_{i_{(T-2)}} \\ & + a_{18} q_{j_{(t,T)}} + a_{19} Q_{j_{(T)}} + a_{20} q_{j_{(t-1,T)}} + a_{21} q_{j_{(t-2,T)}} + a_{22} Q_{j_{(T-1)}} \\ & a_{23} q_{j_{(t,T-1)}} + a_{24} q_{j_{(t,T-2)}} + a_{25} Q_{j_{(T-2)}}, \end{aligned} \tag{6.36}$$

where $i, j = 1, 3$, and 4; $i \neq j$. Similarly, one can write another equation of the system for the output variable $q_{j_{(t,T)}}$.

Consecutively, the third layer has a single system of three equations with 34 terms in each equation.

The best models are selected for each season separately in the form system of equations by using the prediction criterion. The seasonal predictive models for the external disturbances are identified as it was in the case of annual models. Using the prediction criterion, the best F =10 systems of equations are selected for the leading variable q_4.

Next, the complete sorting of the annual and seasonal models are made and evaluated by the balance-of-predictions criterion,

$$b_i^2 = [Q_{(yr)} - \frac{1}{4}(q_{(w)} + q_{(sp)} + q_{(su)} + q_{((f)})]_i^2;$$

$$B^2 = \sum_{i=N_1}^{N_2} b_i^2 \rightarrow \min, \tag{6.37}$$

where b_i is the balance relation for the ith year and B is the total balance on the interval N_1 to N_2.

The overall sorting on the selected models from the annual as well as seasonal models involves 3×10 predictions. The optimal system of equations for the leading variable q_4 is selected according to the balance of predictions criterion.

A measure of success of the modeling is the global minimum achieved on the balance criterion; the error must not exceed 5%. If it exceeds, the freedom-of-choice of models at both the stages needs to be broadened to increase the volume of sorting and the divergence can also be reduced by including more input variables in the descriptions.

3.2 Example—ecosystem modeling using rank correlations

Example 3. Modeling of ecosystems of Kakhovka and Kremenchug reservoirs using the rank correlations [62], [66].

The objective systems analysis (OSA) used in the multilevel iterative algorithm which is described in the above example serves for singling out the least biased systems of equations by establishing the relations of the modeled object. This determines the set of characteristic variables of the object and its simplest physical model which is suitable for a short-range forecast. This set is used to find a nonlinear model of the object which improves the forecast. In case of a large amount of data, it would be easy to take into account more delayed values of the arguments, and to obtain dynamic and nonlinear models.

The best structures of the individual equations are obtained by using the single-layered combinatorial algorithm. Although the inductive learning algorithms are developed for use with a very short data samples, there are limitations in using the single-layered combinatorial algorithm. Experiments spread over time and experiments spread over space are assumed to be equivalent to increase the number of conditional equations. This means that the observations that are spread in space (over the stations) and the observations that differ in season number possess equal validity. The variables of the object are selected from the general set of variables indicated at the start of the investigation of the hydrobiological experimentations and are usually very large in number. If all the variables are considered along with their delayed arguments in space and time, it increases the computational volume of the combinatorial algorithm. A rank correlation method is used to classify the initial variables. A modified version of the OSA algorithm allows one to eliminate the arguments whose rank correlation is of modulus unity. The equations with high collinearity between the arguments are eliminated and the ill-conditioned matrices are avoided.

The *modified objective systems analysis* was used independently for analysing the ecological systems of Kakhovka and Kremenchug reservoirs. Here we present briefly the general procedure used in the analysis. The averaged seasonal values are calculated from the data obtained at six stations in the middle and lower part of the reservoirs. The initial data obtained on the avearge seasonal values of 37 state variables and four external influences in spring, summer and fall of a year has of 18 years. The following notations are used for some variables;

x_2 – chromaticity (degrees),
x_6 – oxygen content (mg/liter),
x_8 – bichromatic oxidizability (mg/liter),
x_9 – suspended particles (mg/liter),
x_{11} – nitrates (mg/liter),
x_{12} – nitrites (mg/liter),
x_{16} – soluble iron (mg/liter),
x_{17} – total phosphorus (mg/liter),
x_{21} – zinc (μg/liter),
x_{25} – organic phosphorus (mg/liter),
x_{27} – biomass of blue-green algae (mg/liter),
x_{30} – number of phytoplankton (millions of cells/liter),
x_{31} – number of blue-green algae (millions of cells/liter),
u_1 – water temparature (°C), and
u_2 – effective sun's energy

For calculating the rank correlation, all seasonal mean observations are given three values of rank (1 for the minimum value, 2 for the medium value, and 3 for the maximum of the three values). The paired rank correlation coefficients are determined from the Spearman's formula [72],

$$\rho_{x_j,x_k} = 1 - \frac{6S(d_i^2)}{N^3 - N}, \tag{6.38}$$

where $S(d_i^2) = \sum_{i=1}^{N} d_i^2$, d_i is the difference of ranks of the corresponding pairs of observations of the variables x_j and x_k, and N is the number of observations. All variables can be divided into three groups (classes I, II and III) with respect to the existing three points (spring, summer, and fall).

The delayed values of the arguments are not taken into account because of the small amount of data. According to the modified OSA, the complete equations from which the sorting of systems of equations is generated belong to one of the following forms:

$$\begin{aligned} x^I &= a_0 + a_1 x^{II} + a_2 x^{III} + a_3 u_1 + a_4 u_2 \\ x^{II} &= b_0 + b_1 x^I + b_2 x^{III} + b_3 u_1 + b_4 u_2 \\ x^{III} &= c_0 + c_1 x^I + c_2 x^{II} + c_3 u_1 + c_4 u_2, \end{aligned} \tag{6.39}$$

where x^I, x^{II}, and x^{III} are the variables from the first, second, and third classes, respectively.

The single-layered combinatorial algorithm is used with the symmetrical criteria of η_{bs} and $\Delta(C)$ to select the set of possible system of equations with various composition of the output variables. For example, the following optimum system of equations is obtained in case of Kremenchug reservoir.

$$\begin{aligned} x_2 &= 78.0 - 2.57u_1, \\ x_6 &= 17.0 - 0.4u_1, \\ x_{30} &= 12400 - 166x_2; \quad \eta_{bs} = 0.006, \end{aligned} \tag{6.40}$$

where x_2, x_6, and x_{30} belong to the classes I, II, and III respectively. We use the notation $x^I = x_2$, $x^{II} = x_6$, and $x^{III} = x_{30}$ further.

Based on the prediction errors, the variable X^{III} is chosen as the leading variable which is predicted better than the other two.

The two-stage scheme. In constructing the forecating model of mean annual values, the following linear form of the complete equation is used;

$$\begin{aligned} X^I_{(T)} = {} & a_0 + a_1 X^I_{(T-1)} + a_2 X^I_{(T-2)} + a_3 X^{II}_{(T)} + a_4 X^{II}_{(T-1)} + a_5 X^{II}_{(T-2)} \\ & + a_6 X^{III}_{(T)} + a_7 X^{III}_{(T-1)} + a_8 X^{III}_{(T-2)} + a_9 U_{1_{(T)}} + a_{10} U_{2_{(T)}} + a_{11} T, \end{aligned} \tag{6.41}$$

where upper-case letters denote the mean annual variables; T is time in years; and the equations for $X^{II}_{(T)}$ and $X^{III}_{(T)}$ are analogous.

These three equations are integrated step-by-step jointly to obtain the yearly forecast.

In obtaining the seasonal average forecasting models at the second stage, the following complete descriptions are used.

$$\begin{aligned} x^I_{(t,T)} = {} & a_0 + a_1 x^I_{(t-1,T)} + a_2 x^I_{(t-2,T)} + a_3 x^I_{(t,T-1)} + a_4 X^I_{(T)} + a_5 x^{II}_{(t-1,T)} \\ & + a_6 x^{II}_{(t-2,T)} + a_7 x^{II}_{(t,T)} + a_8 x^{II}_{(t,T-1)} + a_9 X^{II}_{(T)} + a_{10} x^{III}_{(t-1,T)} + a_{11} x^{III}_{(t-2,T)} \\ & + a_{12} x^{III}_{(t,T)} + a_{13} x^{III}_{(t,T-1)} + a_{14} X^{III}_{(T)} + a_{15} u_1 + a_{16} u_2, \end{aligned} \tag{6.42}$$

where t and T denote the season and year, correspondingly; equations for $x^{II}_{(t,T)}$ and $x^{III}_{(t,T)}$ are written analogously.

The system of mean seasonal equations are integrated season by season to obtain the forecast.

For example, the following system of equations for annual model of Kremenchug reservoir is obtained as one of the best:

$$\begin{aligned} X^I_{(T)} &= -18.0 + 0.00147 X^{III}_{(T)} + 5.58 U_1, \\ X^{II}_{(T)} &= 7.78 + 0.284 X^{II}_{(T-1)} - 0.00052 X^{III}_{(T-1)}, \\ X^{III}_{(T)} &= -817 - 67.5 X^{II}_{(T)} + 214 U_1 + 0.811 T. \end{aligned} \tag{6.43}$$

Similarly four more sytems of equations are selected for future evaluation; i.e., $F = 5$. The values of the external disturbances are taken according to the given scenario.

By the combinatorial algorithm, the seasonal forecasting models for the reservoirs are found; there are three equations in the system for each season. A total of five systems of equations ($F = 5$) are chosen for each season. The best one among them is given below for the Kremenchug reservoir:

(i) winter:

$$\begin{aligned} x^I_{(t,T)} &= 0.00926 x^{III}_{(t-1,T)} - 2.59 x^{II}_{(t-1,T)} + 0.729 x^I_{(t,T-1)} + 1.15 X^I_{(T)} - 26.1 u_1, \\ x^{II}_{(t,T)} &= 1.06 X^{II}_{(T)} + 0.706 u_1, \\ x^{III}_{(t,T)} &= -0.0155 x^I_{(t,T)} + 0.108 x^{II}_{(t,T)} - 0.185 x^{II}_{(t-1,T)} \\ &\quad -0.0223 x^{III}_{(t,T-1)} - 0.00002 X^{III}_{(T)} + 0.769 u_2; \end{aligned}$$

(ii) spring:

$$\begin{aligned} x^I_{(t,T)} &= -4.12 x^{II}_{(t,T)} - 0.0839 x^{III}_{(t,T)} + 0.243 x^I_{(t-1,T)} + 1.15 X^I_{(T)} + 2.61 u_2 + 7.06, \\ x^{II}_{(t,T)} &= 2.02 x^{III}_{(t-1,T)} + 0.834 X^{II}_{(T)}, \\ x^{III}_{(t,T)} &= -0.501 x^I_{(t-1,T)} - 1.11 x^{II}_{(t-1,T)} - 2.75 x^{III}_{(t-1,T)} \\ &\quad -0.0432 x^{III}_{(t,T-1)} - 0.0543 X^{III}_{(T)} + 23.1 u_1 - 10.1 u_2; \end{aligned}$$

(iii) summer:

$$
\begin{aligned}
x^{I}_{(t,T)} &= -2.35x^{II}_{(t-1,T)} + 0.0267x^{III}_{(t-1,T)} + 0.584x^{I}_{(t,T-1)} + 1.04X^{I}_{(T)},\\
x^{II}_{(t,T)} &= 0.0669x^{I}_{(t-1,T)} + 0.482X^{II}_{(T)},\\
x^{III}_{(t,T)} &= -13.8x^{I}_{(t,T)} - 3.32x^{I}_{(t-1,T)} + 2.93x^{III}_{(t-1,T)}\\
&\quad -0.0906x^{III}_{(t,T-1)} + 2.68X^{III}_{(T)} + 48.8u_1 - 13.1u_2;
\end{aligned}
$$

(iv) fall:

$$
\begin{aligned}
x^{I}_{(t,T)} &= -0.00205x^{III}_{(t,T)} + 0.834X^{I}_{(T)},\\
x^{II}_{(t,T)} &= 0.00228x^{III}_{(t-1,T)} + 0.391X^{II}_{(T)},\\
x^{III}_{(t,T)} &= 4.95x^{I}_{(t,T)} + 15.6x^{I}_{(t-1,T)} + 62.0x^{II}_{(t-1,T)}\\
&\quad +0.32x^{III}_{(t-1,T)} - 0.333x^{III}_{(t,T-1)} + 0.892X^{III}_{(T)} + 298.0u_2.
\end{aligned}
\tag{6.44}
$$

In the next step, the balance criterion is used in sorting the forecasts of the leading variable X^{III}.

$$
b_i = X_i - \frac{1}{4}(x_{(w)} + x_{(sp)} + x_{(su)} + x_{(f)})_i; \quad B^2 = \sum_{i=1}^{T_a} b_i^2 \rightarrow \min, \tag{6.45}
$$

where T_a is the anticipated time.

It is necessary to use different schemes as applied above in obtaining the annual and seasonal forecasts; i.e., the annual forecasts are obtained by a one-dimensional linear scheme, whereas the seasonal forecasts are obtained with a two-dimensional time readout.

In evaluating the balance criterion, a total of 25 variants of forecasts (5 seasonal and 5 annual) are carried out and the optimal annual and seasonal models are obtained for the leading variable.

Overall we can say that the calculation of the rank correlation coefficient of data helps in eliminating the collinear factors and thus allows the reduction of computational volume substantially.

4 MODELING OF ECONOMICAL SYSTEM

In economical system modeling, variables are divided into exogenous or external; i.e., those specified outside the model and introduced into it, and endogenous; i.e., those obtained within the model. This means that the input variables are exogenous, while outputs and state variables are endogenous. Sometimes, one variable may be exogenous to a particular model and endogenous to another. In the beginning of the experimentation, it is often not known which variables will be selected for inclusion in the equations; and it is worth adopting a technique to distinguish the varaibles. The ratio of the variables exogenous to endogenous and their participation in the model determines the extent to which the model is open or closed.

The objective systems analysis (OSA) makes it possible to find an autonomous system of linear algebraic or finite-difference equations that is optimal for the given objective criterion assuming that all variables are the system (internal) variables, the "status quo" scenario. The exogenous variables are not indicated *a priori* at the beginning of the analysis. This technique avoids the *a priori* resolution of the difficult and the controversial question: what extent is a variable exogenous. Specialists select the exogenous variables from the set of variables that figured out in the system of equations picked by the computer.

Step-by-step integration of such an autonomous system of equations yields short-range predictions without any special control of the system. This first level scenario often proves useful for showing the use of control to get better processes in the object being modeled, or to get a better estimate for the near future. To proceed with other scenarios, the variables need to be divided into endogenic and exogenic variables or output and input variables. Furthermore, the input variables are to be divided into external disturbances and control variables. This separation of variables is usually done on the basis of physical considerations, which contain the element of subjectivity.

Here various examples of economical system modeling which represent different scenarios in obtaining the optimal systems of equations are presented. They correspond to the modeling of British economy and USA economy based on the studies conducted by Ivakhnenko and his coworkers [40], [41], [38], and Klein, Mueller and Ivakhnenko [76].

In the first example (Example 4, modeling of British economy) the descriptions are considered with linear static elements described by algebraic equations. It was proven that the possibility of changing the control actions is severely limited in such systems.

The second example (Example 5) is the continuation of the first one with an introduction of finite-difference equations with one delayed argument. It is proven that the modified systems analysis is useful in identifying the macroeconomic variables for long-range predictions of up to 10 to 15 years. The predictions of these variables can serve as a basis for estimating other economic indices used in macroeconomic modeling of a country. A solution for the problem of control action promises to design more efficient systems based on the control criteria and control actions.

The third example (Example 6) is meant for briefing the idea of modeling US economy by extending the delayed arguments in the descriptions. The selected system of equations is used to measure the prediction accuracy on the average quarterly values.

The fourth example (Example 7) is the result of extended studies on British economy to adopt a special procedure that measures the cause-effect relationships among the variables. The degree of exogenicity of the variables is defined on the basis of a harmonic criterion. Here the external influences are included in the descriptions for studying the possible changes in the system.

4.1 Examples—modeling of British and US economies

Example 4. The OSA at the level of trends for modeling the British economy.

Usually, a stationary process is represented as the sum of low-frequency and high-frequency components:

$$q(t) = f(t) + \sum_{i=1}^{m} (A_i \cos w_i t + B_i \sin w_i t), \tag{6.46}$$

where the low-frequency part $f(t)$ is called the quasistatic part or the trend, and the high-frequency part is called the dynamic remainder. The former is represented as polynomials or sums of exponentials, whereas the latter is expressed as a finite-difference equation or a harmonical trend. The structural complexity of these functions is uniquely determined by the inductive learning algorithms. One should understand the uniqueness of the selected trend as a single completely defined optimal structure of the equation of the trend and that this uniquely determines the dynamic remainder.

In the sense of objective systems analysis, a system of algebraic equations describing the trends of the output variables is used at the level of trends, and a system of finite-difference

equations is used at the level of the dynamics. These both results of analyses are different because they are related to different mechanisms of statics and dynamics. The coefficients of the former give the interrelationships among the variables, whereas the latter give the interrelationships among their differences.

Frequently, the object itself suggests the type of analysis required. For example, the activity of an aquatic ecosystem is not determined by the level of just one variable like the biomass of phytoplankton, but rather by the rate of its change under the action of external influences. This means that the analysis of dynamic interrelationships is important for ecosystems. For economical systems, it is more important in analyzing the coarse and interaction of the variables.

In general, the operation of the basic inductive learning algorithms is very similar in obtaining the optimal trends and the dynamic remainders. In a given set of multivariables, a nonautonomous physical model in a state space that does not contain the time coordinate is obtained with any algorithm. In case of multilevel algorithm, objective system analysis is used as a first step to obtain a set of output variables characterizing the object and then the two-stage approach is used in time space to obtain an autonomous nonphysical predicting model of the trend or remainder.

The general scheme of the multilevel algorithm at the level of trends is discussed below: (i) the table of initial data sample is supplied, (ii) objective systems analysis with linear algebraic equations (not including the time coordinates) is conducted and a characteristic set of output variables is determined, (iii) by the objective systems analysis, the best F_1 mean annual models are identified using the systems of nonlinear algebraic equations including the time as an argument $(T, T^2, \cdots)$ for the output variables, (iv) the best $F_2(< F_1)$ systems of equations are selected from the above annual models using the stability criterion of multiple correlations, (v) by the objective systems analysis, the best F_3 quarterly systems of nonlinear equations including the time coordinates $(t, T, t^2, T^2, \cdots)$ are identified for the output variables, (vi) the balance criterion of predictions is evaluated by sorting the predictions of the F_2 annual models and the F_3 quarterly models, and (vii) the system of equations with optimal complexity is given with their predictions.

The notations x and u denote the quarterly values, X and U denote the annual values of the variables and of the external influences, and t and T denote the quarterly and annual time coordinate values. The data table should at least contain the data of one external influence variable to maintain the static stability of the system of linear equations. This rigid necessity does not arise in case of nonlinear equations. The data is separated into three sets A, B, and C.

The single layer combinatorial algorithm is used to select the best structure of each equation at each layer of the objective systems analysis according to the minimum-bias η_{bs} and regularity $\Delta(B)$; and the system criterion of minimum- bias $\eta_{s_{(bs)}}$ and prediction criterion $\Delta(C)$ are used in further analysis of the equations. Keeping in view the drawbacks of the modeling such as multiple values of the predictions and the presence of noise in the data, a confidence level D is fixed so that $\eta_{s_{(bs)}} < D$. Usually the confidence level is kept below the value of 10^{-4}. This explains why the algorithm chooses an optimal system with the least noise having the most complete informational basis with the necessary state variables and the external disturbances. In an ideal data (without noise) the algorithm indicates that all systems of equations with complete information are of equal value.

Identifying the process of inflation in case of British economy

The following mean annual values of 26 variables are:

X_1 – the national product in million pound-sterlings,
X_2 – the energy consumption in million tons of conventional coal,
X_3 – the steel production in millions of tons,
X_4 – the automobile production in thousands,
X_5 – the companies' revenues in million pound-sterlings,
X_6 – the individuals revenues in million pound-sterlings,
X_7 – the savings in million pound-sterlings,
X_8 – the capital investment in million pound-sterlings,
X_9 – the capital investment in the public sector in million pound-sterlings,
X_{10} – the capital construction in percentages,
X_{11} – the index of industrial production in percentages,
X_{12} – the wholesale costs for materials and fuel in percentages,
X_{13} – the volume of retail trade in percentages,
X_{14} – the index of retail costs in percentages,
X_{15} – the registration of new automobiles in thousands,
X_{16} – the purchasing value of the pound in percentages,
X_{17} – the avearge wages in percentages,
X_{18} – the number of unemployed in thousands,
X_{19} – the number of employed persons in millions,
X_{20} – the number of employement vacancies in thousands of working places,
X_{21} – the labor productivity in percentages,
X_{22} – the exports in million pound-sterlings,
X_{23} – the imports in million pound-sterlings,
X_{24} – the current balance in million pound-sterlings,
X_{25} – the money supply (group 1) in million pound-sterlings,
X_{26} – the money supply (group 2) in million pound-sterlings,
U_1 – the tax rate on the company's revenues,
U_2 – the tax rate on the individual's incomes,
U_3 – the government expenditure in millions of pound-sterlings, and
U_4 – the cost of oil in percentages.

Data that covered 15 years (1964 to 1978) were used from the *"Economic Trends: Annual Supplement of 1980 Edition,* London." The data is divided into three sets as $A + B + C = 6 + 6 + 3$.

The results of the OSA and the best systems of equations obtained as a result of the chosen confidence level are given below:

The system of equation at the first layer is

(I)

$$X_{14} = 33.11 - 63.88U_2 + 1.575U_4; \qquad (6.47)$$
$$\eta_{s_{(bs)}} = 0.0002534$$

The other number of systems obtained below the confidence level are:

(II)

$$\begin{aligned} X_6 &= -22.28 + 3.906U_1 + 1.198X_{14}, \\ X_{14} &= 18.7 - 3.437U_1 + 0.8334X_6, \\ X_{22} &= -1.89 + 0.2259X_6; \end{aligned} \qquad (6.48)$$
$$\eta_{s_{(bs)}} = 0.0000225;$$

(III)

$$\begin{aligned} X_6 &= -22.28 + 3.906U_1 + 1.198X_{14}, \\ X_8 &= 4.729 + 1.967U_1 + 0.4364X_6 - 0.2653X_{14}, \\ X_{14} &= 18.7 - 3.437U_1 + 0.8334X_6; \\ \eta_{s_{(bs)}} &= 0.00003024. \end{aligned} \tag{6.49}$$

The system IV consists of the equations for X_6 and X_{14} as they are in the above systems. The systems II, III, and IV have achieved the limits of confidence level in this analysis; the system IV has attained the global minimum. The drawback of using this system is that it is statically unstable and it is admitted in the later works that the selection of control actions applied in the system to make it statically stable is not quite correct. Further developments of the modeling of British economy is described in the Examples 5 and 7.

Example 5. Modeling of the British economy using one delayed argument and without specifying the external influences.

This example illustrates the extension the use of the OSA in modeling the British economy with the following features:

(i) It is assumed that the study consists not only of linear static elements described by algebraic equations, but also of links with one delayed argument, which is also called the first-order linear link.

This can be more generalized by considering the second- and higher-ordered links which can be replaced by a sequential combination of first-order links.

(ii) The exogenous varaibles are not indicated *a priori* at the beginning of the analysis.

(iii) The delayed arguments in the equations enable the selection of equations on the basis of the prediction criterion i or $\Delta(C)$, where i indicates the step-by-step prediction integration and $\Delta(C)$ indicates the error on an examining set C:

$$\Delta^2(C) = \sum_{i=1}^{N_C} \frac{(x - \hat{x})_i^2}{(x - \bar{x})_i^2} \leq 1.0, \tag{6.50}$$

where $x, \hat{x}$, and $\bar{x}$ indicate the actual, estimated and average values of the variable x.

The selected equations satisfy the objective characteristics of both the criteria of minimum-bias and prediction; this means that the system criteria are used wherever applicable in the analysis of the systems of equations.

The data and the data separation are the same as in the previous example.

The *OSA algorithm* is given as below:

The algorithm consists of several layers with gradual increase in complexity of the equations. The single-layered combinatorial algorithm with the two criteria described above is used at all layers. During the selection, the equations with $\Delta(C) > 1.0$ are excluded from the search as they contain false information.

The following general form of the polynomials are considered:

(i) the *first layer*, the equations with not more than three terms of the form

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 T, \tag{6.51}$$

where T is the time coordinate in years;

(ii) the *second layer* , the equations with not more than five terms of the form

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 T + a_3 X_{j_{(T)}} + a_4 X_{j_{(T-1)}}; \tag{6.52}$$

(iii) the *third layer*, the quations of not more than seven terms of the form

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 T + a_3 X_{j_{(T)}} + a_4 X_{j_{(T-1)}} + a_5 X_{k_{(T)}} + a_6 X_{k_{(T-1)}}, \qquad (6.53)$$

and so on until the number of terms not exceeding 18 to 20 as the limit of the combinatorial algorithm or the layers are extended while there is at least one equation among all in the last layer satisfying the condition $\eta_{bs} < D$, where D is a certain "confidence level" chosen on the basis of experience.

Systems of equations for prediction and control

The above OSA is applied in modeling the British economy. In addition to the 26 variables used in the previous example, an additional four economic indices are used as input to the algorithm. The indices picked by the computer are

X_{27} – the budget deficit in million pound-sterlings,
X_{28} – the trade balance in million pound-sterlings,
X_{29} – the current balance in million pound-sterlings,
X_{30} – the overall government expenditure in million pound- sterlings,
X_{31} – the tax on private income in million pound-sterlings, and
X_{32} – the corporate tax in million pound-sterlings.

At the first layer, one equation below the confidence level ($D \leq 10^{-4}$) is selected,

$$X_{22_{(T)}} = -1.017 + 1.293 X_{22_{(T-1)}}, \qquad (6.54)$$
$$\eta_{bs} = 0.00004062.$$

At the second layer, ten equations are selected and the best among them is

$$X_{22_{(T)}} = -1.032 + 1.297 X_{22_{(T-1)}} - 0.003908 X_{29_{(T)}}, \qquad (6.55)$$
$$\eta_{bs} = 0.00002226.$$

At the third layer, eight equations are selected and the best among them is

$$X_{14_{(T)}} = 18.64 + 1.064 X_{6_{(T)}} - 0.2526 X_{6_{(T-1)}} + 0.0653 X_{22_{(T)}}, \qquad (6.56)$$
$$\eta_{bs} = 0.0000376.$$

The selection is terminated further because the minimum-bias begins to increase from the fourth layer onward.

The important feature of the algorithm is that the variables selected at the previous layer are refined by supplying additional output variables and arguments at the successive layer.

In the example given here, the most representative output variables X_1, X_8, X_{11}, X_{14}, X_{17}, and X_{22} are selected at the first and second layers. At the third layer, variables X_6, X_{23}, and X_{30} are added to this list. The best equations with least minimum-bias from the three levels are obtained as below:

$$X_{1_{(T)}} = -10.99 + 0.145 X_{1_{(T-1)}} + 0.002955 X_{25_{(T)}} + 0.001323 X_{26_{(T-1)}};$$
$$\eta_{bs} = 0.8902E - 5,$$
$$X_{6_{(T)}} = -16.76 + 1.033 X_{14_{(T)}} + 0.4645 X_{23_{(T)}};$$
$$\eta_{bs} = 0.1.524E - 5,$$
$$X_{8_{(T)}} = -8.381 + 0.002038 X_{25_{(T-1)}};$$

$$
\begin{aligned}
& \eta_{bs} = 2.78E-5, \\
X_{11_{(T)}} &= 25.87 + 0.7407X_{11_{(T-1)}}; \\
& \eta_{bs} = 6.936E-4, \\
X_{14_{(T)}} &= 18.64 + 1.061X_{6_{(T)}} - 0.2526X_{6_{(T-1)}} + 0.06531X_{28_{(T)}}; \\
& \eta_{bs} = 0.376E-5, \\
X_{17_{(T)}} &= -5.045 - 1.518T + 2.159X_{1_{(T)}} - 0.02077X_{27_{(T)}} \\
&\quad +0.06109X_{27_{(T-1)}}; \\
& \eta_{bs} = 2.568E-5, \\
X_{22_{(T)}} &= -1.032 + 1.297X_{22_{(T-1)}} - 0.003908X_{29_{(T)}}; \\
& \eta_{bs} = 2.226E-5, \\
X_{23_{(T)}} &= -3.392 - 0.007218X_{23_{(T-1)}} + 1.013X_{22_{(T)}} - 0.09933X_{28_{(T)}}; \\
& \eta_{bs} = 2.21E-5, \\
X_{30_{(T)}} &= -5.6 + 0.001638X_{26_{(T-1)}}; \\
& \eta_{bs} = 4.253E-5.
\end{aligned} \tag{6.57}
$$

The algorithm has yielded a system of nine equations with 14 variables. Values of five variables $X_{25}, X_{26}, X_{27}, X_{28}$, and X_{29} are to be selected based on the nature of the problem or by using auxiliary equations. The time variables in some of the equations, primarily of those selected from a set of arguments are assigned by specialists for prediction of norms. The equations are tested for their stability; the roots of the characteristic matrix equation lie within the limits of the unit circle. Step-wise predictions reveal that the processes converge rapidly with the trends exhibited by them. This means that the obtained system of equations can be used for identifying the system structures and for short-term predictions.

Here we are not presenting the two-stage approach used for long-term quantitative predictions as used before, but the solution of a control problem specified in the current example. Usually in solving the economic control problem, experts indicate not only the variables related to control actions, but also their number and interrelations in the form of objective criteria.

Solution of a control problem

Control variables are chosen by the specialists either from the set of characteristic variables from the OSA or from the variables that correlate fairly with the characteristic variables. This is checked according to the values of the external criteria for the auxiliary equations formed for these variables. It is also necessary to check that the selected control variables are adequate to satisfy the controllability conditions (according to Kalman). These conditions might require expansion of the control variables or a change in the control objective.

For example, let us assume that the variable X_{27} (budget deficit) is chosen by experts as the control action and the control objective is defined as maintaining the ratio of the retail price index to the average wages (X_{14}/X_{17}) at the year 1978 level. The following basic system's equations are used to build up a closed system with the selected control variable X_{27} as one of the arguments.

$$
\begin{aligned}
X_{1_{(T)}} &= -10.99 + 0.145X_{1_{(T-1)}} + 0.002955X_{25_{(T)}} + 0.001323X_{26_{(T-1)}}; \\
& \eta_{bs} = 0.8902E-5, \\
X_{17_{(T)}} &= -5.045 - 1.518T + 2.159X_{1_{(T)}} - 0.02077X_{27_{(T)}} + 0.06109X_{27_{(T-1)}}; \\
& \eta_{bs} = 2.568E-5,
\end{aligned}
$$

$$X_{14_{(T)}} = 0.6X_{17_{(T)}}, \quad \text{and}$$
$$\Delta X_{14_{(T)}} = 0.6\Delta X_{17_{(T)}}. \tag{6.58}$$

The ratio is given as 0.6; that does not mean that wages rise faster than prices. The fact is that the British economic statistics take the year 1975 as the base year for computing the average wage. If the same base year were to be used for the control problem, then the ratio would be 1.56; this means that the price rise outstrips the average wage as compared to 1976.

The system of auxiliary equations is obtained using the combinatorial algorithm with the minimum-bias criterion and the prediction criterion. The reference function considered in the algorithm takes into account all the variables with one delayed argument.

$$X_{i_{(T)}} = a_0 + a_1 T + a_2 T^2 + a_3 X_{i_{(T-1)}} + \sum_{j\,(j\neq i)} (b_j X_{j_{(T)}} + c_j X_{j_{(T-1)}}), \tag{6.59}$$

where $i = 25, 26, 27$; and $j = 1, 11, 14, 17, 23, 26, 27$.

The system of equations identified are

$$\begin{aligned}
X_{25_{(T)}} &= 2729 + 0.4277X_{25_{(T-1)}} + 131.4X_{1_{(T)}} - 47.79X_{14_{(T)}} + 58.27X_{17_{(T)}} \\
&\quad +0.1648X_{26_{(T)}} - 0.4907X_{26_{(T-1)}}; \quad \eta_{bs} = 1.258E-3, \\
X_{26_{(T)}} &= 6024 + 0.9666X_{26_{(T-1)}} + 566.7X_{1_{(T-1)}} - 310.3X_{14_{(T)}} \\
&\quad -611.2X_{14_{(T-1)}} + 2.315X_{25_{(T)}} + 24.77X_{27_{(T-1)}}; \quad \eta_{bs} = 9.29E-4, \\
X_{27_{(T)}} &= 88.29 + 1.188T^2 - 1.61X_{17_{(T)}} + 1.304X_{17_{(T-1)}} - 0.013X_{25_{(T)}} \\
&\quad +0.009885X_{25_{(T-1)}} - 0.004399X_{26_{(T)}}; \quad \eta_{bs} = 4.306E-1.
\end{aligned} \tag{6.60}$$

The resulting closed system contains six equations with six variables. Kalman's controllability conditions are satisfied. Step-wise integration of the system yields the predictions to achieve the ratio (X_{14}/X_{17} = constant) through variation of budget deficit X_{27} (Figure 6.12).

Prediction of variables other than the characteristic variables.

Out of forty variables fed into the computer while modeling the British economy, only 14 were selected as the characteristic variables. All other variables were rejected as unsuitable for modeling continuous consistent features, both as output variables and as constituents of a set of arguments. This does not mean that the rejected variables are impossible to identify using the inductive learning algorithms; this can be done as a supplementary analysis. They can be predicted as a function of time and the variables selected by computer. For example, the following full polynomial can be used in predicting the number of unemployed:

$$\begin{aligned}
X_{18_{(T)}} &= a_0 + a_1 T + a_2 T^2 + a_3 X_{18_{(T-1)}} + a_4 X_{18_{(T-2)}} \\
&\quad + \sum_{j} (b_j X_{j_{(T)}} + c_j X_{j_{(T-1)}} + d_j X_{j_{(T-2)}}); \quad j = 1, 11, 14, 17.
\end{aligned} \tag{6.61}$$

The combinatorial algorithm with the criteria minimum-bias and prediction selected the following equation:

$$\begin{aligned}
X_{18_{(T)}} &= 238.9 - 0.5211X_{18_{(T-2)}} - 10.18X_{1_{(T)}} + 26.55X_{1_{(T-1)}}; \\
&\eta_{bs} = 0.0989.
\end{aligned} \tag{6.62}$$

The suggested aproach of OSA described in the example affords a novel approach to select a system of indices used at various layers of aggregation. Conventional models

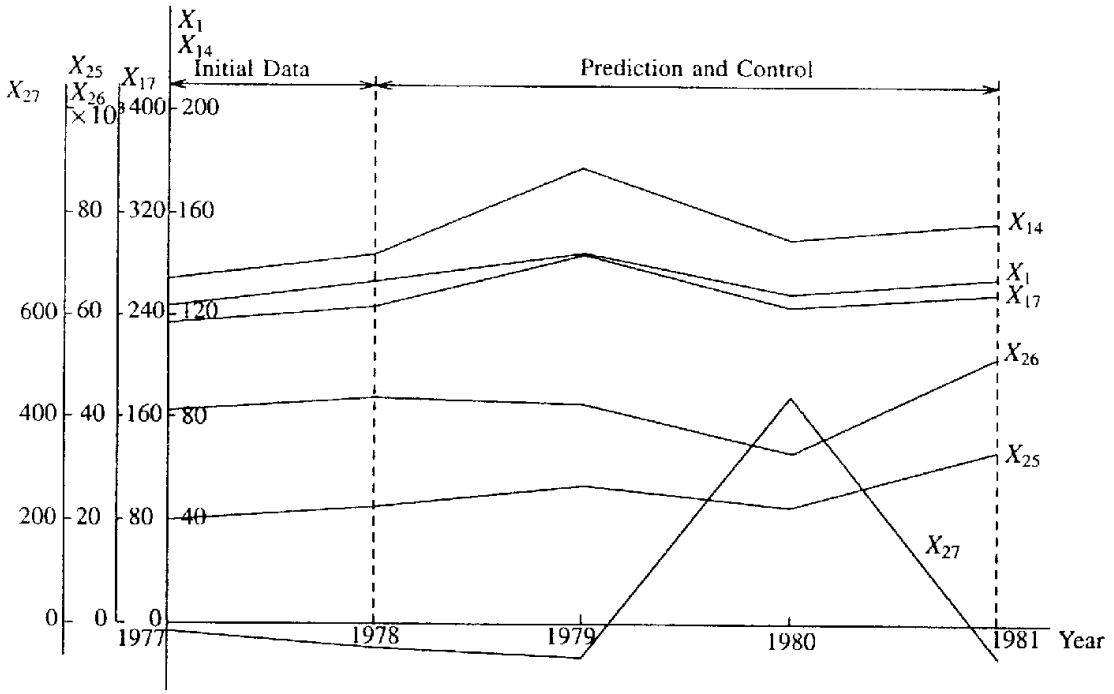

Figure 6.12. Prediction and control action in achieving X_{14}/X_{17} = const. through variation of budget deficit X_{27}

normally use many variables and equations describing their interrelations. The system of indices in a national economic model usually is included with 17 to 130 variables depending on the extent of process elaboration. For instance, in the project headed by L.P. Klein (University of Pennsylvania, USA), the model for British economy contains 226 equations, 206 for the USA economy, and 183 for Canadian economy.

Case studies of the design of the control system have shown that the method described here can be used to design more efficient systems based on other control criteria and control actions. This opens up a vast area of research. It would also be of interest to stimulate control criteria of the form $X_{14}/X_{17} = 0.6 + K(T - 15)$, where T is the time coordinate in years ($T_{(1978)} = 15$), and assigning different values for the coefficient $K = \cdots, -0.2, -0.1, 0, 0.1, 0.2, \cdots$ to determine the effect of economic processes of relative rise or drop in prices.

One can note that in the proposed method of OSA, prediction and control is also applicable in studying other complex systems.

Example 6. Self-organization modeling of US economy.

The objective systems analysis is used in obtaining the system of finite- difference equations. The multilayer algorithm is used to identify each equation. The variables are chosen using the system criterion of minimum-bias and the number of equations is chosen with the prediction criterion using all data. The data used in the algorithm is quarterly data from 1969 to 1974 [76] for the following variables:

x_1 – national product,

x_2 – real national product,

x_3 – national income,

x_4 – personal income,
x_5 – deflation of prices of the national product,
x_6 – deflation of personal prices,
x_7 – consumer price index,
x_8 – whole sale price index,
x_9 – private production in man-hour,
x_{10} – earnings per man-hour,
x_{11} – rate of unemployement,
x_{12} – net running export,
x_{13} – money supply M1,
x_{14} – money supply M2,
x_{15} – rate of 3-month treasury notes,
x_{16} – usage rate of corporate funds,
x_{17} – rate of 6-month commercial paper,
x_{18} – rate of general growth of corporate funds,
x_{19} – rate of personal savings,
x_{20} – pre-tax corporate earnings, and
x_{21} – federal surplus.

The data is separated as even and odd points for the roles of sequences A and B and used in calculating the minimum-bias criterion.

The OSA algorithm is realized in the example below.

At the *first step* of the algorithm, the finite-difference models of all variables are synthesized using the minimum-bias criterion.

$$x_i^t = a_0 + \sum_{j=1}^{8} a_j x_i^{t-j} + \sum_{k=1(k\neq i)}^{21} \sum_{j=0}^{8} b_{kj} x_k^{t-j}. \tag{6.63}$$

All variables are measured with eight delayed values. Each equation is synthesized using the multilayer algorithm. In the first layer of the multilayer algorithm 780 partial models are generated, of which 30 of the most unbiased models are allowed to proceed to the next layer. The selection continues (from 435 models, 30 models are selected) until the minimum-bias criterion ceases to decrease. The number of layers never exceeds ten for any model.

At the *second step*, further selection of system of equations continues from among the obtained models of first step. Here the models that yield a good prediction have priority in the selection process; i.e., the models are selected according to the step-by-step prediction criterion. From among the models obtained from the first step, eight models corresponding to the output variables $x_1, x_3, x_4, x_5, x_6, x_{10}, x_{13}$, and x_{14} are selected and considered as the sought system, the model of the US economy.

The selection threshold is established such that the obtained system of equations would be autonomous and convenient for step-wise predictions. After the selection, the estimates of the equations are adapted using the whole data sample. The system of equations are given below.

$$x_1^t = -114.86 + 0.894x_1^{t-1} + 0.05625x_4^{t-5} - 2.3476x_5^{t-3} + 23.587x_5^{t-6}$$

$$
\begin{aligned}
& -18.275x_5^{t-7} + 5.376x_5^{t-8} - 5.5217x_6^{t-4} - 0.8283x_{13}^{t-7};\\
x_3^t = {} & -451.37 + 0.2563x_4^{t-3} - 1.3502x_5^{t-7} - 3.231x_6^{t-6} + 8.1061x_6^{t-8}\\
& -140.92x_{10}^{t-2} - 236.96x_{10}^{t-3} + 87.045x_{10}^{t-7} + 3.5427x_{13}^{t-1}\\
& +2.8526x_{13}^{t-2} + 1.2499x_{13}^{t-5} + 1.1724x_{13}^{t-6};\\
x_4^t = {} & -238.64 - 84.163x_{10}^{t-5} + 82.507x_{10}^{t-7} + 3.7134x_{13}^{t-2} + 0.6916x_{14}^{t-2};\\
x_5^t = {} & 11.895 + 1.088x_5^{t-1} - 0.00551x_1^{t-8} + 0.293x_3^{t-6} - 0.2725x_6^{t-3}\\
& -0.086x_{13}^{t-7} + 0.03914x_{14}^{t-7} - 0.00515x_{14}^{t-8};\\
x_6^t = {} & 5.585 + 0.1306x_6^{t-2} - 0.01343x_3^{t-7} - 0.00243x_4^{t-3} + 1.2604x_5^{t-1}\\
& -0.2224x_5^{t-2} - 0.04754x_{13}^{t-5} - 0.2118x_{13}^{t-6} + 0.0876x_{14}^{t-6};\\
x_{10}^t = {} & 0.19596 + 0.6092x_{10}^{t-2} + 0.3788x_{10}^{t-4} + 0.1746x_{10}^{t-7} - 0.34245x_{10}^{t-8}\\
& -0.00001x_3^{t-8} + 0.000145x_{13}^{t-1} - 0.00242x_{13}^{t-5}\\
& -0.000319x_{14}^{t-6} + 0.003319x_{14}^{t-7};\\
x_{13}^t = {} & 67.893 + 0.0645x_1^{t-1} + 0.0465x_4^{t-3} - 1.0195x_5^{t-4} + 1.1876x_5^{t-6}\\
& -0.6898x_5^{t-7} + 35.405x_{10}^{t-5} + 19.755x_{10}^{t-7} - 0.4934x_{13}^{t-6}\\
& +0.18827x_{13}^{t-7} - 0.15198x_{13}^{t-8};\\
x_{14}^t = {} & 4.3623 + 1.0298x_{14}^{t-1} - 18.541x_{10}^{t-1} + 15.657x_{10}^{t-3} + 45.187x_{10}^{t-7}\\
& -0.4316x_{13}^{t-1} - 0.4448x_{13}^{t-4} + 0.1071x_{13}^{t-6}.
\end{aligned}
\tag{6.64}
$$

The prediction accuracy is checked for each equation in the system on the data of the years 1975 and 1976. The range of the residual sum of squares vary from 0.28% to 6.87% for one year and from 1.23% to 11.44% for two years.

Example 7. Modeling of the British economy for restoration of the governing laws in the object.

The variables participating in the modeling of a complex object are assumed to have some degree of exogenicity. In this example, the degree of exogenicity of the variables is defined on the basis of a special criterion that is used to find more objective ways of dividing the variables. The objective system-analysis algorithm makes it possible to find an autonomous or closed system of algebraic or finite-difference equations that is optimal for a given criterion by assuming that all the variables in the equations are endogenous or system variables. If we remove those equations from the closed autonomous system whose output variables have proven to be exogenic with the greatest degree, then a governing principle is obtained in the form of an underdetermined system of equations or an approximation of such governing principle, which is suitable for short-range predictions (a "status quo" scenario). Further studies on the system yields a number of other scenarios that are useful for analyzing the object. Obviously, the corresponding equations are not physical laws reflecting the mechanism of objective modeling, but they make it possible to study the possible changes by introducing the external influences with respect to an objective control criterion.

The *OSA* algorithm is given below:

The choice of equations at each layer is made on the basis of two criteria; the minimum-bias (η_{bs}) and prediction (i). The equations with the criterion value of $i^2 > 1.0$ are eliminated from the preliminary selection as the equations providing disinformation.

First layer: Equations of the following form containing not more than four terms on the right side are evaluated.

$$
X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 X_{i_{(T-2)}} + a_3 X_{i_{(T-3)}}. \tag{6.65}
$$

Second layer: Systems of two equations of the following form whose right sides contain not more than eight terms are evaluated.

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 X_{i_{(T-2)}} + a_3 X_{i_{(T-3)}} + a_4 X_{j_{(T)}} + a_5 X_{j_{(T-1)}} + a_6 X_{j_{(T-2)}} + a_7 X_{j_{(T-3)}}. \quad (6.66)$$

Third layer: Systems of three equations of the following form containing not more than 12 terms are evaluated.

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 X_{i_{(T-2)}} + a_3 X_{i_{(T-3)}} + a_4 X_{j_{(T)}} + a_5 X_{j_{(T-1)}} + a_6 X_{j_{(T-2)}} + a_7 X_{j_{(T-3)}} + a_8 X_{k_{(T)}} + a_9 X_{k_{(T-1)}} + a_{10} X_{k_{(T-2)}} + a_{11} X_{k_{(T-3)}}. \quad (6.67)$$

All the systems of equations obtained are autonomous; i.e., the number of variables in them is the same as the number of equations. This enables us to make step-by-step predictions in the system. The layer by layer procedure continues until it reaches the limits of the self-organization modeling according to the basic algorithms used in them. The end result is that one chooses some of the non-contradictory systems of equations, which are attained below the confidence level set for this purpose.

The minimum-bias criterion as a criterion of exogenicity of the variables:

The original data of the variables exhibit information regarding the changes in them. One considers two competitive hypotheses to decide which of two chosen variables X_i and X_j is cause and effect: hypothesis H_1 says that X_i is effect and X_j is cause, and hypothesis H_2 says that X_i is cause and X_j is effect.

These hypotheses can be tested using the single-layered combinatorial algorithm with the criteria of minimum-bias and prediction; one finds two dynamic optimal models using the following complete descriptions:

for H_1,

$$X_{i_{(T)}} = a_0 + a_1 X_{i_{(T-1)}} + a_2 X_{i_{(T-2)}} + a_3 X_{i_{(T-3)}} + a_4 X_{j_{(T)}} + a_5 X_{j_{(T-1)}} + a_6 X_{j_{(T-2)}} + a_7 X_{j_{(T-3)}};$$

for H_2,

$$X_{j_{(T)}} = b_0 + b_1 X_{j_{(T-1)}} + b_2 X_{j_{(T-2)}} + b_3 X_{j_{(T-3)}} + b_4 X_{i_{(T)}} + b_5 X_{i_{(T-1)}} + b_6 X_{i_{(T-2)}} + b_7 X_{i_{(T-3)}}. \quad (6.68)$$

This means that for each hypothesis one finds a model of optimal complexity. The hypothesis is true for finding which minimum-bias criterion is deeper than the other. This procedure is suitable only for the variables of dynamical systems—it works without an error with a sufficiently large sample of experimental data.

A harmonic criterion of exogenicity of variables

Considering each variable in turn as an output variable, let us look at the above dynamic models. By compiling tables of data approximating the change in the two variables, one can make expert evaluations on the models obtained earlier. The problem is reduced to considering one pair of variables as cause and effect, or the other way around.

The period of the lowest-frequency component of each variable [11], [73] is found using

$$S(w) = \frac{2}{N}\left[\sum_{t=1}^{N} X_k^2(t) + 2\sum_{\nu=1}^{N-1}\sum_{t=1}^{N-\nu}(X_k(t) - \bar{X}_k)(X_k(t+\nu) - \bar{X}_k)\cos(\nu w)\right], \quad (6.69)$$

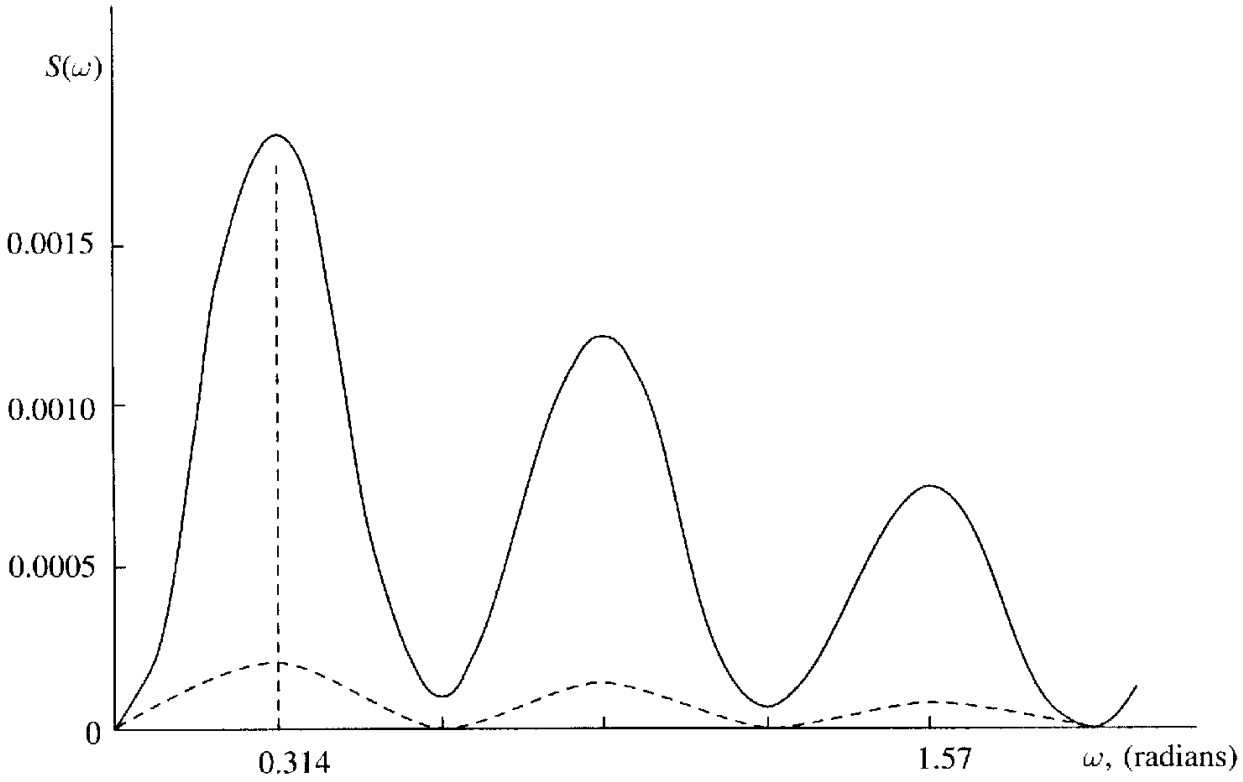

Figure 6.13. Spectrograms of the variables x_1 (solid curve) and x_2 (dashed curve) for determining the lowest frequency

where N is the number of measured data points of the variables $X_k(t)$, $k = 1, 2; \bar{X}_k$ is the average of $X_k(t)$ on N points.

Assigning a sequence of frequencies $0 \leq w \leq \pi$, spectrogram $S(w)$ is graphed until the first local maximum is obtained. This gives us the period T and the frequency w_M of the lowest harmonic. The period of this harmonic should be the same for the two variables (Figure 6.13). The phase of each of the harmonic components is then identified by using the regression equations of the following form:

$$X_i = a_0 + (a_1 \sin w_M t + a_2 \cos w_M t) = a_0 + A \sin(w_M t + \theta_i);$$
$$X_j = b_0 + (b_1 \sin w_M t + b_2 \cos w_M t) = b_0 + B \sin(w_M t + \theta_j);$$

$$\theta_i = \arctan \frac{a_1}{a_2}; \quad \theta_j = \arctan \frac{b_1}{b_2}, \tag{6.70}$$

where w_M is the frequency corresponding to the first local maximum of the spectrogram. The criterion of exogenicity for the two variables is given as $\Delta E = \theta_i - \theta_j$. In dynamical systems, the cause cannot overtake the effect in regard to phase, and consequently, the hypothesis H_1 is chosen if $\Delta E > 0$ (the cause is X_i and the effect is X_j), and H_2 if $\Delta E < 0$ (the cause is X_j and the effect is X_i).

This type of analysis of cause and effect makes it possible to eliminate those equations formed with the output variables that are formally causes rather than effects. Ultimately, the determinate system of equations becomes underdetermined, because the number variables exceeds the number of equations.

The results of modeling of the British economy

The above described OSA algorithm is applied for modeling the British economy. All 26 variables participating in the algorithm, both the endogenic variables X and the four exogenic external disturbances U are treated as equally significant. For the description of the variables and the data, refer to Example 1. The following autonomous system of equations and the values of criteria are obtained as below; the values of the prediction criterion i belong to the set C:

$$
\begin{aligned}
X_{1_{(T)}} &= -0.5739 + 0.5197X_{1_{(T-1)}} + 2.766X_{8_{(T)}} - 15.17U_{2_{(T-3)}},\\
&\quad \eta^2_{bs} = 1.389E-3 \text{ and } i = 2.821E-1;\\
X_{6_{(T)}} &= -0.5937 + 2.518X_{8_{(T)}} + 3.104X_{8_{(T-1)}} + 38.74U_{1_{(T-1)}}\\
&\quad -5.379U_{1_{(T-3)}} - 64.54U_{2_{(T-1)}},\\
&\quad \eta^2_{bs} = 1.590E-3 \text{ and } i = 2.331E-1;\\
X_{8_{(T)}} &= 0.4487 + 0.02953X_{1_{(T-1)}} + 0.2508X_{6_{(T)}} - 0.555X_{8_{(T-2)}}\\
&\quad -4.606U_{1_{(T-1)}} + 12.25U_{2_{(T-3)}},\\
&\quad \eta^2_{bs} = 4.176E-4 \text{ and } i = 5.838E-2;\\
X_{14_{(T)}} &= 4.076 + 5.229X_{8_{(T-1)}} + 22.94U_{1_{(T-1)}} - 33.22U_{1_{(T-2)}}\\
&\quad +86.97U_{2_{(T-2)}} - 54.13U_{2_{(T-3)}},\\
&\quad \eta^2_{bs} = 4.694E-4 \text{ and } i = 3.014E-1;\\
X_{22_{(T)}} &= -5.769 + 0.03826X_{1_{(T-2)}} - 0.2461X_{6_{(T)}} + 2.814X_{8_{(T-1)}}\\
&\quad -17.69U_{1_{(T-2)}} + 7.553U_{1_{(T-3)}},\\
&\quad \eta^2_{bs} = 2.902E-2 \text{ and } i = 8.91E-2;\\
U_{1_{(T)}} &= 0.2981 - 0.01838X_{1_{(T)}} + 0.01304X_{6_{(T)}} + 0.006959X_{6_{(T-2)}}\\
&\quad -0.8259U_{2_{(T-2)}},\\
&\quad \eta^2_{bs} = 1.893E-2 \text{ and } i = 2.123E-3;\\
U_{2_{(T)}} &= 0.1425 - 0.01253X_{1_{(T-1)}} + 0.03903X_{8_{(T-1)}} - 0.00358X_{8_{(T-2)}}\\
&\quad +0.005462X_{14_{(T-1)}} - 0.2325U_{1_{(T-1)}},\\
&\quad \eta^2_{bs} = 3.442E-3 \text{ and } i = 8.210E-4.
\end{aligned}
\tag{6.71}
$$

The step-by-step predictions of the variables are checked to the year 1981 without taking into account any hypothesis to control the system for improving it.

The degree of exogenicity

The degree of exogenicity is used in the search for boundaries in dividing a significant variable into endogenic and exogenic variables. This is measured for each variable by $E_i = \sum_{j=1}^{7} \Delta E_{ij}$, $i = \overline{1,7}$, $i \neq j$. The harmonic criterion of exogenicity is analyzed for the variables of the above system and the results are tabulated in the Table 6.2.

According to the rank of the degree of exogenicity, the variables U_1 and U_2 proved to be the most exogenic. By removing the corresponding equations from the system, it becomes an underdetermined system of equations or governing principle of the system under the requirements that: (i) it is necessary that the selected candidate models be the result of true reference functions of the system and (ii) it is necessary that the initial data be sufficiently accurate. If the first requirement is not true, then the system of difference equations obtained is only an optimal approximation of the governing laws under consideration. With reference to the second requirement, it should be noted that with the increase of noise in the initial data, the OSA algorithm chooses systems with fewer equations. Thus, the physical model corresponding to the object can be obtained with the minimum-bias criterion only

Table 6.2. Results of the analysis on harmonic criterion of exogenicity

Variable	ΔE							E	Rank
	X_1	X_6	X_8	X_{14}	X_{22}	U_1	U_2		
X_1	-	0	0.123	0.033	0.19	0.2	-0.097	0.449	5
X_6	0	-	-0.036	0	-0.38	0.092	0.0066	0.0245	4
X_8	-0.123	0.036	-	-0.22	-0.05	0.83	-0.081	0.492	6
X_{14}	-0.033	0	0.22	-	0	1.04	0.3	1.527	7
X_{22}	-0.19	0.38	0.05	0	-	-1.51	0.031	-0.234	3
U_1	-0.2	-0.092	-0.83	-1.04	0.51	-	0.247	-1.405	1
U_2	0.097	-0.0066	0.082	-0.3	-0.031	-0.247	-	-0.4056	2

in exact data. In the presence of noise in the data, the algorithm gives only the optimum approximation of these laws under the given conditions.

The underdetermined system of equations representing the governing laws or their optimal approximation can be used for constructing other scenarios for short-range normative predictions that are of interest to us.

Principles involved in short-range and long-range predictions

In the above analysis, when the term "governing law" is used, it refers to some mathematical description ensuring a sufficiently accurate short-range predictions of one to three steps ahead. This means that it is interpreted as an approximation of the characteristics of an object in a narrow sense. To restore such a governing law, the inductive approach is used with the minimum-bias criterion.

To restore a governing law in the broad sense and ensure exact long-range prediction with the inductive approach, one uses the minimum-bias and balance-of-predictions criteria. The originality of the multilevel approach is indicated first of all by the fact that these criteria are used at different levels of the analysis. No one mathematical language can exactly express the true physical or other governing laws that are suitable for long-range predictions. A single language can be used only for obtaining a physical model using exact data for short-range predictions. This is substantiated by the existence of a limiting attainable prediction time for all individually accepted mathematical languages. Only multilevel algorithms based on the balance criterion ensure long-range predictions; i.e., these can be applied for obtaining the governing laws of an object in the broad sense.

The governing laws corresponding to an economical system should be sought in the form of a consistent system of annual and quarterly models that ensures an exact qualitative long-range prediction.

5 AGRICULTURAL SYSTEM STUDIES

Various studies [65], [74] on agricultural systems reveal that fulfilling agricultural production forecasts, particularly in large irrigation systems remains a difficult problem. Difficulties arise which seemingly cannot be overcome by conventional modeling techniques.

Problems concerning the control and planning of irrigated forms are divided into three basic groups:

The *first group of problems* is related to the selection of varieties and hybrids of farm products that can be effectively cultivated in soil irrigated with a controlled quantity of

water, and the optimum planning of the structure of irrigated areas for the entire crop rotation cycle.

A crop or a hybrid is assumed to be superior if it provides a maximum average yield ($\bar{y}$) on an observable series of meteorological factors $\{X = (x_1, x_2, \cdots, x_n)\}$ for standard values of the vector of control variables $\{U = (u_1, u_2, \cdots, u_k\}$ and a specified limit (δ) imposed on the standard deviation (s) of the yield:

$$\begin{aligned} \bar{y} &= \tfrac{1}{T} \textstyle\sum_{i=1}^{T} y_i(X, U) \rightarrow \max, \\ s &= \tfrac{1}{T} \textstyle\sum_{i=1}^{T} \sqrt{(y_i - \bar{y})^2} \leq \delta, \end{aligned} \tag{6.72}$$

where T is the number of years in which the meteorological factors are observed and y_i is the crop yield calculated from the meteorological factors of the ith year.

The *second group* is of problems dealing with crop rotation cycle on each field for maximizing the production with the given limitations of water and fertilizers. The task of this group is to maximize the gross yield of a crop-rotation area in hundred-kilograms of grain (y_m) for a known structure of the sowing area:

$$y_m = \sum_{j=1}^{m} y_j(X, U) \rightarrow \max, \tag{6.73}$$

where m is the number of fields included in the crop-rotation area; y_j is the yield of crop j, and $X(x_1, x_2, \cdots, x_n)$ are predicted values of the meteorological factors. The structure of the optimal model contains the selected arguments from the vector of controlled variables (U) for specified crop-yield planning intervals.

The *third one* encompasses operative production control/ planning of the irrigation-crop rotation system with the effective use of fertilizer, water, climatic, and technical resources. Here the goal is to determine the optimum distribution of water resources, fertilizers, and other yield control factors (U_i) that can be used during the ith interval:

$$y_i = \sum_{j=1}^{m} y_{ji}(X_i, U_i) \rightarrow \max, \tag{6.74}$$

where $i = 1, 2, 3$ is the control interval number (intervals between watering); $j = 1, 2, \cdots, m$ is the field number included in the crop-rotation area; and y_{ji} is the yield model of crop j during control interval i.

On each control interval, a separate yield model is used and the vector X consists of a combination of measured and predicted climatic factors. The control intervals are selected to be compatible with the phases of development of various crops, but are not longer than the period between the waterings.

Each of these problem groups requires different arguments of the yield model, control intervals, and allowed accuracy of solutions. These requirements are usually non-contradictory, requiring one to synthesize a complex of yield models for each group of problems. Solving these problems on the basis of the experimental field data and by using trial and error methods yields no guaranteed solutions to the problem of effective use of irrigation systems for a given period. But these problems can be adequately solved by using the production models synthesized on the basis of experimental data obtained from soil and climatic conditions of a specific irrigation system. This is adressed in the work of Khomovnenko and Kolomiets [74] by the use of inductive learning algorithms. In this work, the self-organization modeling is described for modeling of winter wheat productivity. Various problems are organized independently to model them with partial models using small groups of arguments. The experimental data used in this study is collected from one of the agricultural experimental stations located in Ukraine during 1967 to 1975.

5.1 Winter wheat modeling using partial summation functions

Example 8. Modeling of winter wheat productivity. Agricultural production depends on the number of natural factors and on the agro-technology used in the process; i.e., on the possibility of selecting the control factors. Natural factors can be divided into soil and climate factors. Soil factors are mechanical composition and depth of profile (which characterize mechanical strength, speed of infiltration, water conductivity of unsaturated soil, and the characteristic of water retention), aeration, fertility, salinity (which influences structure, weight potential, and toxicity), temperature, and water levels of the soil. Climatic factors are temperature and humidity of air, sedimentation, wind, light intensity, length of the day, and length of the growing period. Some of these factors are independent and the changes in some make changes in others. These factors influence the soil-climate zone and determine the possible selection of varieties and hybrids of plants in the irrigated plant cycle. They have various effects on the production of agricultural products depending on the period of vegetation.

Three main situations in irrigated soil management are pointed out below:

(i) photosynthesis is limited by factors not related to water delivery, but the water is limited by water reserves. An increase in water use due to sprinkling either does not increase the harvest much or has no effect on it;

(ii) if both the photosynthesis and water consumption are limited by water supplies, then the harvest increases with an increase of water use, and sprinkling is most effective; and

(iii) photosynthesis is affected by factors other than water supply; the water consumption is not affected by the water supply in the soil. In this case, it is necessary to use other methods of controling crop production; for example, by increasing the mineral feeding.

In this example of modeling winter wheat, the following identifiable phenophases of development are considered:

1 – from planting to sprouting,
2 – from sprouting to the beginning of tilling,
3 – period of wintering,
4 – from the end of tilling to stem formation,
5 – head formation,
6 – milk formation,
7 – waxy milk formation,
8 – waxy formation, and
9 – complete ripeness.

The model of winter wheat harvest is represented in a general form of output and participating input variables as

$$\begin{aligned} y = f(&B, Q, H, N, P, K, T^{\circ}_{min}, h, l, C, \\ &W_1, \cdots, W_9, t_1, \cdots, t_9, T^{\circ}_1, \cdots, T^{\circ}_9, t^{\circ}_1, \cdots, t^{\circ}_9, \\ &S_1, \cdots, S_9, E_1, \cdots, E_9, R_1, \cdots, R_9, N_{r_1}, \cdots, N_{r_9}, N_{h_1}, \cdots, N_{h_9}, \\ &t_{N_1}, \cdots, t_{N_9}, t_{P_1}, \cdots, t_{P_9}, t_{K_1}, \cdots, t_{K_9}), \end{aligned} \tag{6.75}$$

where y is the wheat harvest in 100 kg-units (220.462 lbs)/ hectare; B is the index of the predecessor; Q is the soil fertility; W is the water in the soil; H is the amount of seed;

C is the period of planting; N, P, K are the amount of nitrogen, potassium, and calcium introduced into the soil; T°_{min} is the sum of minimal temperatures below $-10^{\circ}C$ during wintering; h is the depth of frost in the soil; l is the amount of snow cover; t_i is the length of the phenophase; T°_i is the sum of effective midday tempeartures during the phenophase; t°_i is the average daily air temperature; S_i is the number of sunshine hours; E_i is the sum of evaporation from the water surface during the phenophase; R_i is the amount of rainfall; N_{r_i} is the number of rain days; N_{e_i} is the number of days having a relative air humidity of less than 30%; $t_{N_i}, t_{P_i}, t_{K_i}$ are the periods of fertilizing; and $i = 1, 2, \cdots, 9$ are the index numbers for indicating the phenophases.

The factors considered in the general form are regularly observed variables at the farm's experimental station and the standard observations made by the national meteorological network. One can also include other factors into the model [64]. The above general form is analyzed for various aspects of the modeling given below. All models are obtained using the single-layered combinatorial algorithm.

A special aspect of this modeling is that the described model is replaced by a selection of partial models, each of which is either linear or a quadratic polynomial with small number of arguments. The action of the remaining factors is averaged and used in representing the precision of the model. The importance of this approach is indicated by the demands of the problems under study and also of the availablity of measuring factors in determining the productivity of crops. The major advantage is that it is simple and considered carrying experiments in accordance with the theory of planning experiments.

Modeling of wheat harvest (y) as a function of the time of planting (C) and the rate of seeding (H)

The data used in this modeling is obtained during 1967 to 1971 for five years of 525 experiments. The seeding rate was maintained as $H = 1, 2, \cdots, 7$ million seeds per hectare and the planting time C was from September 25th to October 25th with three variants of water use; without sprinkling, watering of 1200 m^3/ hectare, and watering plus vegetational sprinkling assuring soil moisture not lower than 80% of the least capacity.

The combinatorial algorithm is used with the combined criterion of minimum-bias plus regularity. All 525 data points are used in the analysis. The best model obtained is given below:

$$y = 24.7 + 4.9H - 0.58H^2 + 0.03HC - 0.0056C^2 \tag{6.76}$$

The quadratic character of the model indicates that the model achieves the maximum value of the optimal norm of planting and periods of planting guaranteeing a maximum harvest. However, the large bias term and the mean squared error (ε) of 33% on all points indicate the influence of other factors like the water regime and the climatic factors that are not part of the model.

To determine the degree of the effect of the water regime, the experimental data points are divided into three sets of 175 points each and partial models are identified for three separate water systems:

$$\begin{aligned} y_1 &= 6.32H - 0.51H^2 + 6.96C - 0.17HC - 1.09C^2, \quad \varepsilon_1 = 41\%, \\ y_2 &= 6.89H - 0.70H^2 + 15.01C - 0.22HC - 2.46C^2, \quad \varepsilon_2 = 31\%, \\ y_3 &= 11.88H - 0.98H^2 + 12.37C - 0.49HC - 1.83C^2, \quad \varepsilon_3 = 27\%, \end{aligned} \tag{6.77}$$

where the model for y_1 is constructed using the data on without sprinkling; the model for y_2 is with one watering in the fall; and the model for y_3 is under conditions with optimal water

supply as suggested by agrotechnology. One can see that the accuracy of models increases from 41% to 27%.

The optimal norm of seeding H and periods of planting C have achieved the above model for maximizing the harvest with

$$\begin{aligned} H_1 &= 5.73; \quad C_1 = 2.67; \\ H_2 &= 5.16; \quad C_2 = 3.43; \\ H_3 &= 5.37; \quad C_3 = 2.64. \end{aligned} \tag{6.78}$$

Here the rate of seeding is given in millions of seeds per hectare and the time of planting in days counting from September 1st multiplied by 0.1. In this way, the optimal period of planting for wheat without sprinkling and with an optimal water supply is September 26 to 27, and for wheat with one initial fall watering is October 4 to 5. The maximum harvests of winter wheat for different watering systems are

$$y_{1max} = 27.66; \quad y_{2max} = 41.13; \quad \text{and} \quad y_{3max} = 48.29. \tag{6.79}$$

Using these values of maximum harvests for various watering systems, one can construct a function $y_{max} = f(W)$, where W is the quantity of water in thousands of m^3 per hectare lost in sprinkling. The following water losses are accepted for each case as (i) without sprinkling, 0.1 (moisture supply to assure germination); (ii) one initial watering, 1.2 thousand m^3 per hectare; and (iii) total losses to maintain soil dampness not lower than 80% of the moisture capacity, 3.0 thousand m^3 per hectare. The second ordered function is considered and the estimated model is obtained as

$$y_{max} = 26.02 + 16.14W - 2.90W^2. \tag{6.80}$$

This model is useful in planning the yield depending on the water supply of the sprinkling system, and also in determining the specific losses in obtaining the quantity of agricultural production.

Modeling of wheat harvest (y) as a function of the rate of fertilizing (N, P) and the water supply (W)

The experimental data observed during the years 1969 to 1972 are used; altogether 168 experiments with 14 different mineral feedings and three types of water supplies were conducted. The mineral feeding consists of nitrogen (N = 30, 60, 90, 120, 150 acting units), and phosphorus (P = 30, 60, 90, 120 acting units).

The data points are divided into three sets; each set consists of 56 points for each water supply. The full description of second-order is used for each water system and the best models are selected according to the combined criterion of "minimum-bias plus regularity" of the form:

$$\begin{aligned} y_1 &= 18.1 + 1.2P - 0.069P^2 + 0.98N, \quad c2 = 49.3\%, \\ y_2 &= 23.8 + 1.76P - 0.105P^2 + 1.12N, \quad c2 = 9.8\%, \\ y_3 &= 27.0 + 2.24P - 0.115P^2 + 1.18N, \quad c2 = 12.2\%. \end{aligned} \tag{6.81}$$

The lower accuracy of the model of y_1 with experiments without watering indicates a strong influence of other factors which are not taken into consideration in the model. Among them are the fluctuation of natural soil moisture and other controlling and disturbing actions whose influence is reduced because of improved water delivery in the other models. The increase

in the values of the coefficients in response to the improved water delivery indicates the increase of intensiveness of the mineral feedings introduced into the soil.

Considering that the harvest models for three watering systems have the same structure of the form,

$$y_i = a_{i0} + a_{i1}P - a_{i2}P^2 + a_{i3}N, \quad i = 1,2,3, \tag{6.82}$$

and that the effect of water supply on the production is nonlinear, a second-ordered function is selected for constructing the dependence of coefficients of the above model on the water supply. The general form of such a dependence is given as

$$a_{ij} = b_{ij} + c_{ij}W + d_{ij}W^2, \quad i = 1,2,3; \quad j = 0,1,2,3. \tag{6.83}$$

After finding the dependencies of all coefficients a_{ij} as $f(W)$, the general model of the wheat production taking into account the water supply and fertilizers used is found as

$$\begin{aligned} y = {} & (18.1 + 6.7W - 1.17W^2) + (1.13 + 0.62W - 0.84W^2)P \\ & -(0.065 - 0.05W - 0.011W^2)P^2 + (0.96 + 0.17W - 0.032W^2)N. \end{aligned} \tag{6.84}$$

This model can be used for predicting the harvest depending the nitrogen and phosphorus additives and also on the water delivery system. One can study the effectiveness of using water resources for all periods of the watering season under the conditions of deficient water supplies using the analogous models for crops in plant rotation as functions of fertilizers.

Modeling of wheat harvest as a function of various climatic factors

The data corresponding to the 60 field experiments conducted during 1973 to 1975 are used; the experiments are aimed to define the effects on the harvest of the amount of seed planted and the quantity of nitrogen fertilizers used with an optimal water supply. The meteorological data are obtained from the meteorological station located within a distance of one kilometer from the experimental fields.

(i) *Models based on the duration of phenophases.* To determine critical phases of development in winter wheat and dominating meteorological factors, the models are constructed by taking into account the dynamics of separate meteorological factors in the main phases of vegetation.

Keeping in view the limited quantity of data points, including those obtained under conditions of good wintering, the models for phenophases are synthesized. Linear models which characterize the dependence of the main quantitative characteristics of the harvest are considered on the duration of development phases in days:

$$\begin{aligned} y_1 &= 4.75t_5 - 1.96t_6 + 1.29t_7 - 2.83t_8 - 2.78t_9, \quad \varepsilon_1 = 15.6\%, \\ y_2 &= 75.8 + 93.9t_7 - 25.6t_8 - 45.1t_9, \quad \varepsilon_2 = 9.1\%, \\ y_3 &= 37.5 - 0.076t_9, \quad \varepsilon_3 = 4.8\%, \end{aligned} \tag{6.85}$$

where y_1 is the harvest in 100 kg/ hectare; y_2 is the weight of thousand seeds in grams; and y_3 is the number of seeds in a head.

The selected models serve only for characterizing the tendency toward a decrease in wheat harvest (y_1) with an increase of the length of the milk (t_6), waxy (t_8), and complete ripeness (t_9), at the expense of a decrease in kernel weight (y_2). Similarly, it is better to mention that an increase of harvest (y_1) with the milky wax degree of ripeness (t_7) occurs at the expense of an increase in kernel weight (y_2). This is because of the positive coefficients of t_7 in the first two models for y_1 and y_2.

(ii) *Models based on the number of sunshine hours.* The purpose of extending the above modeling is to find the climatic factors that exert more influence than the duration of the phenophase on the production. Linear models are synthesized depending on the amount of sunshine S during the different phases. The optimal models are obtained as below:

$$\begin{aligned} y_1 &= 3.83S_5 - 0.313S_6 - 3.65S_7 + 0.88S_9, \quad \varepsilon_1 = 10.5\%, \\ y_2 &= 42.8 - 0.313S_5 - 0.685S_7 + 1.114S_9, \quad \varepsilon_2 = 3.2\%, \\ y_3 &= 1.51S_5 + 0.056S_6 + 0.474S_9, \quad \varepsilon_3 = 11.1\%. \end{aligned} \tag{6.86}$$

According to the length of the phases and the amount of sunshine from the above models, one can state that the amount of sunshine is the essential element in the harvest. Sunshine hours during wheat head formation S_5 affect the quantity of kernels in the head y_3. It is also correct to state that an increase in S_5 means a slight decrease in the weight of the 1000 kernels y_2.

(iii) *Models based on the temperature.* Models are constructed for the sum of effective temperatures for the phases in the winter wheat development. Additional arguments are included into the models for determining the effect of the early phases of development in wheat production. The resulting models obtained for y_1 and y_2 have the form

$$\begin{aligned} y_1 &= 6.35L_5 + 1.76t_5^\circ - 0.07T_6^\circ + 0.88T_9^\circ, \quad \varepsilon_1 = 14.6\%, \\ y_2 &= 2.92t_5^\circ - 0.08T_5^\circ + 0.445T_7^\circ, \quad \varepsilon_2 = 2.5\%, \end{aligned} \tag{6.87}$$

where t_i° and T_i° are the average daily air temperature and the sum of effective midday temperatures during the ith phenophase, correspondingly; L_5 is the degree of development of leafy surface during the phase of head formation (for the April 9th).

These models take the second place to the models depending on the duration of sunshine hours. Here also the strong dependence of productivity on the conditions of head formation is evident.

(iv) *Models based on the water evaporation.* The evaporation capacity of the atmosphere is an important indicator of the conditions of an agricultural production. One can use various meteorological indicators such as the relative and absolute humidity of the air, its temperature, atmospheric pressure, air dryness, evaporation from the water surface, and various combinations of these and other indicators in constructing the models (also unified into one model with the variable coefficients using the inductive learning algorithms) for climate zones with appropriate adaptation for each irrigated field.

Here the variables concerning the total and midday evaporation from the water surface of the standard evaporation tank are used in models for each phase of winter wheat development. The following models are obtained:

$$\begin{aligned} y_1 &= 12.3E_5 - 8.71E_7 - 8.83E_9, \quad \varepsilon_1 = 10.1\%, \\ y_2 &= 113.5 - 0.464E_5 - 3.09E_6 - 10.3E_9, \quad \varepsilon_2 = 2.1\%, \\ y_3 &= 3.54E_5 - 0.368E_9, \quad \varepsilon_3 = 9.3\%. \end{aligned} \tag{6.88}$$

According to the structures of these models, these are analogous to the models that represent the dependence of the harvest using the number of sunshine hours. These models also trace the relationship with E_5 during the head formation. One can draw the conclusion that the number of kernels is established during the period of head formation, and that an increase in evaporation causes an increase in the number of kernels and better harvest.

Models with respect to the end of the head formation phase.

From the above modeling results, it is clear that climatic factors affect the wheat harvest and its characteristics mainly prior to the stage of head formation. This makes it possible to build up a good prediction model for the wheat harvests for the end of the head formation phase using the amount of evaporation from the water sutrface (E_5), the number of sunshine hours (S_5), the duration of the phenophase (t_5), and the relative air humidity (e_5).

The linear models obtained are as follows:

$$\begin{aligned} y_1 &= 5.91S_5 - 2.06t_5 + 11.13e_5 - 3.65E_5, \quad \varepsilon_1 = 10.1\%, \\ y_2 &= -0.56S_5 + 0.12t_5 + 16.19e_5, \quad \varepsilon_2 = 2.6\%, \\ y_3 &= 0.51t_5 + 6.26e_5, \quad \varepsilon_3 = 5.1\%. \end{aligned} \tag{6.89}$$

The models can be used for the evaluation of agricultiral-climatic resources of a specific irrigation system for determining the possibilities of raising a particular crop for which the model is identified. In this case, the observed data can be averaged to the available series of years. The same models can be used for predicting the future harvest with known factors of crop development for the end of the head formation phase or earlier—for example, after wintering with a favorable prediction of the sunshine, evaporation of water surface, and length of the head formation phase.

It is necessary to supplement with certain appropriate biological indicators for a more objective prediction of wheat harvest after the beginning of spring growth. The variables such as the degree of development of the leafy surface L and the quantity of dry matter V at the beginning and end of the head formation are used in constructing the second-order polynomial models; the optimal models have the form

$$\begin{aligned} y_1 &= 17.6L_4 + 0.39V_4 - 1.29L_4^2, \quad \varepsilon_1 = 6.7\%, \\ y_2 &= 77.5 + 0.006V_4^2 + 3.19L_4 - 5.85L_4V_4, \quad \varepsilon_2 = 4.7\%, \\ y_3 &= 37.6 - 0.00066V_4^2 + 0.0065L_4^2, \quad \varepsilon_3 = 4.7\%; \end{aligned} \tag{6.90}$$

$$\begin{aligned} y_1 &= 40.55 - 0.0023V_5 + 0.086V_5L_5 - 0.325L_5^2, \quad \varepsilon_1 = 13.6\%, \\ y_2 &= 43.17 - 0.06V_5 + 1.12L_5, \quad \varepsilon_2 = 6.7\%, \\ y_3 &= 37.69 - 0.00003V_5^2 + 0.0014V_5L_5, \quad \varepsilon_3 = 4.9\%. \end{aligned} \tag{6.91}$$

The latter which is a more precise model of the harvest specifies the dependence of wheat production on the degree of development of the leafy surface at the end of the stalk formation phase. The dependence of the production on the quantity of dry matter and the amount of leaf surface at the end of the head phase decreases because of biological changes occurring in the head stage.

This means that it is sufficient to indicate the amount of leaf surface and the amount of dry matter at the end of the the stalk formation phase for building up the prediction models. The linear models that consider the climatic factors during the head stage have the form

$$\begin{aligned} y_1 &= -77.5 + 2.89L_4 + 3.07S_5 - 0.032t_5 - 0.139e_5 + 17.19E_5, \quad \varepsilon_1 = 6.8\%, \\ y_2 &= 0.048V_4 - 0.503S_5 + 0.079L_5 + 16.05E_5, \quad \varepsilon_2 = 2.5\%, \\ y_3 &= 32.94 - 0.061V_4 + 0.563e_5, \quad \varepsilon_3 = 5.1\%. \end{aligned} \tag{6.92}$$

The inclusion of the variables of biological indicators of development at the end of stalk formation permit a higher degree of accuracy in the models. The variable corresponding to

the amount of dry matter is discarded by the computer in the model of wheat harvest (y_1) and the variable corresponding to the surface of green foliage is left in the model.

The above selected models can be used effectively for predicting the wheat harvest since the measurement of foliage area can be carried out with high precision by indirect methods—for example, with the aid of aerial photography.

The studies on agricultural productions with irrigation are related to a large number of meteorological and agro-technical factors. The majority of the factors are cross-correlated and the construction of one universal model for studying various aspects of agricultural production is inadequate. This study indicates that as a result of directed selection of partial models constructed for a large group of arguments, it is possible to get models of production corresponding to certain practical considerations.

Example 9. Adaptation of yield models to crop regionalization.

The principal sources of model error in the above example are the limited usage of the variables. Here is a trial to further extend modeling of the yield of agricultural crops in adapting to regionalization. The organization of the modeling process objectively takes into account the expansion of input variables. The modeling errors depend on the noise in the experimental data, assumptions made in the design of the complete model (i.e., the maximum complexity of the model and correspondence to the physical process), the method used to divide the experimental data into sets, the choice of control intervals (averaging periods of input variables), the criterion of model selection, etc.

A particular feature of crop rotation-structure planning to solve the different groups of problems in adapting yield models to crop regionalization is addressed in the work of Khomovnenko [75] with the exclusive use of predicted input variables. The intermediate yield models are developed with the use of calender (monthly, seasonal, or average of several months) values of climatic factors.

For example, considering the modeling of wheat harvest as $y = f(H, C)$, 20 more variables which characterize the climatic conditions are supplemented to the original list of variables H and C; average daily temperatures $t_{(i-j)}$, precipitation $R_{(i-j)}$, sunshine hours $S_{(i-j)}$, and lack of air humidity $e_{(i-j)}$ are summed over the period from ith to jth month and used as an input in the model (when $i = j$, only one subscript is used).

The meteorological data used in modeling is collected during 1945 to 1974 by one of the agricultural weather stations that is located about 100 km away from the experimental farm fields. This a situation in which time and spatial extrapolation of input variables could be used.

The data is divided into four sets: training (A), testing (B), and two examin sets (C_1, C_2, and $C = C_1 \cup C_2$). The second examin set is produced from the sequence of meteorological data, measured in the region for which the yield model is to be used. The regularity criterion $\Delta(A \cup B)$ is the squared error measured on the training and test sets, $\Delta^2(A \cup B) = \sum_{i \in A \cup B}(y - \hat{y})_i^2 < 0.02$; $\Delta(C_1)$ is the squared error measured on the first examin set, and $\Delta(C_2)$ is the squared error measured on the second examin set. The model accuracy is improved further by an appropriate choice of averaging intervals of the input variables—for example, by making these intervals equal to the development phases of the crop, and the model adaptive with respect to the climatic conditions of the irrigation system. To adapt the model, the step-by-step correction of equations is used by excluding the extremum test data points. The algorithm with orthogonalized complete description (a generalized algorithm) is used [112] (also refer to Chapter 2).

In the first case, the criteria used in estimating the potential effectiveness are the average annual yield ($\bar{y}$) and the mean square deviation $\Delta(A \cup B)$. A complete sifting of all

intermediate models yields the model

$$y = 53.41 - 0.000203 e_{(5)} t^{\circ}_{(6)} - 4588.8 \frac{1}{H e_{(3-5)}} + 0.048 R_{(3-5)}, \tag{6.93}$$

which is the optimal yield model for winter wheat; the average annual yield computed from this model is $\bar{y} = 23.25$ 100- kg./hectare and the error $\Delta(C) = 0.004$. In computing the criterion error, the variable y is replaced by $\bar{y}$.

Similarly, in the second case the yield models used for annual planning of water distribution have different lists of input variables that characterize watering and fertilization schedules and climatic factors. The models obtained are as:

$$\begin{aligned}
y_1 &= 69.1 - 0.0000023 \frac{S^2_{(5)}}{t^{\circ}_{(10)}} + 551.1 \frac{N}{t^{\circ}_{(4)}} - 0.0023 \frac{t^{\circ}_{(5)} t^{\circ}_{(3-4)}}{P t^{\circ}_{(4)}} - 6.0 \frac{S_{(4)}}{t^{\circ}_{(4)}}, \\
y_2 &= 30.9 + 0.0000079 R_{(12-2)} N S_{(4)} t^{\circ}_{(5)} - 57.7 \frac{1}{N S_{(3)}} - 0.94 \frac{1}{R_{(11-2)}}, \\
y_3 &= 31.91 + 0.007 N S_{(4)} + 0.33 N R_{(3-5)} - 0.022 \frac{1}{N^2 R_{(3-5)}} + 1.72 \frac{1}{N^2 S_{(3)}}.
\end{aligned} \tag{6.94}$$

The models for y_2 and y_3 are obtained as optimal and acceptable solutions, but the intermediate models for y_1 do not yield any acceptable solutions due to the limited nature of the experimental series. However, in order to illustrate the potential effectiveness of the above selected models for y_1, y_2, and y_3, which characterize different irrigation schedules, the yields are simulated for a period of 30 years (1945 to 1974).

The results of the investigation of the models prove to be physically sound and agree with the experimental results conducted by various scientists. A further increase in accuracy can be achieved by optimizing the average intervals of the input variables.

6 MODELING OF SOLAR ACTIVITY

Model as a sum of trend and remainder

The random processes being modeled can be represented as the sum of trend and a remainder.

$$y(t) = Q(t) + q(t), \tag{6.95}$$

where $Q(t) = \frac{1}{T_m} \sum_{t=1}^{T_m} y(t)$ is the moving average about the center of the averaging interval and $q(t)$ is the remainder (called "anamoly" in meteorology). Both are modeled using one of the iductive learning algorithms. The process averaging interval T_m gradually increases until the sum of the prediction errors of the trend and of the remainder decreases. The global minimum of the sum is the optimal value of T_m.

Additive-multiplicative trend

Alternatively, more complex is the additive-multiplicative trend of the form

$$y(t) = Q_1(t) + Q_2(t) q_1(t) + Q_3(t) q_3(t), \tag{6.96}$$

where Q_1, Q_2, and Q_3 are polynomials in time with the dimensions of the moments of a random function—the mathematical expectation, the variance, and the third moment, respectively. Such trends are first of its kind in the literature to be used.

The limiting admissible prediction time with respect to a trend is estimated on the basis of the averaging interval; $T_{tr} \leq T_m$, where T_{tr} is the averaging interval of trend. Usually this interval is sufficiently large. For autoregression single factor models of the form $q_t = f(q_{t-1}, q_{t-2}, \cdots)$ the limiting admissible prediction time of the remainder in the case of a single-level prediction does not exceed the correlation interval $T_{rem} \leq T_k$, where T_{rem} is the averaging interval used in the remainder part.

To determine the limiting admissible prediction time for multifactor models of the form $q_t = f(q_{t-1}, q_{t-2}, \cdots, u_t, u_{t-1}, \cdots)$, it is necesssary to choose according to the larger of the autocorrelation and cross-correlation intervals.

Multistep prediction as a transient process

Long-range multistep prediction has many features of transient process. In the prediction region of a random process, the predicting model shifts from a steady state regime of continuous step-by-step renewal of information (regime of observation) into another steady-state regime for which no new information of the object is being fed in (regime of prediction). As in the theory of tracking systems of servomechanisms, the predicting model can be represented as the sum of two components—the trend and remainder.

In the interpolation interval, the prediction differs only slightly from the actual data since it is based on the minimization of the mean-square error of the residuals. In turn, in the prediction interval, it is convenient to represent the remainder in the form as a series

$$q(t) = q_1(t) + q_2(t) + q_3(t) + q_4(t), \tag{6.97}$$

where $q_1(t)$ is the exponential component of the remainder, $q_2(t)$ is the attenuating transient error, $q_3(t)$ is the nonattenuating component of the remainder, and $q_4(t)$ is the constant component.

After singling out the trend $Q(t)$ from the actual data, it is expected that harmonic components exist in the remainder with the same frequencies as in the trend because only nonattenuating oscillations of given frequencies are singled out in the trend. The presence of nonattenuating component $q_3(t)$ in the remainder is the result of imperfection of the algorithm for singling out the harmonic trend; in an ideal case $q_3(t) = 0$. In long-range predictions, there is a small steady state angular tracking error ($\theta = \text{const}$) by which the prediction differs from the trend. As in the servosystems, this can be determined without integrating the differential equations of the model, but by substituting the steady state forcing function such as $Q(t) = \sin \omega t$ and the response, $y(t) = A \sin(\theta + \omega t)$ for determining the angle θ.

The analysis of steady-state regime is simple in a single-frequency trend and complicated in several frequencies. In particular, one needs to solve nonlinear equations to determine the tracking error without integrating the equations. But it is simpler to first integrate the equations of the predicting model and, thus, find the tracking error and all other components of the prediction. Once the angular tracking error is determined, then it is easier to determine the quality factor of the predicting model as a ratio of the angular frequency of the trend to the tracking error.

Correlation interval of the transient component

Correlation function of a typical non-steady-state process can be represented as a sum of steady-state and transient components:

$$y(t) = y_{ss}(t) + y_{tr}(t), \tag{6.98}$$

where

$$y_{ss}(t) = Q(t) + q_3(t) + q_4(t) \text{ and } y_{tr}(t) = q_1(t) + q_2(t).$$

The corresponding autocorrelation functions can be found for the three terms in the above equation as

$$A_{yy}(\tau) = A_{ss}(\tau) + A_{tr}(\tau). \tag{6.99}$$

The basic part of the steady-state component is the trend and the remainder is the transient component.

If we construct a correlation function of steady-state process, we will notice that the prediction time of a steady-state process does not exceed the length of its correlation interval; $T_{pr} \leq \tau_c$, where τ_c is the coherence time (refer chapter 2).

The time interval throughout which the correlation function exceeds the value of the delta function, $\delta = 0.05$. The value of $A_{yy} = 0.05$ is taken from the experiments of "tossing a coin." The correlation interval for a purely random process is given as less than unity—$\tau_c \leq 1$; this means that the prediction is impossible from observations. From these experiments, we can conclude that the steady-state component can achieve infinite prediction time and the transient component can achieve the prediction time of less than or equal to unity with the help of the best inductive learning algorithms.

An example of modeling solar activity

The basic steps in the algorithm for predicting oscillatory processes are listed as follows:

1. spectral analysis of the process,
2. singling out the harmonic trend and the remainder,
3. calculating the steady-state and transient components,
4. obtaining the optimal difference equation of the transient component using an inductive learning algorithm,
5. predicting the process as the sum of the steady-state and transient components, and
6. determining the accuracy of the prediction in case of prediction time equal to the correlation interval of transient and evaluating the results.

Here the problem of predicting solar activity characterized by Wolf numbers is considered. The data are taken for the years 1700 to 1978.

1. *Spectral analysis of the process.* The spectral analysis on the series of data reveals that it contains a sharp harmonic component with period $T_0 = 11.2$ years, multiple harmonics with periods $T_1 = (1/2)(11.2)$, $T_2 = (1/3)(11.2)$, $T_3 = (1/4)(11.2)$, and $T_4 = (1/5)(11.2)$ years, and also low frequency harmonics with periods $T_{-1} = 2(11.2)$, $T_{-2} = 3(11.2)$, $T_{-3} = 4(11.2)$, and $T_{-4} = 5(11.2)$ years.
2. *Singling out harmonic trend and the remainder.* The optimal harmonic trend is obtained by gradually increasing harmonic components until it leads to the lowering of the approximation error on the remainder. In this example, the trend obtained using all initial data has only a single harmonic with period $T_0 = 11.2$ years (frequency $\omega_0 = 2\pi/T_0 = 0.56$ radians/ year).

$$Q(t) = 49.9 - 8.3 \sin \frac{2\pi}{11.2} t + 25.4 \cos \frac{2\pi}{11.2} t \tag{6.100}$$

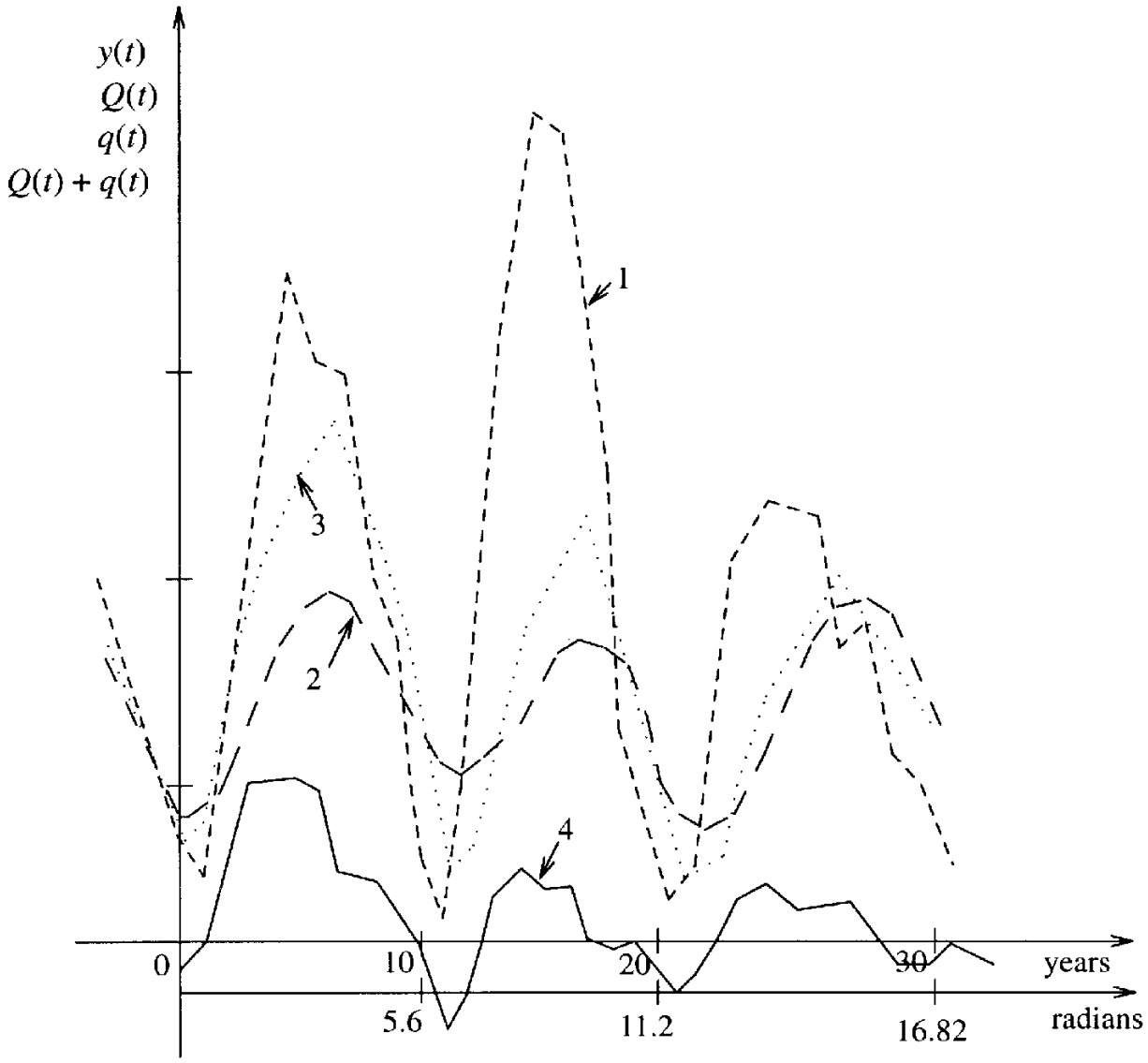

Figure 6.14. Processes being investigated. (1) the actual data of Wolf numbers, (2) the harmonic trend $Q(t)$, (3) the prediction with the trend and the estimated remainder $Q(t)+q(t)$, and (4) the output of the difference equation obtained with the inductive learning algorithm for the remainder data

It is aimed at preserving the possibility of checking the accuracy of long-range prediction over the course of 40 years with the starting year of 1943; i.e., where $t = 0$.

The residual data, which yields the remainder is calculated as the difference between the actual data and the trend. Figure 6.14 illustrates the trend $Q(t)$ and the remainder $q(t)$. As it is expected, the remainder is of an attenuating nature. The same frequency $\omega_0 = 2\pi/11.2$ radians/ year is found as it is used in the trend.

3. *Obtaining a difference model of the remainder.* Here combinatorial algorithm is used to obtain the optimal difference equation; the complete polynomial has 15 lagging arguments and the model is chosen in the plane of two criteria $\eta_{bs}^2 - i^2(N)$ with the constraint of $i^2 \leq 1.0$. The following model is obtained:

$$q(t) = 0.153q_{t-3} - 0.144q_{t-5} + 0.526q_{t-11} - 0.225q_{t-15}. \qquad (6.101)$$

Figure 6.14 exhibits the approximation of the remainder by step-by-step integration of this equation.

4. *Singling out the nonattenuating harmonic part of the remainder.* It is proceeded by the singling out of the harmonic trend and obtained the following harmonic trend, as in step 2,

$$q_3(t) = 1.184 \sin \frac{2\pi}{11.2} t + 5.046 \cos \frac{2\pi}{11.2} t, \tag{6.102}$$

the amplitude of which is $A_3 = \sqrt{(1.184^2 + 5.046^2)} = 5.183$. A_1 and A_2 are the amplitudes of the harmonic part of the trend and the remainder, correspondingly. This enables us to find the angular tracking error $\theta = 0.061$ radians. The quality factor of the predicting model is computed as $G = \omega/\theta = 0.56/0.061 = 9$ year^{-1}.

5. *Determining the constant component of the remainder.* This is estimated as $q_4(t) = 2.329 = \text{const}$ by using the least squares technique.
6. *Predicting the solar activity.* The equations are obtained for the trend $Q(t)$ and the remainder $q(t)$ as well as for all four of its components $q_1(t), q_2(t), q_3(t)$, and $q_4(t)$. The difference equation for the remainder gives the step-by-step predictions for it. The sum of the trend and the remainder gives the single-level prediction of solar activity without improving the prediction result.

   In short-range predictions with the prediction interval of $T_{pr} = 11$ years, the accuracy of prediction is 0.64 (without adaptation of the degree of stability) and in long-range predictions with $T_{pr} = 33$ years, it is 1.54.

Correlation function of the transient component of the remainder

The transient component of the remainder is defined as $y_{tr} = q_1(t) + q_2(t)$ and its correlation function is calculated from

$$A_{tr}(\tau) = \sum_{i=1}^{N-\tau} q_i q_{i+\tau}, \quad \tau = 1, 2, \cdots, N. \tag{6.103}$$

All values of $A_{tr}(\tau)$ are normalized with respect to the maximum value and construct the correlational graph. The coherence time obtained is nine years, confirming the prediction interval on which it is possible to increase the accuracy considerably by optimal choice of the degree of the predicting model.

Further analysis on the problem is based on the theory of servosystems with the analysis of the dynamic equations of a tracking system [52], and we leave it to the reader. However, in long-range predictions it makes no sense to increase the time of transient process of the remainder beyond the correlation interval which is approximately three time constants of τ_c, according to the concepts of the information theory.

Two-level algorithm

Alternatively, the modeling of solar activity is conducted through the two-level algorithm of multilevel objective analysis using the balance criterion. The above examined optimization (structure and estimates of the predicting models) is performed for both levels as seasonal and annual predictions. With regard to the compatibility of the predictions, the balance criterion makes it possible to raise the prediction time of seasonal predictions to the prediction time of annual predictions.