STUDIES

ON

REVISED GMDH ALGORITHMS WITH APPLICATIONS

（改良形ＧＭＤＨとその応用に関する研究）

by

Tadashi Kondo

Doctoral Dissertation

Department of Precision Engineering

Osaka University

December 1978

ACKNOWLEDGEMENTS

The author would like to express his sincere appreciation to Professor S. Makinouchi and Dr. H. Tamura for their guidance and encouragement during the course of the author's graduate works and the thesis researches. The author also expresses his gratitude to Professors K. Nakagawa and Y. Suzuki of Osaka University for their valuable advices and comments for improving the manuscript. The author wishes to thank Professors H. Kawabe, H. Tsuwa, T. Yamada, T. Tsukizoe and N. Ikawa of Osaka University for their valuable discussions. The author is also greatly indebted to Professor T. Soeda of Tokushima University for his valuable discussions and for offering real air pollution data in Tokushima. The author's many thanks should be extended to Mr. K. Yamagata and other members of Professor Makinouchi's laboratory for their stimulating discussions and kind assistances.

All the numerical computations are carried out at the Computation Center of Osaka University.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

STUDIES

ON

REVISED GMDH ALGORITHMS WITH APPLICATIONS

by

Tadashi Kondo

ABSTRACT

In this thesis, the revised GMDH (Group Method of Data Handling) algorithms and their applications to environmental problems such as air pollution and river pollution problems are discussed.

A basic GMDH algorithm, originally proposed by Ivakhnenko in 1968, which is based on a principle of heuristic self-organization, is a useful technique of data analysis for identifying complex nonlinear systems under the statistical analysis of input-output data. This algorithm has many advantages to deal with modeling of real complex systems, however, the algorithm has many methodological limitations such that the algorithm needs many heuristics, and the identified results depend heavily on these heuristics.

In this thesis, the author proposes two kinds of new revised GMDH algorithms which eliminate the limitations in the basic GMDH. One is

the revised GMDH algorithm which generates optimal partial polynomials automatically in each selection layer, and therefore, much better flexibility for constructing a complete polynomial can be obtained compared with the basic GMDH algorithm. The other is the revised GMDH algorithm which generates optimal intermediate polynomials automatically instead of partial polynomials in each selection layer. The optimal intermediate polynomials express the direct relationship between the input and output variables and they are generated so as to minimize the prediction error evaluated by using all the data. Therefore, the physically meaningful structure can be identified when the characteristics of the system are well reflected in the data.

Then, these two revised GMDH algorithms are applied to environmental problems. As the first example, large-spatial pattern identification of air pollution by a combined model of source-receptor matrix and the revised GMDH algorithm of generating optimal partial polynomials is discussed. By using synthetic data obtained by the computer simulation of air pollution diffusion, the predicted results obtained from the combined model is compared with the results obtained from the source-receptor matrix model only, and also with the results obtained from the combined model of source-receptor matrix and the basic GMDH. As the second example, nonlinear modeling for short-term prediction of air pollution concentration by the revised GMDH of generating optimal partial polynomials is discussed. By using the time series data of SO_2 concentration, the wind velocity and the wind direction in Tokushima, Japan, a suitable model for predicting SO_2 concentration at a few hours

in advance is developed. The predicted results obtained by the revised GMDH model are compared with the results obtained by a linear regression model, a linear autoregressive model and a basic GMDH model. As the third example, nonlinear statistical modeling of steady state river quality by the revised GMDH of generating optimal intermediate polynomials is discussed. By using measured data of river quality such as BOD and DO concentrations in the Bormida river, Italy, two kinds of steady state models of river quality is developed. The predicted results obtained by the revised GMDH model are compared with the results obtained by a conventional physical model.

Each Chapter of this thesis is based on the following papers.

Chapter 2

[1] H. Tamura and T. Kondo: Revised GMDH algorithm using self-selection of optimal partial polynomials and its application to large-spatial air pollution pattern identification, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 13, No. 4, 351-357 (1977)

[2] H. Tamura and T. Kondo: Revised GMDH algorithm using prediction sum of squares (PSS) as a criterion for model selection, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 14, No. 5, 519-524 (1978)

Chapter 3

[3] T. Kondo and H. Tamura: Revised GMDH algorithm of self-selecting optimal intermediate polynomials using AIC, (in Japanese) Trans. Soc. Instr. Control Engineers. (forthcoming)

Chapter 4

[4] H. Tamura and T. Kondo: Large-spatial pattern identification of air pollution by a combined model of source-receptor matrix and revised GMDH, Proc. IFAC Sympo. on Environmental Systems Planning, Design and Control, 373-380, Kyoto (Aug. 1977)

[5] H. Tamura and T. Kondo: Nonlinear modeling for short-term prediction of air pollution concentration by a revised GMDH, Proc. International Conference on Cybernetics and Society, IEEE Syst., Man, Cybern. Society, 596-601, Tokyo and Kyoto (Nov. 1978)

Chapter 5

[6] H. Tamura and T. Kondo: Nonlinear modeling for the steady state river quality by a revised GMDH, (in Japanese) Trans. Soc. Instr. Control Engineers. (submitted)

CHAPTER 1 FUNDAMENTAL PRINCIPLES OF GMDH

1.1 Introduction

Recently, the contribution of the systems engineering to the complex large-scale problems such as environmental problems, traffic problems, resource problems, etc. has been eagerly desired. In these real systems, very many variables and parameters are contained, and it is very difficult to identify the systems characteristics exactly by using the knowledges of some specific sciences only. Basic GMDH (Group Method of Data Handling) proposed by A.G. Ivakhnenko, which is based on a method of heuristic self-organization, is a useful technique of data analysis for identifying a completly unknown nonlinear system using the input-output data [7~10].

In the basic GMDH algorithm, the following advantages can be found.

(a) Nonlinear systems can be identified easily by using a small number of input-output data.

(b) The structure of the model can be self-selected by using no a priori information on the system structure.

However, the basic GMDH algorithm includes many disadvantages, and therefore many attempts have been made to improve the algorithm since

it was proposed in 1968. Almost all the improvements on the GMDH are concerned with the procedures of constructing the proper model and with the criterion for the model selection.

In this Chapter, firstly the principle of heuristic self-organization, which is a basic concept of GMDH, is described. Secondly, the basic GMDH algorithm proposed by Ivakhnenko is shown, and its disadvantages are clarified. Then, some revised GMDH algorithms which have been proposed to overcome these disadvantages are shown. Finally, the motivation to this thesis research is mentioned.

1.2 Principle of Heuristic Self-Organization [5,8]

GMDH is based on a principle of heuristic self-organization which is a useful approach to various complex problems. The systems or programs of heuristic self-organization are defined as those which have a multilayered or a hierarchical algorithm and include the generators of random hypothesis, or combinations, and several layers of threshold self-sampling of useful information. In each layer, by applying random combinations to input variables, new variables, whose structures are more complex or whose characteristics are more improved than those of the input variables, are generated, and from these variables more effective variables can be self-selected. These operations are repeated until the desired characteristics of the variables begin to degenerate. By using heuristic self-organization, we can solve the problems which are too complex to trace all input-output relationships throughout the

system. That is, in heuristic self-organization, the notion of general integral influences which has a self-adjustment facility of the system by acting upon the multilayered structure, is used, and particular information of each component of the system is not necessary. The integral influence is a heuristic one which is determined according to the summary result of input and output responses. The simplest realization of integral influences is a threshold unit permitting only some inputs to pass. In the self-organization, heuristics, which are conjectures in evaluating a course of problem solution by man, i.e. are creative thought processes of man, play an important role. Man controls the course of the solution by continuously directing its way to the desired results by means of integral influences. That is why heuristic self-organization ensures an accuracy which could not be reached by the use of routine mathematical methods. From the mass selection of plants and animals, the hypothesis of selection, which is a basis of the heuristic self-organization, can be found. This hypothesis of selection has the threshold type unit of integral influence, each of which has a single optimal setting corresponding to the accuracy in the result. Three examples of self-organizing systems are shown in Fig. 1.1. The first example in Fig. 1.1 (a) is the well-known perceptron, the model of the brain perception function, designed by Rosenblatt [19]. The second example in Fig. 1.1 (b) is the structure of a system designed at the Stanford University, where the problem is to predict the structure of organic molecules [18]. The third example in Fig. 1.1 (c) is the structure of GMDH. In the following section, the basic GMDH algorithm based on the heuristic self-organization is shown.

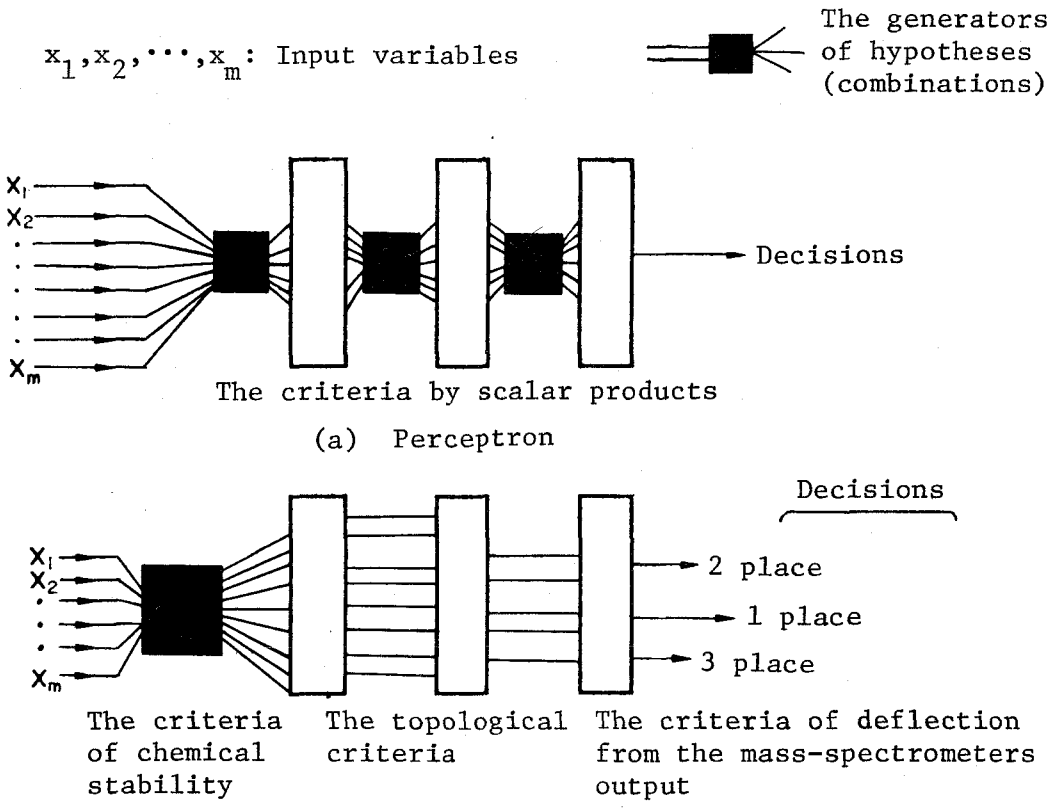


Fig. 1.1 Examples of system structures having heuristic self-organization [8]

1.3 Basic GMDH Algorithm [21]

The relationship between the input variables $(x_1, x_2, \cdots, x_m)$ and the output variable ϕ of the system is assumed to be written by

$$\phi = f(x_1, x_2, \cdots, x_m) . \tag{1.1}$$

Equation (1.1) is called as a complete description of the system. Many kinds of GMDH algorithms can be constructed for various kinds of complete descriptions such as polynomials, Bayes formulas, trigonometrical functions and rational expressions [2,8,11]. Among many kinds of complete descriptions, the Kolmogorov-Gabor polynomial

$$\phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k + \cdots \tag{1.2}$$

is most widely used, because almost all the real systems can be described as eq. (1.2) equivalently. Equation (1.2) can be constructed by combining the following second order polynomials of the two variables in multilayers.

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j + b_4 x_i^2 + b_5 x_j^2 \tag{1.3}$$

Here, y_k is called as the intermediate variable, and eq. (1.3) is called as the partial polynomial. The basic GMDH algorithm of constructing a proper Kolmogorov-Gabor polynomial is written as follows [9,10]:

Step 1:

Determine the input variables x_i $(i=1,2,\cdots,m)$ and the output variable ϕ. Normalize each variable if necessary. That is, each variable is transformed as

$$x'_{i\alpha} = \frac{x_{i\alpha} - \bar{x}_i}{\sqrt{V_{x_i}}} \quad , \quad (i=1,2,\cdots,m) \quad ; \quad \phi'_\alpha = \frac{\phi_\alpha - \bar{\phi}_\alpha}{\sqrt{V_\phi}} \quad . \qquad (1.4)$$

Here, $x_{i\alpha}$ is the α-th data of the input variable x_i, and $\bar{x}_i$ and V_{x_i} denote the mean value and the variance of x_i, respectively.

Step 2:

Divide the original data into two groups; the training data for estimating the coefficients of the partial polynomials, and the checking data for selecting the intermediate variables. The dividing rule is very heuristic. Usually the training and checking data are taken alternately or on the basis of the variance from the mean value.

Step 3:

For the combination of two variables x_i and x_j, estimate the parameters $(b_0,b_1,\cdots,b_5)$ contained in the partial polynomial of eq. (1.3) by using least square estimation for the training data.

Step 4:

Calculate the following mean square error for the checking data,

$$\Delta_{ch} = \frac{1}{N_{ch}} \sum_{\alpha=1}^{N_{ch}} (\phi_\alpha - y_{k\alpha})^2 \qquad (1.5)$$

by using the partial polynomial estimated in Step 3. Here, N_{ch} denotes the number of checking data, and $y_{k\alpha}$ denotes the α-th estimated value of the output under the k-th intermediate variable y_k. Select L intermediate variables which give L smallest mean square errors. This selection rule is also very heuristic. Equation (1.5) is called as a regularity criterion.

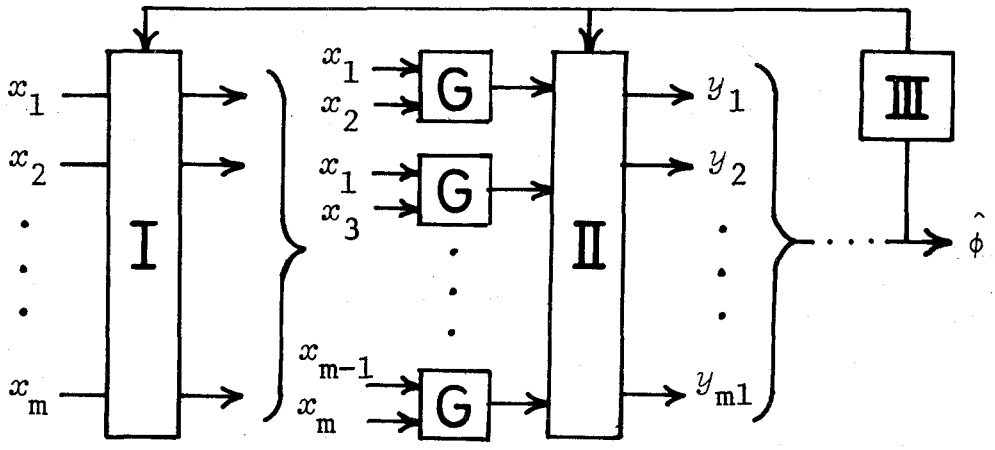
Step 5:

Replace x_i and x_j by y_i and y_j, respectively, and go to Step 3. Repeat Steps 3 to 5 until the smallest mean square error $\Delta_{ch(min)}$ cannot be improved.

In addition to the above procedure, we must optimize heuristics so as to find an optimal complete polynomial. The block diagram of the basic GMDH is shown in Fig. 1.2, where $\hat{\phi}$ denotes the estimated value of the output variable ϕ. In the above procedure, eq. (1.3) is used as a partial polynomial and the 2^T-th order complete polynomial with respect to the input variables can be constructed in the T-th layer. In order to identify the systems with various complexity more easily, some other partial polynomials have been proposed as follows:

(a) First order polynomial

$$y_k = b_0 + b_1 x_i + b_2 x_j \tag{1.6}$$



I : Division of the original data
II : Self-selection of the intermediate variables
III : Optimization of the threshold
G : Generator of the partial polynomial

Fig. 1.2 Block diagram of the basic GMDH

(b) Second order polynomials

As a bilinear partial polynomial

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j \tag{1.7}$$

has been proposed. As a partial polynomial which contain smaller number of parameters

$$X_k = w x_i + (1-w) x_j \qquad \text{w: weight} \tag{1.8.a}$$

$$y_k = b_0 + b_1 X_k + b_2 X_k^2 \tag{1.8.b}$$

has been proposed [4].

(c) High order polynomial [5,6]

$$X_i = b_0 + b_1 x_i + b_2 x_i^2 \tag{1.9.a}$$

$$X_j = b_0' + b_1' x_j + b_2' x_j^2 \tag{1.9.b}$$

$$y_k = c_0 + c_1 X_i + c_2 X_j + c_3 X_i X_j + c_4 X_i^2 + c_5 X_j^2 \tag{1.9.c}$$

We must predetermine the form of the partial polynomial which is used in the GMDH algorithm. This predetermination rule is also very heuristic. As is evident from above discussion, the following heuristics are contained in the basic GMDH algorithm.

H1. Determination for the division rule of the available input-output data into the training data and the checking data.

H2. Determination for the number of intermediate variables selected in each layer.

H3. Determination of the form of partial polynomial.

We must optimize these heuristics so as to find an optimal complete polynomial, and therefore we must repeat the GMDH computational procedure very many times by changing the heuristics. Furthermore, the basic GMDH algorithm involves following limitations to be solved.

L1. The identified model depends heavily on the heuristics H1, H2 and H3.

L2. When a second order or a higher order polynomial is used as a partial polynomial, the system, which has many input variables with low order polynomial, cannot be identified. The identified model with many input variables will become unnecessarily complex.

L3. The identified model fits well to the training data but not well to the checking data.

In the following section, improvements made on the basic GMDH algorithm in order to overcome these limitations are discussed.

1.4 Improvements of the Basic GMDH [21]

The methodological improvements of the basic GMDH algorithm are almost concerned with the procedure of constructing partial polynomials in order to overcome the limitation L2 and with the criterion for the model selection in order to overcome the limitation L3. Firstly, the GMDH algorithms, which are improved in the procedure of constructing partial polynomials, are shown.

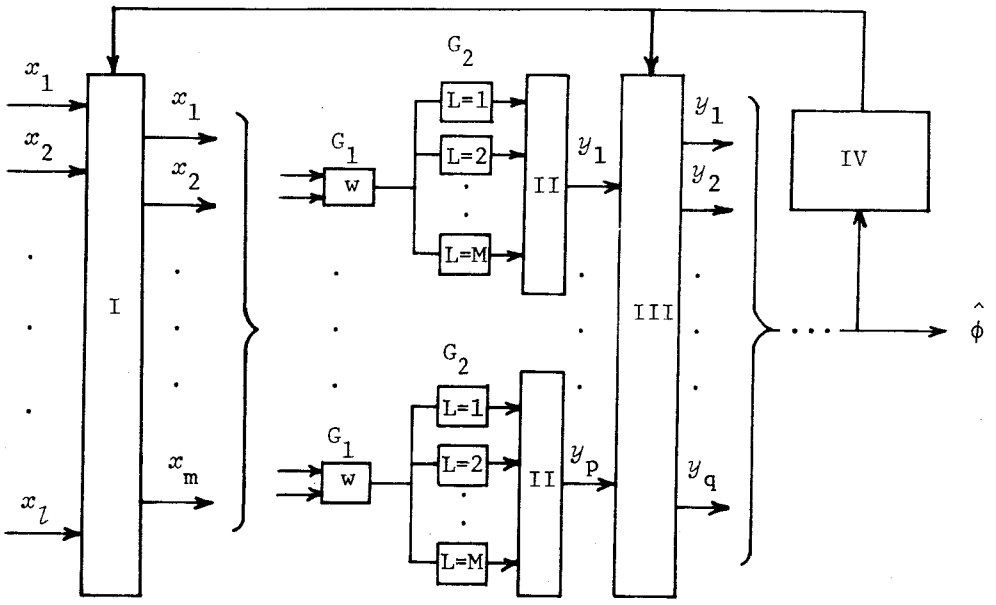
1) Algorithm of constructing optimal partial polynomials by stepwise regression under the statistical test for significance [1]

In this algorithm, heuristics H3 is not necessary, and limitation L2 and a part of limitation L1 are eliminated. But, limitation L3 is not eliminated because the structure of the partial polynomial is

determined by using only the training data.

2) Algorithm of self-selecting optimal partial polynomials so as to minimize the mean square error for the checking data [20]

In this algorithm, heuristics H3 is not necessary, and limitation L2 and a part of limitation L1 are eliminated. Furthermore, limitation L3 is considerably eliminated. The block diagram of this algorithm is shown in Fig. 1.3. The mean square error for the checking data is used



I : Self-selection of input variables
II : Self-selection of optimal partial polynomials
III : Self-selection of intermediate variables
IV : Optimization of threshold
G_1 : Filter, G_2 : Base function with a single input

Fig. 1.3 Block diagram of a revised GMDH [20]

in order to determine the structure of the complete polynomial. In other words, the mean square error for the checking data is used to generate the optimal partial polynomials and these polynomials are used to construct the multilayered structure. The self-selection procedures of optimal partial polynomials, in which the mean square error for the checking data is used as a selection criterion, are considered on the basis of eqs. (1.3) and (1.8). Application of this algorithm to an environmental problem of air pollution can be found in [20].

It has been reported that the revised GMDH algorithms as described above can construct more simplified complete polynomials and obtain better prediction accuracy than the basic GMDH algorithm.

Secondly, the GMDH algorithm, which are revised for the procedure of constructing the partial polynomial and for the criterion for the model selection, are shown.

3) Combination-generating GMDH algorithm using unbiasedness criterion [12,13]

This algorithm was proposed by Ivakhnenko.

(a) Procedure of constructing optimal partial polynomials by generating combinations of input variables [12]

Instead of using second order polynomial of eq. (1.3) as a partial polynomial, a combination, which gives the smallest value of unbiasedness index, is selected from combinations of two variables shown in Table 1.1, and an optimal partial polynomial is constructed.

Table 1.1 Constructing a partial polynomial [12]

	1	2	3	4	5	6
Number of polynomials	$2^0=1$	$2^1=2$	$2^2=4$	$2^3=8$	$2^4=16$	$2^5=32$
Right sides of the partial polynomials	$z_1=a_0$	$z_2=a_1x_2$ z_2+z_1	$z_3=a_2x_2^2$ z_3+z_1 z_3+z_2 $z_3+z_2+z_1$	$z_4=a_3x_1$ z_4+z_1 z_4+z_2 $z_4+z_2+z_1$ z_4+z_3 $z_4+z_3+z_1$ $z_4+z_3+z_2$ $z_4+z_3+z_2+z_1$	$z_5=a_4x_1x_2$ z_5+z_1 z_5+z_2 $z_5+z_2+z_1$ z_5+z_3 $z_5+z_3+z_1$ $z_5+z_3+z_2$ $z_5+z_3+z_2+z_1$ z_5+z_4 $z_5+z_4+z_1$	$z_6=a_5x_1^2$ z_6+z_1 z_6+z_2 $z_6+z_1+z_2$ z_6+z_3 $z_6+z_3+z_1$ .....

(b) Unbiasedness criterion [13]

This criterion is used in order to eliminate the limitation L3. Firstly, the available input-output data are divided into two groups A_1 and A_2. Here, the numbers of the data are R_1 and R_2, respectively. Secondly, the partial polynomial

$$y_k^* = f_1(x_i, x_j) \tag{1.10}$$

is estimated by using A_1 as the training data. The data A_2 are used as the checking data. Then the role of the data is exchanged. That is,

A_1 is used as the checking data and A_2 is used as the training data. The partial polynomial

$$y_k^{**} = f_2(x_i, x_j) \tag{1.11}$$

is estimated by using A_2. The unbiasedness index for the k-th combination in the T-th layer

$$n_{Tk} = \frac{1}{R_1+R_2} \sum_{\alpha=1}^{R_1+R_2} (y_{k\alpha}^{*} - y_{k\alpha}^{**})^2 \tag{1.12}$$

is calculated. Here, $y_{k\alpha}^{*}$ and $y_{k\alpha}^{**}$ denote the α-th values of y_k^{*} and y_k^{**}, respectively. Then, the unbiasedness criterion in the T-th layer

$$N_T = \frac{1}{F} \sum_{k=1}^{F} n_{Tk} \tag{1.13}$$

is calculated. Here, F is the number of the intermediate variables selected in the T-th layer.

(c) Combination-generating GMDH algorithm using unbiasedness criterion [12,13]

In the first layer, the available input-output data are divided into two groups, and for each combination of the two input variables the unbiasedness indexes for all the polynomials shown in Table 1.1 are calculated. Then the optimal partial polynomial, which gives the smallest unbiasedness index, is constructed. Then the F intermediate variables, which give F smallest unbiasedness indexes, are selected

and the unbiasedness criterion in eq (1.13) is calculated. In the second layer and above, the procedure in the first layer is repeated. The iterative computation is terminated when the value of N_T cannot be decreased.

In this algorithm, a part of limitation L1, and limitations L2 and L3 are eliminated. But, since the division of the data is used, the heuristics H1 is still necessary. Furthermore much more computation time is needed compared with the basic GMDH.

4) Algorithm using the structural and parametric stability [3]

(a) Structural stability

Firstly, the input-output data are divided into two groups. The partial polynomial is estimated by the least square estimation for the data 1, and the mean square error for the data 2 is calculated. Then, the role of the data is exchanged and the mean square error for the data 1 is calculated. The partial polynomial, which gives smaller sum of two mean square errors, is defined as structurally stable one.

(b) Parametric stability

In each second order polynomial, the term which has small difference between two values of each parameter estimated for data 1 and 2 is defined as parametrically stable one.

The algorithm using these two stability is shown as follows. The partial polynomial in each selection layer is constructed under the criterion of the parametric stability in the second order polynomial

of eq. (1.3), and the intermediate variables are selected under the criterion of the structural stability. The iterative computation is terminated when the value of the structural stability cannot be improved. The structural stability is considered as the equivalent idea to the unbiasedness criterion proposed by Ivakhnenko.

In this algorithm, a part of limitation L1, and limitations L2 and L3 are eliminated, but heuristics H1 is necessary because we need to divide the data into two groups.

The revised GMDH algorithms described above have been proposed in order to eliminate the three limitations L1, L2 and L3. But, these revised GMDH algorithms do not eliminate the three limitations completely. Especially, all of them need the heuristics H1 and therefore computational procedure of GMDH is to be repeated many times in order to find an optimal heuristics H1. In general, however, it is practically impossible to find the optimal division rule for each problem, and the identified model will depend heavily on the heuristics H1. The revised GMDH algorithm, which does not use the heuristics H1, i.e. which uses all the data as the training and at the same time as the checking data, is desired in order to obtain the optimal model which fits well to all the data.

Some other revised GMDH algorithms have been proposed from practical situations. Subsequently, these revised GMDH algorithms are shown briefly.

5) Algorithm using balance-of-variables criterion for the purpose of the long-term prediction [14,15]

In this algorithm proposed by Ivakhnenko, various procedures such as the procedure of using balance function or the procedure of using direct function and inverse function are included.

6) Sequential GMDH algorithm [6]

In this algorithm, the structure of the system is predetermined by using the basic GMDH algorithm, and when new input-output data are obtained, the estimates of the parameters are updated by using a sequential least square estimation method. It has been reported that this algorithm is useful to obtain stable predicted values for the time series sequence.

Besides the improvements described above, there are many studies on the GMDH by Ivakhnenko, et al., and they are summarized in [16,17].

1.5 Concluding Remarks and Motivation to This Research

In this Chapter, firstly the basic concept of GMDH algorithm which is called the heuristic self-organization is described. The heuristic self-organization is a very useful concept to solve engineering cybernetics problems which have very complex structure with large dimensionality. Secondly, the basic GMDH algorithm proposed by Ivakhnenko is shown. It is clarified that the basic GMDH algorithm involves three main limitations to be eliminated, and there exist many revised GMDH algorithms in order

to overcome this difficulty. But, the three limitations have not yet been eliminated completely.

In the following Chapters, the author will propose two kinds of new revised GMDH algorithms which eliminate these three limitations completely where some prediction error criterions will be used for model selection. By using the prediction error criterions, we try to develop the algorithms which do not need to divide the original data into two groups; the training data and the checking data. Furthermore, by using self-selection of optimal partial or intermediate polynomials in each selection layer, we try to eliminate the three limitations completely contained in the basic GMDH algorithm.

The revised GMDH algorithm of generating optimal partial polynomials under the prediction error criterion will be proposed in Chapter 2. This algorithm is supposed to be useful for identifying a very complex system as a statistical model, where we cannot, in general, obtain a physical interpretation for the model identified. Then, the revised GMDH algorithm of generating optimal intermediate polynomials under the prediction error criterion will be proposed in Chapter 3 where the intermediate polynomials generated in each selection layer express the direct relationship between the input and output variables. This algorithm is supposed to be useful for identifying physically meaningful structure of a relatively simple system when the characteristics of the system are well reflected in the input-output data.

REFERENCES

[1] Duffy, J.J. and M.A. Franklin: A learning identification algorithm and its application to an environmental system, IEEE Trans. Syst., Man, Cybern., Vol. SMC-5, No. 2, 226-240 (1975)

[2] Dylbokova, D.L. and I.S. Dylbokov: Prediction of trends of development of digital computers by GMDH using linear-fraction partial descriptions and balance-of-variables criterion, Soviet Automatic Control, Vol. 8, No. 2, 24-30 (1975)

[3] Endo, A.: Identification of a nonlinear system with the modified GMDH, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 14, No. 2, 130-135 (1978)

[4] Ihara, J.: Improved GMDH — A case of dynamical world population models, (in Japanese) Systems and Control, Vol. 19, No. 4, 201-210 (1975)

[5] Ikeda, S. and Y. Sawaragi: GMDH (Heuristic self-organization) and identification and prediction of complex systems, (in Japanese) Proc. Soc. Instr. Control Engineers, Vol. 14, No. 2, 185-195 (1975)

[6] Ikeda, S., M. Ochiai and Y. Sawaragi: Sequential GMDH algorithm and its application to river flow prediction, IEEE Trans. Syst., Man, Cybern., Vol. SMC-6, No. 7, 473-479 (1976)

[7] Ivakhnenko, A.G.: The group method of data handling, A rival of the method of stochastic approximation, Soviet Automatic Control, Vol. 1, No. 3, 43-55 (1968)

[8] Ivakhnenko, A.G.: Heuristic self-organization in problems of engineering cybernetics, Automatica, Vol. 6, No. 2, 207-219 (1970)

[9] Ivakhnenko, A.G.: Polynomial theory of complex systems, IEEE Trans. Syst., Man, Cybern., Vol. SMC-1, No. 4, 364-378 (1971)

[10] Ivakhnenko, A.G., et al.: Group handling of data in identification of the static characteristic of a multi-extremal plant, Soviet Automatic Control, Vol. 2, No. 2, 30-37 (1969)

[11] Ivakhnenko, A.G. and M.M. Todua: Prediction of random processes using self-organization of the prediction equations, Part 1 Problems of simple medium-term prediction, Soviet Automatic Control, Vol. 5, No. 3, 35-52 (1972)

[12] Ivakhnenko, N.A. and M.Z. Kvasko: Combination-generating GMDH algorithms in which the regularity of both symmetrical and nonsymmetrical polynomials is checked, Soviet Automatic Control, Vol. 5, No. 5, 33-38 (1972)

[13] Ivakhnenko, A.G., et al.: Discovery of physical laws by GMDH method with the absence-of-bias criterion, Soviet Automatic Control, Vol. 6, No. 6, 32-45 (1973)

[14] Ivakhnenko, A.G., et al.: Long-term prediction of random processes by GMDH algorithms using the unbiasedness criterion and balance-of-variables criterion, Part 1, Soviet Automatic Control(S.A.C.), Vol. 7, No. 4, 40-45 (1974); Part 2, S.A.C., Vol. 8, No. 4, 24-38 (1975); Part 3, S.A.C., Vol. 9, No. 2, 28-42 (1976); Part 4, S.A.C., Vol. 9, No. 4, 16-27 (1976)

[15] Ivakhnenko, A.G. and B.K. Svetal'skiy: Self-organization of world dynamics model according to Forrester's data and control synthesis by selecting the vertices of the hypercube of feasible controls, Soviet Automatic Control, Vol. 8, No. 1, 25-40 (1975)

[16] Ivakhnenko, A.G.: Present state of the theory of computer-aided self-organization of mathematical models (survey), Soviet Automatic Control, Vol. 8, No. 5, 18-26 (1975)

[17] Ivakhnenko, A.G.: The group method of data handling in prediction problems, Soviet Automatic Control, Vol. 9, No. 6, 21-30 (1976)

[18] Paterson, D.: Computers that hypothesize, New Scient., Vol. 39, No. 612, 442-443 (1968)

[19] Rosenblatt, F.: Principles of Neurodynamics, Spartan Books, Washington (1962)

[20] Tamura, H. and T. Kondo: Revised GMDH algorithm using self-selection of optimal partial polynomials and its application to large-spatial air pollution pattern identification, (in Japanese) Trans. Soc. Instr. Control Enginners, Vol. 13, No. 4, 351-357 (1977)

[21] Tamura, H. and T. Kondo: Recent trends in GMDH algorithms and their applications, (in Japanese) Operations Research, Vol. 23, No. 2, 104-111 (1978)

CHAPTER 2 REVISED GMDH OF GENERATING OPTIMAL PARTIAL POLYNOMIALS UNDER THE PREDICTION ERROR CRITERION

2.1 Introduction

The GMDH algorithm, which is based on a method of heuristic self-organization [10], is a useful technique of data analysis for identifying a completely unknown nonlinear system using the input-output data. In the basic GMDH [11,12] developed by Ivakhnenko, the concept of so called regularization is introduced for the purpose of avoiding the overfitting for the past data. Namely, the available input-output data are divided into the training data for estimating the coefficients in the partial polynomials, and the checking data for selecting the intermediate variables. In the basic GMDH algorithm, we need the following heuristics.

(a) Predetermination of the structure of the partial polynomials

(b) Division of the original data into two sets; the training data and the checking data

(c) Predetermination of the number of the intermediate variables

These heuristics are to be changed so as to find an optimal complete polynomial. Therefore, the computational procedure of the basic GMDH

must be repeated many times, but the complete polynomial obtained is not always an optimal one. Furthermore, the identified results depend heavily on these heuristics.

In this Chapter, we propose a revised GMDH algorithm which does not need any heuristics. In the basic GMDH algorithm, the structure of the partial polynomials is fixed to a predetermined description for all possible combinations of two variables. The revised GMDH algorithm proposed in this Chapter is the one which automatically generates optimal partial polynomials in each selection layer, and the polynomials as such are used to construct a complete polynomial in the multilayered structure. Therefore, the identified results do not depend on the heuristics of determining the structure of the partial polynomials and much better flexibility for constructing a complete polynomial can be obtained compared with the basic GMDH algorithm. Furthermore, in the revised GMDH algorithm proposed in this Chapter, all the data can be used as the training data and at the same time as the checking data, where instead of the mean square error for the checking data the Prediction Sum of Squares (PSS) [4] or Akaike's Information Criterion (AIC) [1,2,3] calculated from these data can be used as a criterion for generating partial polynomials, for selecting intermediate variables and for stopping the multilayered calculation. Therefore, the identified results do not depend on the heuristics of dividing the data into two sets. In the revised GMDH algorithm, the number of the intermediate variables is preferred to be as large as possible in order

to minimize PSS or AIC. That is, the number of the intermediate variables is determined not by the heuristics but by the upper limit of the memory capacity of the computer.

Firstly, we discuss the partial descriptions used in the previous GMDH algorithms. In the previous GMDH algorithms, the mean square error for the checking data has not been used for determining the structure of the optimal partial polynomials and for estimating the parameters in the partial polynomials. Therefore, the valuable information contained in the checking data is disregarded to construct the partial polynomials, and, as the results, the identified model does not fit well to the checking data. Secondly, the methods of computing the prediction errors; Prediction Sum of Squares (PSS) and Akaike's Information Criterion (AIC) are shown. By using these prediction errors as a criterion for model selection, we can construct an optimal model which fits well to all the data. Then, the revised GMDH algorithm of self-selecting partial polynomials under the criterion of PSS or AIC is developed. Since any heuristics are not needed in this revised GMDH, we do not need to repeat the computational procedures of the revised GMDH. The revised GMDH algorithm is applied to a simple illustrative example and the results are compared with those obtained by the basic GMDH algorithm.

2.2 Partial Polynomials Used in the Previous GMDH Algorithms [14]

There are many kinds of GMDH algorithms in which many kinds of

complete descriptions such as polynomials, rational expressions, Bayes formulas are used. Here, we use the Kolmogorov-Gabor polynomial

$$\phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k + \cdots \tag{2.1}$$

as a complete description of the system. In what follows, we show some kinds of polynomials which have been proposed as partial polynomials.

1) First order polynomial

$$y_k = b_0 + b_1 x_i + b_2 x_j \tag{2.2}$$

By using this polynomial, we can construct a first order complete polynomial.

2) Second order polynomial

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j \tag{2.3}$$

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j + b_4 x_i^2 + b_5 x_j^2 \tag{2.4}$$

By using these second order polynomials, we can construct a 2^T-th order polynomial after passing the T-th selection layer. As a second order polynomial, which contains a smaller number of parameters than eqs. (2.3) and (2.4), the following partial

polynomial has been proposed [8].

$$X_k = wx_i + (1-w)x_j \qquad \text{w: weight} \tag{2.5.a}$$

$$y_k = b_0 + b_1 X_k + b_2 X_k^2 \tag{2.5.b}$$

3) High order polynomial [9]

$$X_i = b_0 + b_1 x_i + b_2 x_i^2 \tag{2.6.a}$$

$$X_j = b_0' + b_1' x_j + b_2' x_j^2 \tag{2.6.b}$$

$$y_k = c_0 + c_1 X_i + c_2 X_j + c_3 X_i X_j + c_4 X_i^2 + c_5 X_j^2 \tag{2.6.c}$$

By using this polynomial, we can construct a 4^T-th order polynomial after passing the T-th selection layer.

4) Optimal partial polynomials for each combination of two variables

(a) Optimal partial polynomial in which parametric unstable terms contained in eq. (2.4) are eliminated [7].

(b) Optimal partial polynomial in which unnecessary terms contained in eq. (2.4) are eliminated by applying stepwise regression method [5] using residual sum of squares (RSS) for the training data [6].

By using these optimal partial polynomials, we can construct a complete polynomial of various order between the first and the 2^T-th order after passing the T-th selection layer.

The relationship between the number of selection layer and the order of a complete polynomial is shown in Fig. 2.1. The order of the complete polynomial constructed by the optimal partial polynomials in 4) (a) and (b) is shown by round mark. By using partial polynomials of eqs. (2.3)~(2.6), the system, which is represented by a polynomial having many input variables with low order, cannot be identified because the order of the complete polynomial is doubled in each selection layer. By using optimal partial polynomials in 4) (a) and (b), a system as such can be identified. That is, much broader kinds of systems can be identified by using optimal partial polynomials but not by using predetermined polynomials for all possible combinations of two variables. Furthermore, the number of terms contained in a complete polynomial can be decreased by using optimal partial polynomials. But, in 4) (a), a lot of computation time are needed in order to construct an optimal partial polynomial. And, in 4) (b), it is difficult to find the optimal standard value of variable selection, and furthermore the valuable information in the checking data cannot be used to construct the partial polynomials.

In order to cope with the disadvantages contained in the previous partial polynomials as described above, we propose a revised GMDH algorithm which generates in each selection layer an optimal partial polynomial which minimize the prediction error.

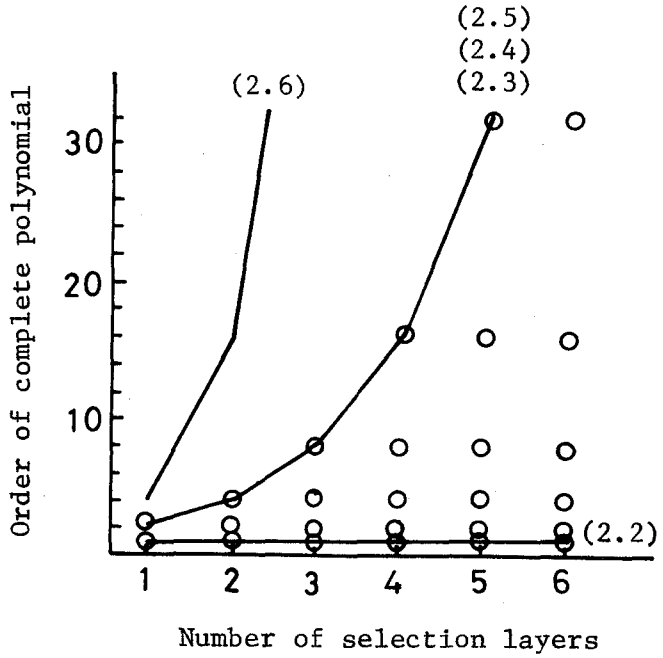


Fig. 2.1 Order of complete polynomials obtained for different partial polynomials

2.3 Prediction Sum of Squares (PSS) and Akaike's Information Criterion (AIC)

1) Computation of PSS [4]

In a multiple regression analysis, PSS is used as a criterion for selecting the independent variables, and the optimal regression equation

$$\hat{z}_\alpha = b_0 + \sum_{i=1}^{m} b_i x_{i\alpha}, \qquad \alpha=1,2,\cdots,n$$

is constructed so as to minimize PSS.

PSS is defined as

$$PSS = \sum_{\alpha=1}^{n} (z_\alpha - \hat{z}_\alpha^*)^2 \tag{2.7}$$

where

$$\hat{z}_\alpha^* = b_{0\alpha} + \sum_{i=1}^{m} b_{i\alpha} x_{i\alpha}, \qquad \alpha=1,2,\cdots,n \quad .$$

Here, n denotes the data length, z_α is the α-th actual value, and $\hat{z}_\alpha^*$ is the α-th estimated value obtained by a multiple regression analysis of all the data except the α-th data. In order to compute PSS of eq. (2.7), the multiple regression analysis must be repeated n times, therefore the amount of computation increases as the increase of the number of data. For this reason, when there are many data, it is not practical to compute PSS in the form of eq. (2.7).

PSS of eq. (2.7) can be reduced to [13]

$$PSS = \sum_{\alpha=1}^{n} \left(\frac{z_\alpha - \hat{z}_\alpha}{1 - \underline{x}_\alpha^T (X^T X)^{-1} \underline{x}_\alpha}\right)^2 \tag{2.8}$$

where

$$\hat{z}_\alpha = b_0 + \sum_{i=1}^{m} b_i x_{i\alpha}, \qquad \alpha=1,2,\cdots,n$$

$$\underline{x}_\alpha^T = (1, x_{1\alpha}, x_{2\alpha}, \cdots, x_{m\alpha})$$

$$X^T = [\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n] \quad .$$

Here, $\hat{z}_\alpha$ is the α-th estimated value obtained by a regression analysis of all the data. In this procedure, we do not need to repeat the regression analysis.

2) Computation of AIC [1,2,3]

In a multiple regression analysis, AIC is also used as a criterion for selecting the independent variables, and the optimal regression equation is constructed so as to minimize AIC. The basic statistics of AIC is defined as

$$\text{AIC} = -2 \log_e (\text{Maximum Likelihood}) + 2k, \tag{2.9}$$

where k is the number of parameters in the model to be adjusted to attain the maximum of the likelihood. Our identification procedure is realized by adopting the model which gives the minimum of AIC within a set of possible alternative complete polynomials. By this procedure, we are trying to minimize the expected deviation of the fitted distribution from the true distribution as measured by Kullback-Leibler's mean amount of information for discrimination. The information theoretic justification of the use of AIC for this purpose for independent observations can be found in [2,3]. For a linear regression analysis, AIC is reduced to

$$\text{AIC} = n \log_e S_k^2 + 2k + C \tag{2.10}$$

$$s_k^2 = \frac{1}{n} \sum_{\alpha=1}^{n} (z_\alpha - \hat{z}_\alpha)^2 \qquad (2.11)$$

where n denotes the data length, C is a constant, $\hat{z}_\alpha$ is the α-th estimated value obtained by a regression analysis of all the data, and z_α is the α-th observed value. Here, it is assumed that the noises contained in the model are mutually independent and normally distributed.

2.4 Revised GMDH Algorithm Using PSS or AIC as a Criterion for Model Selection [15]

In a GMDH algorithm, PSS or AIC calculated from all the data can be used as a criterion for generating optimal partial polynomials in each selection layer, for selecting intermediate variables and for stopping the multilayered iterative computation. The significant advantage of using PSS or AIC for model selection is that it is not necessary to divide the available data into the training data and the checking data. All the data can be used for constructing the model and at the same time for evaluating the prediction error, since PSS and AIC have an ability to evaluate the prediction error incurred by the model. Therefore, the identified results do not depend on the heuristics for dividing the data into the training data and the checking data, as it does in the basic GMDH algorithm. Furthermore, much better

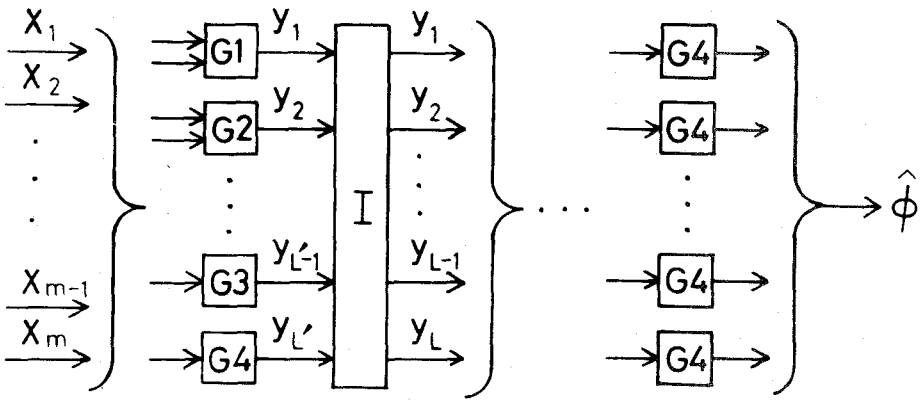
flexibility for constructing a complete polynomial can be obtained compared with the basic GMDH algorithm.

The block diagram of the revised GMDH algorithm using AIC is shown in Fig. 2.2. Here, it is assumed that the complete description of the system can be written as the Kolmogorov-Gabor polynomial

$$\phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \cdots \quad . \tag{2.12}$$

The revised GMDH algorithm is constructed by the following four procedures:

1) Generating optimal partial polynomials in each selection layer



I : Self-selection of intermediate variables

Fig. 2.2 Block diagram of the revised GMDH of generating optimal partial polynomials

The optimal partial polynomials can be generated through G1, G2, G3 and G4 applying a stepwise regression procedure [5] for the input variables to the following second order polynomial,

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j + b_4 x_i^2 + b_5 x_j^2 \ . \tag{2.13}$$

In this stepwise regression procedure, PSS or AIC is used as a criterion for selecting dominant variables in eq. (2.13). The normal equation for this polynomial can be written as

$$X^T X \, \underline{B} = X^T \underline{Y} \tag{2.14}$$

where $\underline{B}=(b_0, b_1, \cdots, b_5)^T$ and $\underline{Y}=(\phi_1, \phi_2, \cdots, \phi_n)^T$. For the normal equation (2.14), 7×13 matrix

$$\left[\begin{array}{c|c|c} X^T X & X^T \underline{Y} & I \\ \hline \underline{Y}^T X & \underline{Y}^T \underline{Y} & \underline{0}^T \end{array}\right]$$

$$= \left[\begin{array}{cccc|c|c} \sum 1 & \sum x_{i\alpha} & \cdots & \sum x_{j\alpha}^2 & \sum \phi_\alpha & \\ \sum x_{i\alpha} & \sum x_{i\alpha}^2 & \cdots & \sum x_{i\alpha} x_{j\alpha}^2 & \sum \phi_\alpha x_{i\alpha} & I \\ \vdots & \vdots & & \vdots & \vdots & \\ \sum x_{j\alpha}^2 & \sum x_{j\alpha}^2 x_{i\alpha} & \cdots & \sum x_{j\alpha}^4 & \sum \phi_\alpha x_{j\alpha}^2 & \\ \hline \sum \phi_\alpha & \sum \phi_\alpha x_{i\alpha} & \cdots & \sum \phi_\alpha x_{j\alpha}^2 & \sum \phi_\alpha^2 & \underline{0}^T \end{array}\right] \tag{2.15}$$

is constructed, where I is a unit matrix, $\underline{0}^T$ is a zero vector, and the 7-th row is supplemented for computing RSS (Residual Sum of Squares) which expresses the accuracy of fitting for all the data.

By using this matrix, we can select the dominant input variables contained in eq. (2.13) easily. That is, when the m-th variable in eq. (2.13) is to be entered in the partial polynomial, the (m+1)-th column is reduced to the unit vector of the (m+8)-th column using a pivoting operation. On the other hand, when the m-th variable in eq. (2.13) is to be deleted from the partial polynomial, the (m+8)-th column is reduced to the unit vector of the (m+1)-th column using a pivoting operation. These selection procedures are repeated alternately based on PSS of eq. (2.8) or AIC of eq. (2.10), where the dominant input variables are selected so as to minimize PSS or AIC. Optimal partial polynomials can be constructed by using the selected input variables. Four kinds of the generators of the optimal partial polynomials are shown in Fig. 2.3.

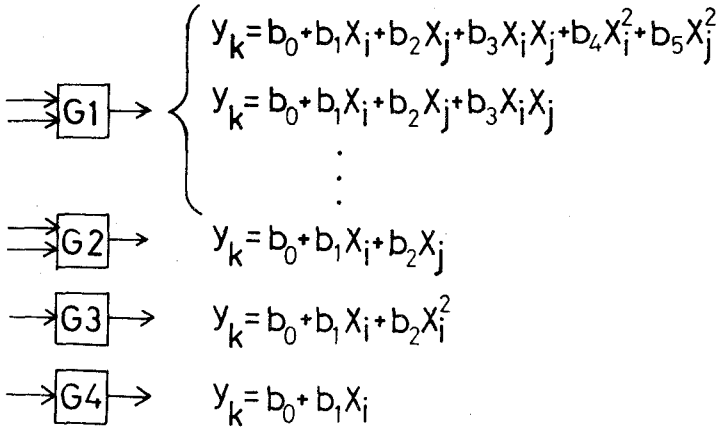


Fig. 2.3 Generators of the optimal partial polynomials

The generator G4 generates the same intermediate variable as that generated in the previous layer.

2) Selecting the intermediate variables

The L intermediate variables, which give the L smallest PSS or AIC, are selected from all the intermediate variables. The number L is preferred to be as large as possible in order to minimize PSS or AIC. That is, L is not determined in a heuristic manner but by taking into account the upper limit of the memory capacity of the computer.

3) Stopping the multilayered iterative computation

When all the generators of the optimal partial polynomials in the selection layer become G4, the iterative computation of the revised GMDH is terminated, because PSS or AIC cannot be decreased any more.

4) Computation of the predicted values

The prediction model is obtained as a weighted average of complete polynomials which are constructed by the intermediate variables remaining in the final layer. Since we can compare the predicted values obtained from several complete polynomials in the final layer, it is possible to exclude the abnormal predicted values before we obtain the final predicted value as a weighted average. Therefore, a stable prediction can be realized.

Since the revised GMDH algorithm described in this Chapter does not need any heuristics, we do not need to repeat the computational procedure for different heuristics.

2.5 Numerical Example [15]

Input and output relationship is assumed as

$$\phi = (0.1 + 0.2x_1 + 0.3x_2 + 0.4x_3)^2 . \qquad (2.16)$$

As input variables, $x_i (i=1,\cdots,4)$ are used. Here, the variable x_4 is not contained in eq. (2.16). The data used for modeling are shown in Table 2.1, and the data used for model validation are shown in Table 2.2.

Firstly, the numerical results obtained by the basic GMDH are shown. Four variables, which give the four smallest mean square errors for the checking data, are selected as the intermediate variables. Thirteen data in Table 2.1 are used as the interpolation points. The interpolation points are divided into the training data and the checking data in proportion of 7 : 6 and two cases are considered as follows:

Case 1: (Tr.) 1~7-th data
(Ch.) 8~13-th data

Case 2: (Tr.) odd-numbered data
(Ch.) even-numbered data

Table 2.1 Input data at the interpolation points

No.	x_1	x_2	x_3	x_4	ϕ
1	0.0	0.0	5.0	5.0	4.4
2	1.0	3.0	1.0	4.0	2.6
3	2.0	5.0	4.0	3.0	13.0
4	3.0	2.0	2.0	2.0	4.4
5	4.0	0.0	3.0	1.0	4.4
6	5.0	4.0	2.0	0.0	9.6
7	0.0	5.0	4.0	1.0	10.2
8	1.0	1.0	1.0	2.0	1.0
9	2.0	0.0	5.0	3.0	6.2
10	3.0	2.0	0.0	4.0	1.7
11	4.0	5.0	4.0	5.0	16.0
12	5.0	3.0	1.0	4.0	5.8
13	0.0	0.0	3.0	3.0	1.7

Table 2.2 Input data at the prediction points

No.	x_1	x_2	x_3	x_4	ϕ
1	1.0	1.0	2.0	2.0	2.0
2	2.0	5.0	3.0	1.0	10.2
3	3.0	4.0	1.0	0.0	5.3
4	4.0	0.0	4.0	1.0	6.2
5	5.0	3.0	0.0	2.0	4.0
6	0.0	5.0	5.0	3.0	13.0
7	1.0	2.0	1.0	4.0	1.7
8	2.0	0.0	4.0	5.0	4.4
9	3.0	4.0	2.0	4.0	7.3
10	4.0	5.0	3.0	3.0	13.0

Prediction models identified by the basic GMDH are as follows:

Case 1 : (First layer)

$$y_1 = 4.277 + 0.762x_1 - 3.386x_2 + 0.239x_1x_2 - 0.161x_1^2 + 0.900x_2^2$$

Here, as a Case 1', we describe the prediction model of the second layer in order to show the increase of complexity according to the increase of the layer.

Case 1': (Second layer)

$$z_1 = 1.100 - 1.301y_3 + 2.025y_4 - 0.226y_3y_4 + 0.273y_3^2 - 0.068y_4^2$$

$$y_3 = -13.80 + 2.642x_2 + 9.713x_3 - 0.605x_2x_3 + 0.195x_2^2 - 1.214x_3^2$$

$$y_4 = -3.847 - 1.929x_1 + 9.115x_3 + 0.177x_1x_3 + 0.241x_1^2 - 1.423x_3^2$$

Case 2 : (First layer)

$$y_1 = 18.74 + 3.929x_1 - 9.910x_4 + 1.089x_1x_4 - 1.619x_1^2 + 1.409x_4^2$$

The variables selected in the model are as follows:

Case 1 : $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_1\underline{x}_2$, $\underline{x}_1^2$, $\underline{x}_2^2$

Case 1': $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$, $\underline{x}_1\underline{x}_2$, $\underline{x}_1\underline{x}_3$, $\underline{x}_2\underline{x}_3$, $\underline{x}_1^2$, $\underline{x}_2^2$, $\underline{x}_3^2$, $x_1x_2x_3$, $x_1x_2^2$, $x_1x_3^2$, $x_1x_2x_3^2$, $x_1x_2^2x_3$, $x_1x_3^2$, $x_1^2x_2$, $x_1^2x_3$, $x_1^2x_2x_3$, $x_1^2x_2^2$,

$x_1^2x_3^2$, x_1^4, x_1^3, $x_1^3x_3$, $x_2x_3^2$, $x_2^2x_3$, $x_2x_3^3$, $x_2^2x_3^2$, x_2^4, x_2^3, $x_2^3x_3$, x_3^3, x_3^4

Case 2 : $\underline{x_1}$, $\underline{x_1^2}$, x_4, x_1x_4, x_4^2

where the variables contained in the proper model eq. (2.16) are shown with underline. Comparing the result of Case 1 with that of Case 1', we can see that the model becomes very complex according to the increase of the layer. The model of Case 2 is essentially different from eq. (2.16).

The accuracy at the interpolation points can be shown as follows: The mean square errors for the training data and the checking data for Cases 1 and 1' are shown in Table 2.3 (a). The mean square errors for the training data and the checking data for Case 2 are shown in Table 2.3 (b). From Table 2.3 (a), we can see that the fitting for the training data is very accurate but for the checking data is very inaccurate. From Table 2.3 (b), we can see that the model of Case 2

Table 2.3 Change of mean square error at the interpolation points

(a) Case 1 and 1'

	1-st layer	2-nd layer
Training data	0.335	8.57×10^{-7}
Checking data	2.37	40.7

(b) Case 2

	1-st layer	2-nd layer
Training data	3.22	7.32×10^{-2}
Checking data	60.7	4.06×10^{4}

is not identified properly. The accuracy at the prediction points is evaluated by

$$J = \left(\frac{1}{10} \sum_{\alpha=1}^{10} \left| \frac{\phi_\alpha - \hat{\phi}_\alpha}{\phi_\alpha} \right| \right) \times 100 \qquad (2.17)$$

for ten data in Table 2.2, where $\hat{\phi}_\alpha$ denotes predicted value. The prediction accuracy obtained is

Case 1 : J = 30 %

Case 1': J = 129 %

Case 2 : J = 175 % .

The numerical results obtained by the revised GMDH are shown. Four variables, which give the four smallest values of PSS, are selected as the intermediate variables. Thirteen data in Table 2.1 are used as the interpolation points.

Prediction models identified by the revised GMDH are as follows: Two intermediate variables are remained in the final layer.

Prediction model 1: (Weight w_1=0.520)

$$v_1 = -0.080 + 0.571z_1 + 0.442z_2$$

$$z_1 = -0.534 + 0.932y_2 + 0.060y_2y_4 - 0.033y_2^2$$

$$z_2 = -0.935 + 0.868y_1 + 0.282y_4$$

$$y_1 = 4.171 - 3.632x_2 + 0.258x_1x_2 + 0.973x_2^2$$

$$y_2 = 0.237 + 0.996x_3 + 0.365x_2^2$$

$$y_4 = 1.198 + 0.960x_3 + 0.461x_1x_3$$

Prediction model 2: (Weight w_2=0.480)

$$v_2 = -0.113 + 0.707z_1 + 0.311z_4$$

$$z_1 = -0.534 + 0.932y_1 + 0.060y_2y_4 - 0.033y_2^2$$

$$z_4 = 0.000 + 1.000y_1$$

$$y_1 = 4.171 - 3.632x_2 + 0.258x_1x_2 + 0.973x_2^2$$

$$y_2 = 0.237 + 0.996x_3 + 0.365x_2^2$$

$$y_4 = 1.198 + 0.960x_3 + 0.461x_1x_3$$

The variables contained in the prediction models are

$$\underline{x}_2,\ \underline{x}_3,\ \underline{x}_2^2,\ \underline{x}_3^2,\ \underline{x}_1\underline{x}_2,\ \underline{x}_1\underline{x}_3,\ x_1x_3^2,\ x_2^2x_3,\ x_1x_2^2x_3,\ x_2^4.$$

The accuracy at the interpolation points can be shown as follows: The mean square errors for all the data (RSS/13) and the values of PSS (PSS/13) are shown in Table 2.4. From Table 2.4, we can see that

the difference between RSS and PSS in higher layers is smaller than in lower layers. The accuracy at the prediction points is again evaluated by eq. (2.17) for ten data in Table 2.2. The prediction accuracy obtained is

$$J = 10.8\ \%.$$

The combinations of the intermediate variables in each selection layer are shown in Fig. 2.4. The linear generators (G2,G4) are appearing more often in higher layer. The iterative computation of the revised GMDH is terminated at the fifth layer.

From the numerical example described above, the following results are obtained.

(a) From the results of Cases 1 and 1', the prediction model obtained by the basic GMDH becomes very complex as the selection layer increases. From the results of Cases 1 and 2, the identified results depend heavily on the way of dividing the original data into the training data and the checking data. The prediction model of Case 2 is essentially different from eq. (2.16) because even numbered data cannot be used for modeling and odd numbered data used for modeling do not contain sufficient information. Furthermore, in the basic GMDH, the identified model fits well to the training data, but not to the checking data.

(b) The prediction model obtained by the revised GMDH is constructed in the fifth layer but is not complex. Furthermore, a uniform accuracy for all the data can be obtained. The prediction accuracy of the revised GMDH is much better than that of the basic GMDH.

Table 2.4 Changes of RSS and PSS

	1-st layer	2-nd layer	3-rd layer	4-th layer	5-th layer
RSS / 13	1.29	0.642	0.450	0.450	0.450
PSS / 13	2.26	1.00	0.689	0.574	0.574

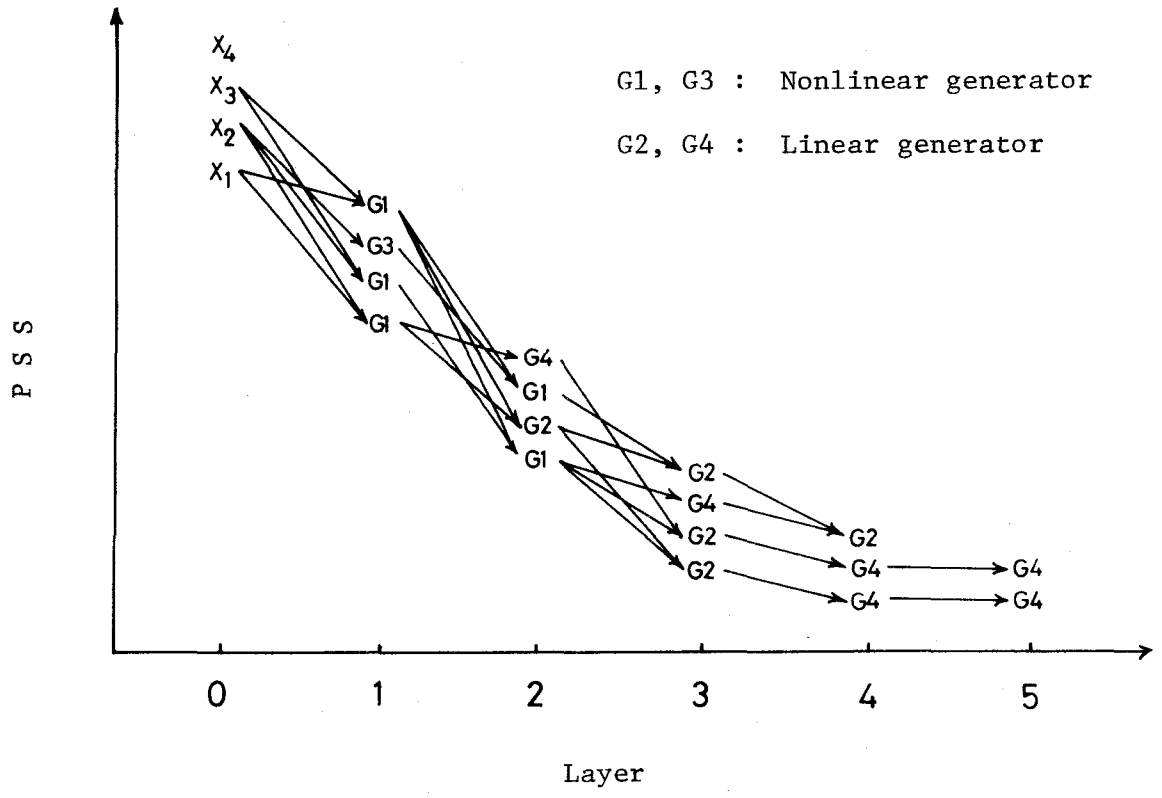


Fig. 2.4 Combinations of intermediate variables in each selection layer

2.6 Concluding Remarks

In this Chapter, a revised GMDH algorithm of generating optimal partial polynomials under the prediction error criterion is proposed. The algorithm is applied to a simple illustrative example and compared with the results obtained by the basic GMDH algorithm. The advantages of the revised GMDH compared with the basic GMDH are as follows:

(a) The revised GMDH algorithm based on PSS or AIC does not use the heuristics to divide the original data into two groups; the training data and the checking data. That is, all the data can be used as the training data and at the same time as the checking data. Therefore, a uniform accuracy for all the data can be obtained.

(b) The revised GMDH algorithm generates optimal partial polynomials in each selection layer so as to minimize PSS or AIC. Therefore, much better flexibility for constructing a complete polynomial can be obtained.

(c) Since any heuristics are not contained in the revised GMDH algorithm, we do not need to repeat the computational procedure for the different heuristics and the identified results do not depend on the heuristics.

The application of the revised GMDH in this Chapter to air pollution problems will be discussed in Chapter 4.

REFERENCES

[1] Akaike, H.: A new look at the statistical model identification, IEEE Trans. Automatic Control, Vol. AC-19, No. 6, 716-723 (1974)

[2] Akaike, H.: Information theory and an extension of the maximum likelihood principle, Proc. 2nd Int. Symposium on Information Theory, Akademiai Kiado, Budapest, 267-281 (1973)

[3] Akaike, H.: Automatic data structure search by the maximum likelihood, in "Computer in Biomedicine" Suppl. to Proc. 5th Hawaii Int. Conf. on System Sciences, 99-101 (1972)

[4] Allen, D.M.: The relationship between variable selection and data augmentation and a method for prediction, Technometrics, Vo. 16, No. 1, 125-127 (1974)

[5] Draper, N.R. and H. Smith: Applied Regression Analysis, Wiley, New York (1966)

[6] Duffy, J.J. and M.A. Franklin: A learning identification algorithm and its application to an environmental system, IEEE Trans. Syst., Man, Cybern., Vol. SMC-5, No. 2, 226-240 (1975)

[7] Endo, A.: Identification of a nonlinear system with the modified GMDH, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 14, No. 2, 130-135 (1978)

[8] Ihara, J.: Improved GMDH — A case of dynamical world population models, (in Japanese) Systems and Control, Vol. 19, No. 4, 201-210 (1976)

[9] Ikeda, S., M. Ochiai and Y. Sawaragi: Sequential GMDH algorithm and its application to river flow prediction, IEEE Trans. Syst., Man, Cybern., Vol. SMC-6, No. 7, 473-479 (1976)

[10] Ivakhnenko, A.G.: Heuristic self-organization in problems of engineering cybernetics, Automatica, Vol. 6, No. 2, 207-219 (1970)

[11] Ivakhnenko, A.G.: Polynomial theory of complex systems, IEEE Trans. Syst., Man, Cybern., Vol. SMC-1, No. 4, 364-378 (1971)

[12] Ivakhnenko, A.G., Y.V. Koppa, I.K. Tymchenko and N.O. Ivakhnenko: Group handling of data in identification of the static characteristic of a multi-extremal plant, Soviet Automatic Control, Vol. 2, No. 2, 30-37 (1969)

[13] Okuno, T., et al.: Multivariate Analysis, Continued, (in Japanese) Nikka-Giren, Tokyo (1976)

[14] Tamura, H. and T. Kondo: Revised GMDH algorithm using self-selection of optimal partial polynomials and its application to large-spatial air pollution pattern identification, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 13, No. 4, 351-357 (1977)

[15] Tamura, H. and T. Kondo: Revised GMDH algorithm using prediction sum of squares (PSS) as a criterion for model selection, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 14, No. 5, 519-524 (1978)

CHAPTER 3 REVISED GMDH ALGORITHM OF GENERATING OPTIMAL INTERMEDIATE POLYNOMIALS UNDER AKAIKE'S INFORMATION CRITERION

3.1 Introduction

In Chapter 2, we have proposed the revised GMDH algorithm [8] which generates optimal partial polynomials in each selection layer automatically by using prediction errors [1,2] as a criterion for model selection, and it is shown that this revised GMDH algorithm has many advantages compared with the previous GMDH algorithms [4,5,6]. Very complex systems, which contain many variables, can be identified by using the revised GMDH algorithm in Chapter 2, but, in general, it is difficult to identify physically meaningful structure between the input and output variables, because the partial polynomials, in which the intermediate variables are the input variables in each selection layer, have been estimated and accumulated in the multilayered structure.

In this Chapter, a revised GMDH algorithm, which generates optimal intermediate polynomials automatically instead of partial polynomials in each selection layer, is proposed. The optimal intermediate polynomials express the direct relationship between the input and output variables,

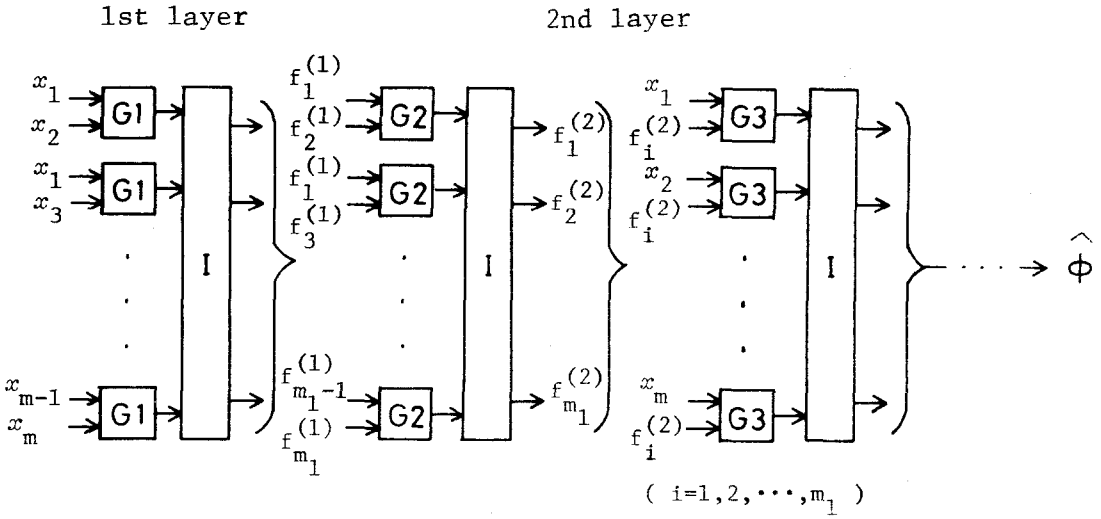
and they are generated so as to minimize the Akaike's Information Criterion (AIC) [1] evaluated by using all the data. Therefore, the physically meaningful structure can be identified when the characteristics of the system are well reflected in the data. This revised GMDH algorithm is applied to the input-output data observed in a simple kinetic system, and we try to discover the Newton's second law. The result obtained is compared with that obtained by the revised GMDH algorithm of generating partial polynomials.

3.2 Revised GMDH Algorithm of Generating Optimal Intermediate Polynomials [7]

In this section, we propose a revised GMDH algorithm which generates optimal intermediate polynomials automatically in each selection layer. In this algorithm, AIC calculated from all the data is used as the criterion for generating optimal intermediate polynomials in each selection layer, for evaluating intermediate polynomials and for stopping the multilayered iterative computation. Here, the heuristics of dividing the available data into two groups; the training data and the checking data, is not needed, and the structure and the parameters of intermediate polynomials are determined so as to minimize the prediction errors evaluated by using all the data. Namely, we select optimal intermediate polynomials in which unnecessary variables are eliminated by applying a stepwise regression procedure [3] using AIC as a criterion for model selection, and we terminate the iterative computation when the value of

AIC cannot be decreased any more. Here, the significance of using AIC as a criterion for generating intermediate polynomials is to obtain the best model by using smaller number of the input variables.

The block diagram of the revised GMDH algorithm is shown in Fig. 3.1, where m is a number of input variables, m_1 is a number of optimal intermediate polynomials selected in each selection layer and L_1 is a maximum number of the terms of the intermediate polynomials in each selection layer. The revised GMDH algorithm is constructed by the following procedures:



I : Self-selection of the optimal intermediate polynomials

G1, G2, G3 : Generators of the optimal intermediate polynomials

Fig. 3.1 Block diagram of the revised GMDH of generating optimal intermediate polynomials

A. Procedure in the first layer

In the first layer, the optimal intermediate polynomial is generated by applying a stepwise regression procedure for the input variables to the following p-th order polynomial,

$$\phi = a_0 + \sum_{i=1}^{2} a_i x_i + \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} a_{i_1 i_2} x_{i_1} x_{i_2} + \cdots + \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} \cdots \sum_{i_p=1}^{2} a_{i_1,i_2,\cdots,i_p} x_{i_1} x_{i_2} \cdots x_{i_p} . \qquad (3.1)$$

Here, eq. (3.1) is constructed with L terms. In this stepwise regression procedure, AIC is used as a criterion for selecting dominant variables in eq. (3.1). The normal equation for this polynomial can be written as

$$X^T X \underline{A} = X^T \underline{\Phi} \qquad (3.2)$$

where $\underline{A} = (a_0, a_1, \cdots, a_{L-1})^T$. For the normal equation (3.2), the $(L+1)\times(2L+1)$ matrix

$$\left[\begin{array}{c|c|c} X^T X & X^T \underline{\Phi} & I \\ \hline \underline{\Phi}^T X & \underline{\Phi}^T \underline{\Phi} & \underline{0}^T \end{array}\right] \qquad (3.3)$$

is constructed, where I denotes a L×L unit matrix, $\underline{0}^T$ denotes a zero vector, and (L+1)-th row is supplemented for computing RSS (Residual Sum of Squares) which expresses the accuracy of fitting for all the data. By using this matrix, we can select the combination of the dominant input variables which minimize AIC, and we can construct optimal intermediate polynomials from this combination.

Firstly, by applying Gauss-Jordan elimination procedure to the matrix (3.3), the first column is reduced to a unit vector by eliminating the non-diagonal elements. Then, we select the dominant input variables contained in eq. (3.1). That is, when the l-th variable in eq. (3.1) is entered in the intermediate polynomial, the $(l+1)$-th column is reduced to a unit vector by using Gauss-Jordan elimination procedure. On the other hand, when the l-th variable in eq. (3.1) is deleted from the intermediate polynomial, the $(L+l+2)$-th column is reduced to a unit vector. These selection procedures are repeated alternately so as to minimize AIC, and the dominant input variables are entered gradually into the intermediate polynomial. We terminate this procedure when the value of AIC cannot be decreased any more or when the (L_1-1) variables are selected in the intermediate polynomial. Optimal intermediate polynomial can be constructed by using the selected input variables. The procedure in the first layer is called as the generator G1 of optimal intermediate polynomials. Then, from ${}_mC_2$ intermediate polynomials generated in the first layer, the m_1 intermediate polynomials, which give the m_1 smallest AIC, are selected.

B. Procedure in the second layer

In the second layer, two kinds of combinations are considered.

1) Combination of two intermediate polynomials selected in the first layer

Let the i-th intermediate polynomial selected in the first layer be

$$\hat{\phi} = f_i^{(1)}(\underline{x}) \quad (i=1,2,\cdots,m_1) \tag{3.4}$$

where $\underline{x}$ is input variables, and it is assumed that eq. (3.4) contains $K_i^{(1)}$ ($\leq L_1-1$) variables. We combine two intermediate polynomials $f_i^{(1)}$ and $f_j^{(1)}$. Let the equation constructed by all the variables contained in $f_i^{(1)}$ and $f_j^{(1)}$ be

$$\hat{\phi} = f_i^{(1)}(\underline{x}) + f_j^{(1)}(\underline{x}) \ . \tag{3.5}$$

The normal equation for eq. (3.5) can be written as

$$[\ X_i^{(1)} \mid X_j^{(1)}\ ]^T [\ X_i^{(1)} \mid X_j^{(1)}\ ]\ \underline{A} = [\ X_i^{(1)} \mid X_j^{(1)}\ ]^T\ \underline{\Phi} \tag{3.6}$$

where $\underline{A} = (a_0, a_1, \cdots, a_{K_i^{(1)}+K_j^{(1)}+1})^T$. For the normal equation (3.6), the following $(K_i^{(1)}+K_j^{(1)}+3)\times(2K_i^{(1)}+2K_j^{(1)}+5)$ matrix

$$\left(\begin{array}{cc|c|cc} X_i^{(1)T}X_i^{(1)} & X_i^{(1)T}X_j^{(1)} & X_i^{(1)T}\underline{\Phi} & I & 0 \\ X_j^{(1)T}X_i^{(1)} & X_j^{(1)T}X_j^{(1)} & X_j^{(1)T}\underline{\Phi} & 0 & I \\ \hline \underline{\Phi}^T X_i^{(1)} & \underline{\Phi}^T X_j^{(1)} & \underline{\Phi}^T\underline{\Phi} & \underline{0}^T & \underline{0}^T \end{array}\right) \tag{3.7}$$

is constructed, where I is a unit matrix, $\underline{0}^T$ is a zero vector.

When $K_i^{(1)}$ variables contained in $f_i^{(1)}(\underline{x})$ are entered into the intermediate polynomial in the second layer, the matrix (3.7) is reduced to

$$\left(\begin{array}{cc|c|cc} I & M_{12} & M_{13} & M_{14} & 0 \\ 0 & M_{22} & M_{23} & M_{24} & I \\ \hline \underline{0}^T & M_{32} & M_{33} & M_{34} & \underline{0}^T \end{array}\right) . \tag{3.8}$$

Here, M_{13}, M_{14}, M_{33} and M_{34} have been already calculated in the first layer as

$$M_{13} = (X_i^{(1)T} X_i^{(1)})^{-1} X_i^{(1)T} \underline{\Phi} \tag{3.9}$$

$$M_{14} = (X_i^{(1)T} X_i^{(1)})^{-1} \tag{3.10}$$

$$M_{33} = \underline{\Phi}^T \underline{\Phi} - \underline{\Phi}^T X_i^{(1)} (X_i^{(1)T} X_i^{(1)})^{-1} X_i^{(1)T} \underline{\Phi} \tag{3.11}$$

$$M_{34} = - \underline{\Phi}^T X_i^{(1)} (X_i^{(1)T} X_i^{(1)})^{-1} . \tag{3.12}$$

The remaining parts of the matrix (3.8) can be obtained as

$$M_{12} = M_{14} (X_i^{(1)T} X_j^{(1)}) \tag{3.13}$$

$$M_{22} = X_j^{(1)T} X_j^{(1)} - (X_j^{(1)T} X_i^{(1)}) M_{12} \tag{3.14}$$

$$M_{32} = \underline{\Phi}^T X_j^{(1)} - (\underline{\Phi}^T X_i^{(1)}) M_{12} \tag{3.15}$$

$$M_{23} = X_j^{(1)T} \underline{\Phi} - (X_j^{(1)T} X_i^{(1)}) M_{13} \tag{3.16}$$

$$M_{24} = - M_{12}^T . \tag{3.17}$$

By using eqs. (3.13)~(3.17), we can construct the matrix (3.8) easily. Then, by applying a stepwise regression procedure to the matrix (3.8) in the same way as in the first layer, we can select a combination of the dominant input variables which minimize AIC, and we can construct an optimal intermediate polynomial from this combination.

In this procedure, when the number of selected variables exceeds (L_1-1), we try to decrease AIC under the following procedure. Firstly, from (L_1-1) variables which have been already contained in the intermediate polynomial, we find the variable, which gives the smallest increase of AIC, and delete it from the intermediate polynomial. Then, from the variables which have not yet been contained in the intermediate polynomial, we find the variable, which gives the biggest decrease of AIC, and enter it into the intermediate polynomial. We repeat this procedure alternately so as to minimize AIC. When the variable, which is deleted from the intermediate polynomial, is entered into the intermediate polynomial immediately, we terminate the iterative procedure and construct the optimal intermediate polynomial by using the selected input variables. The procedure in this part is called as the generator G2 of optimal intermediate polynomials. Then, from ${}_{m_1}C_2$ intermediate polynomials generated in this procedure, the m_1 intermediate polynomials, which give the m_1 smallest AIC, are selected.

2) Combination of the intermediate polynomial and the input variables

Let the i-th intermediate polynomial selected in the preceding combination be

$$\hat{\phi} = f_i^{(2)}(\underline{x}) \qquad (i=1,2,\cdots,m_1) \tag{3.18}$$

where $\underline{x}$ is input variables, and it is assumed that eq. (3.18) contains $K_i^{(2)}$ $(\leq L_1-1)$ variables. We combine the intermediate polynomials

$f_i^{(2)}(i=1,2,\cdots,m_1)$ with the input variables $x_j(j=1,2,\cdots,m)$. Let the equation constructed by all the variables contained in $f_i^{(2)}$ and $x_j f_i^{(2)}$ be

$$\hat{\phi} = f_i^{(2)}(\underline{x}) + x_j f_i^{(2)}(\underline{x}), \qquad (j=1,2,\cdots,m) \tag{3.19}$$

where $x_j f_i^{(2)}$ contains $(K_i^{(2)}+1)$ variables. The normal equation for eq. (3.19) can be written as

$$[\, X_i^{(2)} \mid X_{ji}^{(2)}]^T [\, X_i^{(2)} \mid X_{ji}^{(2)}]\, \underline{A} = [\, X_i^{(2)} \mid X_{ji}^{(2)}]^T\, \underline{\Phi} \tag{3.20}$$

where $\underline{A} = (a_0, a_1, \cdots, a_{2K_i^{(2)}+1})^T$. For the normal equation (3.20), the following $(2K_i^{(2)}+3)\times(4K_i^{(2)}+5)$ matrix

$$\left(\begin{array}{cc|c|cc} X_i^{(2)T}X_i^{(2)} & X_i^{(2)T}X_{ji}^{(2)} & X_i^{(2)T}\underline{\Phi} & I & 0 \\ X_{ji}^{(2)T}X_i^{(2)} & X_{ji}^{(2)T}X_{ji}^{(2)} & X_{ji}^{(2)T}\underline{\Phi} & 0 & I \\ \hline \underline{\Phi}^T X_i^{(2)} & \underline{\Phi}^T X_{ji}^{(2)} & \underline{\Phi}^T\underline{\Phi} & \underline{0}^T & \underline{0}^T \end{array}\right) \tag{3.21}$$

is constructed, where I is a unit matrix, $\underline{0}^T$ is a zero vector. Then, by applying the stepwise regression procedure to the matrix (3.21) in the same way as in the preceding combination, we select a combination of the dominant input variables, which minimizes AIC, and

we construct an optimal intermediate polynomial from this combination. The procedure in this part is called as the generator G3 of optimal intermediate polynomials. Then, from ($m_1 \times m$) intermediate polynomials generated in this part, the m_1 intermediate polynomials, which give the m_1 smallest AIC, are selected.

In the second layer, firstly we select dominant variables from all the variables contained in $f_i^{(1)}$ and $f_j^{(1)}$, and construct the intermediate polynomial $f_i^{(2)}$ by using the selected variables. Then we combine the input variables x_j ($j=1,2,\cdots,m$) with $f_i^{(2)}$ ($i=1,2,\cdots,m_1$) and construct the optimal intermediate polynomial in the second layer. On the other hand, instead of using above linear combinations, we can use a nonlinear combination of $f_i^{(1)}$ and $f_j^{(1)}$ directly such as a second order polynomial of two variables, but the number of the variables, which we must consider in selecting dominant variables, become very large, and it is not desirable in the practical situation. Furthermore it seems that the system can be identified more accurately by using linear combinations than by using a nonlinear combination such as a second order polynomial, because the model is becoming complex gradually in each selection layer.

C. Procedure in the 3rd, 4th, $\cdots$ layers

In the 3rd, 4th, $\cdots$ layers, the same procedure as in the second layer is repeated. The multilayered iterative computation is terminated in one of the following cases.

(a) The AIC is reduced to a very small value and the value of AIC cannot be decreased any more in the next layer.

(b) The structures of m_1 intermediate polynomials are the same forms as those of m_1 intermediate polynomials in the previous layer.

When the multilayered iterative computation is terminated as the result of Case (a), it indicates that without being disturbed by large noises, the nonlinear relationship between the input and output variables can be obtained accurately. That is, it seems to be the most probable that the physically meaningful relationship between the input and output variables is obtained. On the other hand, in Case (b), the relationship obtained between the input and output variables is not a physically meaningful one. When the multilayered iterative computation is terminated, the intermediate polynomial remained in the final layer is adopted as a complete polynomial of the system.

By using these three procedures A, B and C as described above, we can construct the revised GMDH algorithm which generates optimal intermediate polynomials automatically in each selection layer so as to minimize AIC evaluated by using all the data.

The parameters used in the revised GMDH algorithm are as follows:

p : maximum order of the intermediate polynomial in the first layer

L_1: maximum number of the terms in the intermediate polynomial

m_1: number of the intermediate polynomials selected in each layer.

These parameters are preferred to be as large as possible and are determined not by the heuristics but by the upper limit of the memory capacity of the computer. When we apply the revised GMDH algorithm to the real

system, the structure and the parameters of the identified model may be considerably different from those of the real system, because, in general, the measured data of the input and output variables contain noticeable measurement errors. Therefore, we must check the structure of the identified model based on the physical knowledge and check the estimated parameters based on the statistical knowledge. Comparing the estimates of the parameters in the identified model with the width of confidence interval, we can find the existing ranges of actual values of parameters [3]. The $100(1-\gamma)$ percentage confidence interval of the estimated parameters is written as

$$b_i : \hat{b}_i \pm t(n-m-1 ; \gamma)\sqrt{S^{ii}V_e} \qquad (3.22.a)$$

$$V_e = S_e / (n-m-1) \qquad (3.22.b)$$

where V_e is the sample variance, n denotes the data length, m is the number of input variables, S^{ii} is (i,i)-th element in $(X^TX)^{-1}$ and $t(n-m-1;\gamma)$ is the 100γ percentage point of a t-distribution with (n-m-1) degrees of freedom.

3.3 Discovery of Physical Law by the Revised GMDH Algorithm [7]

We assume that a force F (gr. cm/sec^2) is applied to an object of mass m (gr.) which is placed on a perfectly smooth surface, and the displacement x and the velocity v are observed. Suppose we use four kinds

of mass (m = 3,5,7,9 gr.), and observe x and v eight times every one second. The force is altered with respect to time. The observed data for the mass m = 9 are shown in Table 3.1, where it is assumed that the measurement errors contained in the observed variables F(t), x(t) and v(t) are Gaussian white noises with zero mean and standard deviation 0.05. We will find the relationship between the input and output variables by applying the revised GMDH algorithm to these data. As the output variables, two variables x(k+1) and v(k+1) are chosen, and as the input variables eight variables

$$x_1 = m,\ x_2 = 1/m,\ x_3 = F(k),\ x_4 = 1/F(k),\ x_5 = x(k),\ x_6 = 1/x(k),$$
$$x_7 = v(k),\ x_8 = 1/v(k)$$

are chosen.

1) Numerical results obtained by the revised GMDH algorithm of generating optimal partial polynomials

Eight intermediate variables are selected in each selection layer.

Table 3.1 Observed data in a simple kinetic system (m = 9)

t (sec)	0	1	2	3	4	5	6	7
F(t) (gr. cm/sec^2)	2.90	2.08	1.03	0.01	0.95	1.94	2.99	2.00
x(t) (cm)	0.00	0.10	0.59	1.24	1.88	2.65	3.52	4.63
v(t) (cm/sec)	0.00	0.38	0.43	0.55	0.67	0.74	0.97	1.29

Observed data for the mass (m = 3,5,7,9 gr.) are used for interpolation points. The models for the output variable x(k+1) are constructed in the fourth selection layer. We show an example of the model obtained in the fourth selection layer.

$$z_1 = 0.017 + 1.081y_1 - 0.087y_8$$

$$y_1 = 0.146 + 1.098x_7 + 0.952x_5$$

$$y_8 = 0.042 + 4.259x_7 - 12.11x_7x_2 + 1.065x_2^2 \qquad (3.23)$$

The models for the output variable v(k+1) are constructed in the fifth selection layer. We show an example of the model obtained in the fifth selection layer.

$$z_1 = -0.014 + 0.589y_1 + 0.423y_3$$

$$y_1 = 0.249 + 0.842x_7 + 3.452x_7x_2 - 0.362x_2^2$$

$$y_3 = 0.112 + 0.109x_3 + 0.741x_7 + 0.109x_3x_7 + 0.078x_7^2 \qquad (3.24)$$

For the output variable x(k+1), the input variables x(k), v(k), v(k)/m and $1/m^2$ are selected. For the output variable v(k+1), the input variables v(k), $v(k)^2$, F(k), F(k)v(k), v(k)/m and $1/m^2$ are selected. It took 2 seconds for computation in each selection layer and 13 kw

for computer memory, where NEAC 2200/700† is used.

2) Numerical results obtained by the revised GMDH algorithm of generating optimal intermediate polynomials

Three parameters used in the revised GMDH algorithm are as follows: maximum order p of the intermediate polynomial in the first layer is two, maximum number L_1 of the terms in the intermediate polynomial is ten, and the number m_1 of intermediate polynomials selected in each selection layer is eight. Observed data for m = 3, 5, 7, 9 gr. are used as the interpolation points. The model for the output variable x(k+1) is constructed in the third layer as

$$x(k+1) = \underset{(\pm 0.040)}{0.035} + \underset{(\pm 0.018)^5}{0.999x_5} + \underset{(\pm 0.046)^7}{0.993x_7} + \underset{(\pm 0.082)^{2}x_3}{0.432x_2x_3} . \qquad (3.25)$$

The model for the output variable v(k+1) is constructed in the second layer as

$$v(k+1) = \underset{(\pm 0.057)}{0.002} + \underset{(\pm 0.036)^7}{0.992x_7} + \underset{(\pm 0.122)^{2}x_3}{0.994x_2x_3} . \qquad (3.26)$$

Here, the values shown in the parentheses are 95 percentage confidence interval for the estimated parameters. For the output variable x(k+1), the input variables x(k), v(k) and F(k)/m are selected. For the output variable v(k+1), the input variables v(k) and F(k)/m are

---

† The operation time of this computer is about three times longer than that of IBM 370/168.

selected. It took 15 seconds for computation in each selection layer and 16.5 kw for computer memory. Here, we discuss the relationship between the displacement x, velocity v and accelation α. The relationship among these variables are defined as

$$\dot{x} = v \ , \qquad \dot{v} = \alpha \ . \tag{3.27}$$

By using a vector-matrix expression, eq. (3.27) is described as

$$\frac{d}{dt}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \alpha \ . \tag{3.28}$$

This continuous-time system can be transformed to a discrete-time system

$$\begin{bmatrix} x(k+1) \\ v(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x(k) \\ v(k) \end{bmatrix} + \begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix} \alpha(k) \tag{3.29}$$

when we choose sampling time interval of one second. By comparing eq. (3.29) with eqs. (3.25) and (3.26) obtained by the revised GMDH of generating intermediate polynomials, we can find the relationship

$$F(k) \ / \ m \simeq \alpha(k) \ . \tag{3.30}$$

Equation (3.30) shows the Newton's second law. This shows that

the Newton's second law is discovered by the revised GMDH of generating intermediate polynomials.

On the other hand, even if we compare eq. (3.29) with eqs. (3.23) and (3.24) obtained by the revised GMDH of generating partial polynomials, we cannot find the Newton's second law. The reason for this is that, in the revised GMDH of generating optimal partial polynomials, it is difficult to identify a physically meaningful structure between the input and output variables because the partial polynomials, in which the intermediate variables are used as the input variables in each selection layer, are accumulated in the multilayered structure. The computation time of the revised GMDH of generating optimal partial polynomials is much less than that of the revised GMDH of generating optimal intermediate polynomials.

3.4 Concluding Remarks

In this Chapter, a revised GMDH algorithm of generating optimal intermediate polynomials in each selection layer is proposed where AIC is used as a criterion for model selection. In this algorithm, the intermediate polynomials show the direct relationship between the input and output variables, therefore the physically meaningful structure can be identified when the characteristics of the system are well reflected in the data.

The revised GMDH algorithm is applied to the input-output data observed in a simple kinetic system, and we tried to discover

the Newton's second law. The result obtained is compared with that obtained by the revised GMDH algorithm of using optimal partial polynomials, and the effectiveness of the revised GMDH algorithm in this Chapter for the problem of identifying physically meaningful structure is justified.

When we apply the revised GMDH algorithm to the real problem, we must investigate the structure and the parameters of the identified model under the physical and statistical knowledges, respectively, because when the input and output variables are disturbed by the noises, the structure of the identified model may become quite different from that of the real system. For the problem of including very many variables and very complex structure, the revised GMDH algorithm of using optimal partial polynomials is more suitable than the revised GMDH algorithm of using optimal intermediate polynomials. This is because it is difficult to find the physically meaningful relationship between the output variable and each input variable when very many variables are contained in the system. Furthermore, we consider that it is very difficult to identify the structure of the system accurately by using only observed data in the presence of measurement noises. That is, it is necessary to know how to use the GMDH algorithm properly depending upon the characteristics of the problem.

When the characteristics of the problem is completely unknown, the complexity of the system structure can be tested by using the revised GMDH algorithm in this Chapter. If the model with very complex structure is identified, the system should be identified again by using

the revised GMDH algorithm of using optimal partial polynomials.

The advantage of the revised GMDH in this Chapter compared with the multiple stepwise regression analysis with variable selection is now clear. In the multiple regression analysis the amount of computation is increasing very rapidly with the increase of the number and the order of the input variables. On the other hand, in the revised GMDH the increase of the computational burden with respect to the increase of the number of input variables is quite modest.

The application of the revised GMDH in this Chapter to river pollution problem will be discussed in Chapter 5.

REFERENCES

[1] Akaike, H.: A new look at the statistical model identification, IEEE Trans. Automatic Control, Vol. AC-19, No. 6, 716-723 (1974)

[2] Allen, D.M.: The relationship between variable selection and data augmentation and a method for prediction, Technometrics, Vol. 16, No. 1, 125-127 (1974)

[3] Draper, N.R. and H. Smith: Applied Regression Analysis, Wiley, New York (1966)

[4] Ihara, J.: Improved GMDH — A case of dynamical world population models, (in Japanese) Systems and Control, Vol. 19, No. 4, 201-210 (1976)

[5] Ikeda, S., M. Ochiai and Y. Sawaragi: Sequential GMDH algorithm and its application to river flow prediction, IEEE Trans. Syst., Man, Cybern., Vol. SMC-6, No. 7, 473-479 (1976)

[6] Ivakhnenko, A.G., et al.: Discovery of physical laws by GMDH method with the absence-of-bias criterion, Soviet Automatic Control, Vol. 6, No. 6, 32-45 (1973)

[7] Kondo, T. and H. Tamura: Revised GMDH algorithm of self-selecting optimal intermediate polynomials using AIC, (in Japanese) Trans. Soc. Instr. Control Engineers. (forthcoming)

[8] Tamura, H. and T. Kondo: Revised GMDH algorithm using prediction sum of squares (PSS) as a criterion for model selection, (in Japanese) Trans. Soc. Instr. Control Engineers, Vol. 14, No. 5, 519-524 (1978)

CHAPTER 4 APPLICATIONS TO AIR POLLUTION PROBLEMS

4.1 Introduction

In this Chapter, the revised GMDH algorithm of generating optimal partial polynomials, which has been developed in Chapter 2, is applied to two air pollution problems; one is a steady state spatial pattern identification problem and the other is an unsteady state short-term prediction problem. In 4.2, large-spatial pattern identification of air pollution by a combined model of source-receptor matrix and the revised GMDH is discussed [10]. A source-receptor matrix [6], which represents a linear relationship between the multiple air pollution sources and the air pollution concentrations at the multiple monitoring stations (receptors), is estimated by a regression analysis of rough data. This source-receptor matrix is used as a rough model of first-order approximation. Then, the difference between the output of the real system (measured data at the monitoring station) and the output of the rough model is identified by the revised GMDH algorithm using optimal partial polynomials. By using synthetic data obtained by the computer simulation of air pollution diffusion, the predicted result

obtained from the combined model developed in this Chapter is compared with the results obtained from the source-receptor matrix model only, and also with the results obtained from the combined model of source-receptor matrix and the basic GMDH.

In 4.3, nonlinear modeling for short-term prediction of air pollution concentration by the revised GMDH is discussed [11]. By using the time series data of SO_2 concentration, the wind velocity and the wind direction in Tokushima, Japan, we intend to find a suitable model for predicting SO_2 concentration at a few hours in advance. Firstly, a suitable data length for modeling air pollution in Tokushima is investigated. Secondly, three different prediction models obtained by the revised GMDH are compared to find suitable structure and the suitable input variables in the model. The predicted results obtained by the revised GMDH model are compared with the results obtained by a linear regression model, a linear autoregressive model and a basic GMDH model. It is shown that the revised GMDH model developed in this Chapter gives better performance for short-term prediction of air pollution concentration compared with the linear models and the basic GMDH model, and it is also shown that the revised GMDH model obtained is much simpler than the basic GMDH model.

4.2 Large-Spatial Pattern Identification of Air Pollution [10]

4.2.1 Physical and statistical models of air pollution

The air pollution models used for predicting air pollution concentration have been proposed as physical models [3,9] or statistical models [1,5,8]. Some physical models are based on three-dimensional partial differential equations which govern the diffusion phenomena of the pollutant. But, in general, it is not easy to solve these diffusion equations for practical situations. So simplified physical models, such as plume model, puff model and box model, have been proposed and applied to long-term prediction or short-term prediction of air pollution concentration. But these physical models have limitations in practical applications such that some unrealistic assumptions and simplifications are used for obtaining the models. On the other hand, nonphysical statistical models [8] are constructed depending only on the statistical analysis of the data measured at the monitoring stations, and very easily applicable to practical prediction problems. Furthermore, complex factors, which cannot be expressed theoretically, can be taken into account in nonphysical models through measured data. But, in these models, the physical processes are treated as a black box, so the physical meaning of these models is not clear.

Here, a combined model of a source-receptor matrix and a revised GMDH is developed. By using any physical prior knowledge of the system, the source-receptor matrix [6,7], which represents a linear relationship between the multiple air pollution sources and the air pollution concentration at the multiple monitoring stations, can be estimated as a model which has a physical meaning [6]. After eliminating the linear part of the system by using the source-receptor

matrix, the completely unknown nonlinear part of the system is identified as a nonphysical model by using the revised GMDH proposed in Chapter 2.

4.2.2 Source-receptor matrix [6,7]

It is assumed that an air pollution model used for steady state (monthly or yearly average) identification of air pollution concentration can be described by the following equation for single air pollution source.

$$c = f\ q \tag{4.1}$$

where c is the air pollution concentration at the monitoring station, q is the emission intensity of pollution source, and f is a coefficient which is determined by the various factors concerned with the pollution source and the diffusion field. In this paper, f is considered to be an explicit function of relative coordinates between the pollution source and the monitoring station. The other factors, such as the topography and the atmospheric stability, are taken into account implicitly when f is determined by using the measured data.

For multiple sources, the air pollution concentration of each monitoring station is estimated by the following equation

$$c_i = \sum_{j=1}^{N} f_{ij}\, q_j , \qquad i=1,2,\cdots,M \tag{4.2}$$

where

$$f_{ij} = f_j(X_{ij}, Y_{ij}), \qquad i=1,2,\cdots,M;\ j=1,2,\cdots,N$$

$$X_{ij} = X_i^r - X_j^s, \qquad i=1,2,\cdots,M;\ j=1,2,\cdots,N$$

$$Y_{ij} = Y_i^r - Y_j^s, \qquad i=1,2,\cdots,M;\ j=1,2,\cdots,N$$

c_i: air pollution concentration at the i-th monitoring station

q_j: emission intensity of the j-th pollution source

(X_i^r, Y_i^r): coordinates of the i-th monitoring station

(X_j^s, Y_j^s): coordinates of the j-th pollution source

M : number of monitoring stations

N : number of pollution sources .

By using vector-matrix representation, eq. (4.2) can be written as

$$\underline{c} = F\, \underline{q} \tag{4.3}$$

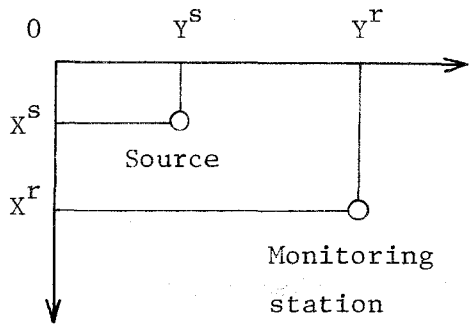
where

$$\underline{c}^T = (c_1, c_2, \cdots, c_M)$$

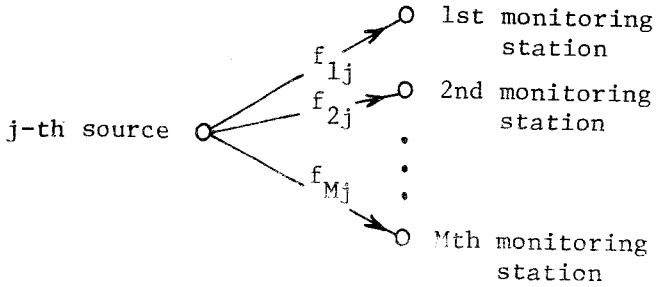
$$\underline{q}^T = (q_1, q_2, \cdots, q_N)$$

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \vdots & \vdots & & \vdots \\ f_{M1} & f_{M2} & \cdots & f_{MN} \end{pmatrix}$$

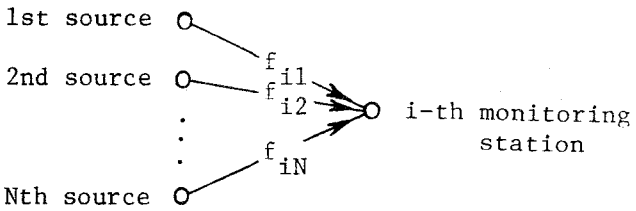
Here, F is called as source-receptor matrix. Figure 4.1 shows the coordinate system and the elements of source-receptor matrix.



(a) Coordinate system



(b) Contribution of one source to many receptors



(c) Contribution of many sources to one receptor

Fig. 4.1 Coordinate system and source-receptor martix

When regional environmental planning and environmental impact assessment are to be performed, it is necessary to estimate the spatial distribution pattern of each air pollution source. For this purpose, the source-receptor matrix, which represents the relationship between each air pollution source and the air pollution concentration at each monitoring station, would be very useful [7].

4.2.3 Estimation of source-receptor matrix by a regression analysis

For estimating each element of the source-receptor matrix, it has been proposed [6] to use physical model such as plume model, but the physical model has limitations in practical applications such that complexity of topography, down wash, and down draught cannot be easily taken into account in the model theoretically. Here, instead of using physical model, each element of the source-receptor matrix is estimated by a regression analysis of the spatially distributed data obtained from e.g. wind tunnel experiments† for a single source. Each element f_{ij} of the source-receptor matrix F in eq. (4.3) is assumed to be described as

$$f_{ij} = a_{0j} + a_{1j}X_{ij} + a_{2j}Y_{ij} + a_{3j}X_{ij}^2 + a_{4j}Y_{ij}^2 + a_{5j}X_{ij}Y_{ij} + a_{6j}e^{-a_j\sqrt{X_{ij}^2 + Y_{ij}^2}} \tag{4.4}$$

---

† If it is hard to execute the wind tunnel experiments in practice, inaccurate data obtained from a physical model could be used for our purpose.

where

$$X_{ij} = X_i^r - X_j^s, \qquad i=1,2,\cdots,M;\ j=1,2,\cdots,N$$

$$Y_{ij} = Y_i^r - Y_j^s, \qquad i=1,2,\cdots,M;\ j=1,2,\cdots,N$$

a_j: a constant

(X_i^r, Y_i^r): coordinates of the i-th monitoring station

(X_j^s, Y_j^s): coordinates of the j-th pollution source

M : number of monitoring stations

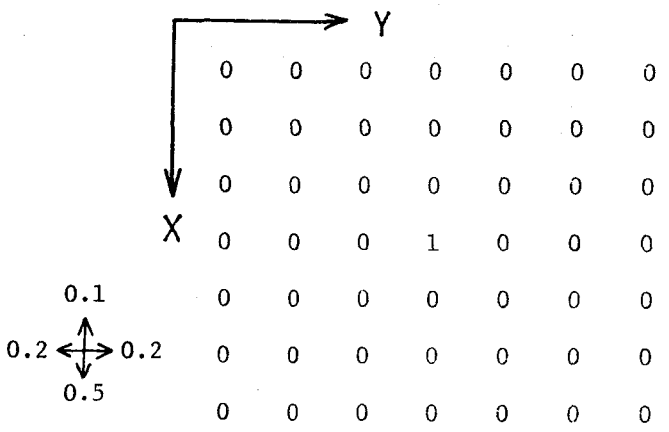
N : number of pollution sources.

For each pollution source , we need to estimate the coefficients a_{0j}, $a_{1j},\cdots,a_{6j}$ in eq. (4.4) by the repetition of regression analysis. Each element f_{ij} is then obtained from eq. (4.4).

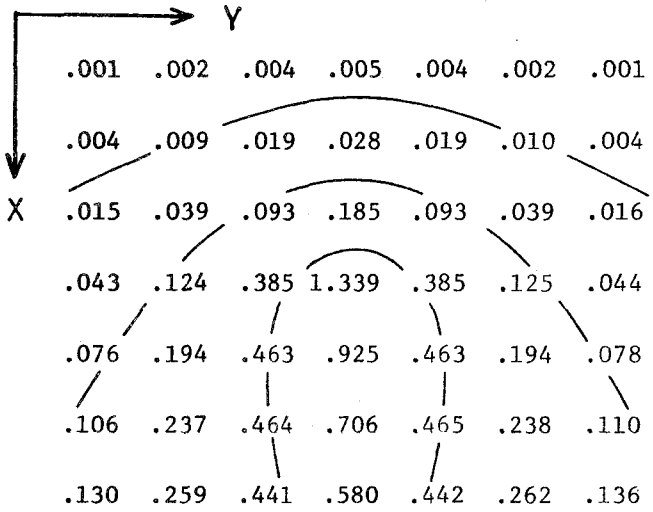
In this paper, instead of using the data obtained from wind tunnel experiments, synthetic data, which are obtained by the computer simulation of air pollution diffusion, are used to estimate each element of the source-receptor matrix. Figure 4.2 shows input and output data of the simulator for a single source. If there exists an air pollution source of intensity one at the coordinates (4,4) as shown in Fig. 4.2 (a), and the diffusion rate of the pollutant are 0.2, 0.2, 0.5 and 0.1, a steady state of the spatially distributed air pollution concentration shown in Fig. 4.2 (b) is obtained. By the multiple regression analysis of the data shown in Fig. 4.2 (b), the coefficients in eq. (4.4) are obtained as

$$f_{ij} = 0.141 + 0.067X_{ij} - 0.015Y_{ij}^2 + 1.029e^{-\sqrt{X_{ij}^2 + Y_{ij}^2}} \qquad (4.5)$$

where the stepwise forward regression analysis [2,4] is used for selecting dominant variables. In order to simplify the procedure, we assume that eq. (4.5) is applicable to all the air pollution sources.



(a) Input data



(b) Output of large-spatial pattern

Fig. 4.2 Input and output of the simulator for single source

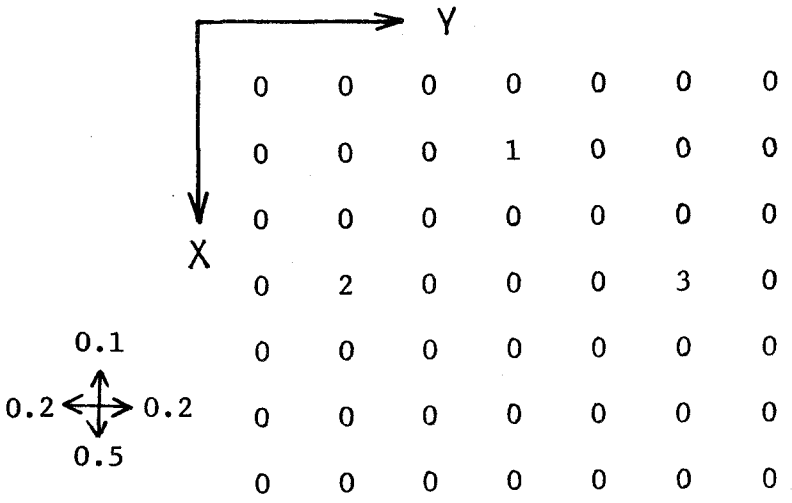
4.2.4 Identification of large-spatial pattern of air pollution by the combined model

In this section, a spatially distributed pattern of air pollutant, which is emitted from the multiple air pollution sources, is identified by the combined approach of source-receptor matrix and the revised GMDH. Suppose there exists three air pollution sources at the coordinates (2,4), (4,2) and (4,6) as shown in Fig. 4.3 (a), a steady state of the spatially distributed air pollution concentration shown in Fig. 4.3 (b) is obtained as the result of the computer simulation of air pollution diffusion. The data underlined in Fig. 4.3 (b) are assumed to have been measured at the monitoring stations. This large-spatial pattern is identified by the combined approach in this section. Firstly, by applying eq. (4.5) to each air pollution source, a source-receptor matrix is determined as

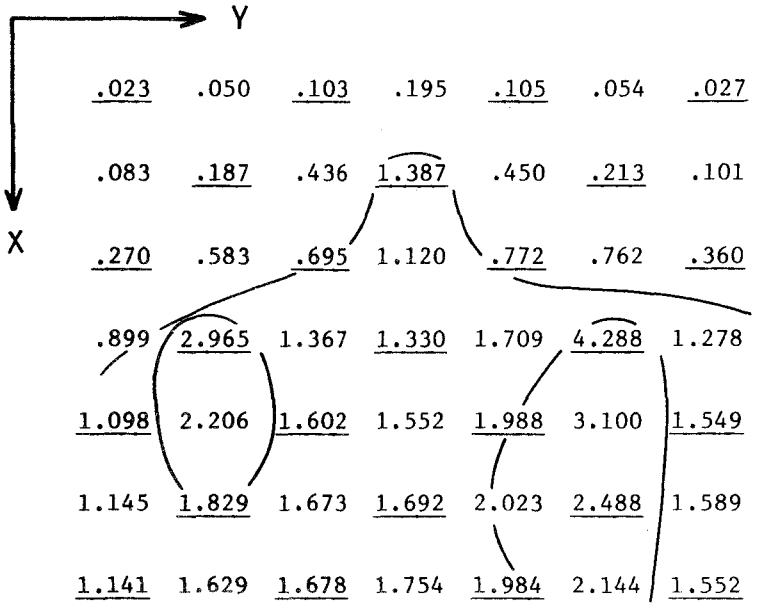
$$F = \begin{pmatrix} -0.0185 & -0.0305 & -0.4352 \\ 0.0559 & 0.1026 & -0.3668 \\ \vdots & \vdots & \vdots \\ 0.3411 & -0.0350 & 0.3696 \end{pmatrix} \tag{4.6}$$

The air pollution model using this source-receptor matrix is described as

$$\underline{c} = F\,\underline{q} \tag{4.7}$$



(a) Input data



(b) Output of large-spatial pattern

Fig. 4.3 Input and output of the simulator for multiple sources

where

$$\underline{c}^T = (c_1, c_2, \cdots, c_{49}\dagger)$$

$$q^T = (q_1, q_2, q_3) .$$

This air pollution model is used as a rough model of first-order approximation, which plays a role of eliminating so called trends from the measured data at the monitoring stations. The large-spatial pattern predicted by the rough model is shown in Fig. 4.4.

After eliminating the linear part of the system by using the rough model, the residual pattern, which is the completely unknown nonlinear part of the system, is described as

$$\Delta c_i = \frac{1}{3} \sum_{j=1}^{3} g_j(X_{ij}, Y_{ij}), \qquad i=1,2,\cdots,49 \tag{4.8}$$

where

$$X_{ij} = X_i^r - X_j^s, \qquad i=1,2,\cdots,49;\ j=1,2,3$$

$$Y_{ij} = Y_i^r - Y_j^s, \qquad i=1,2,\cdots,49;\ j=1,2,3$$

Δc_i: residual data at the i-th point

(X_i^r, Y_i^r): coordinates of the i-th point

(X_j^s, Y_j^s): coordinates of the j-th pollution source .

---

† Here, the number of points includes not only the number of monitoring stations but also the number of prediction points.

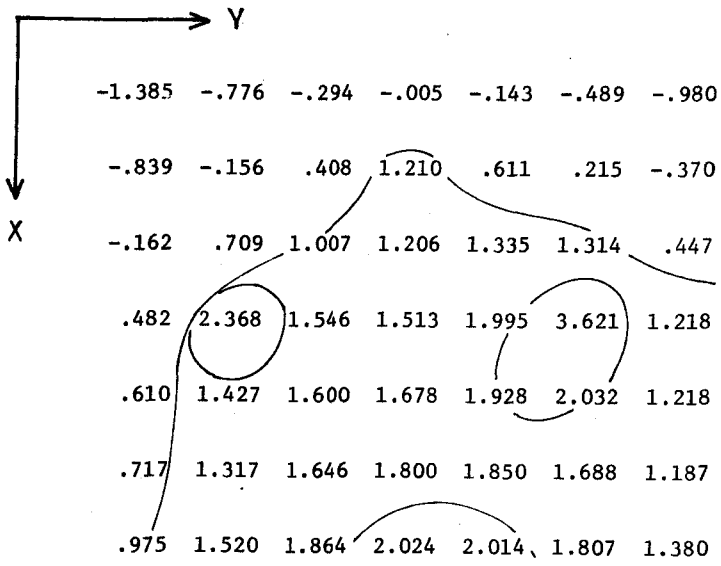


Fig. 4.4 Predicted values of large-spatial pattern using source-receptor matrix

Here, the functions $g_j(X_{ij},Y_{ij})$ (j=1,2,3) whose structures are completely unknown, are assumed to be described by polynomials of a certain order with respect to X_{ij} and Y_{ij}. Equation (4.8) can be obtained as an average of the three models identified by the revised GMDH. The residual data shown in Fig. 4.5, which can be calculated by using the measured data at the monitoring stations, are used as the input data of the revised GMDH. The large-spatial pattern predicted by the rough model is corrected by using eq. (4.8).

The block diagram of the prediction system using the combined model of source-receptor matrix and the revised GMDH is shown in Fig. 4.6.

For comparing the revised GMDH with the basic GMDH, the predicted results obtained by a combined model of source-receptor matrix and the basic GMDH are also shown where five variables

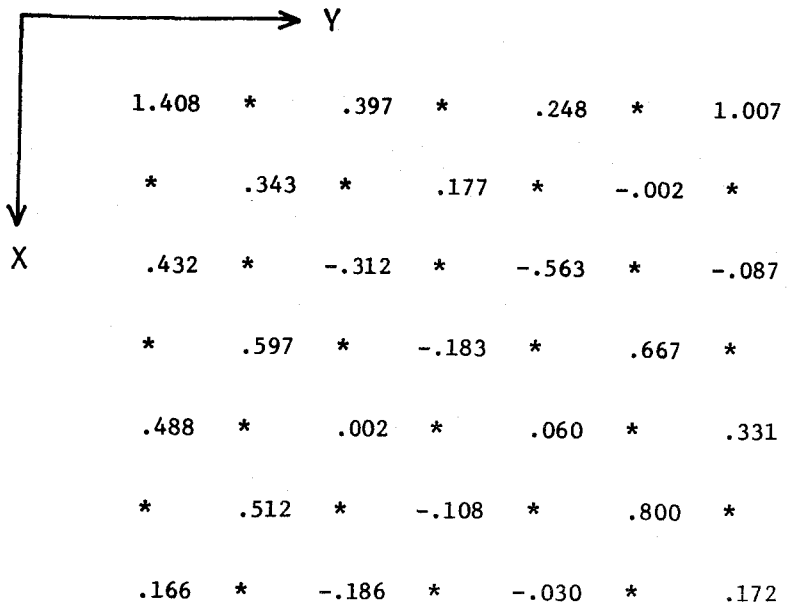


Fig. 4.5 Deviation of the predicted values from the measured data

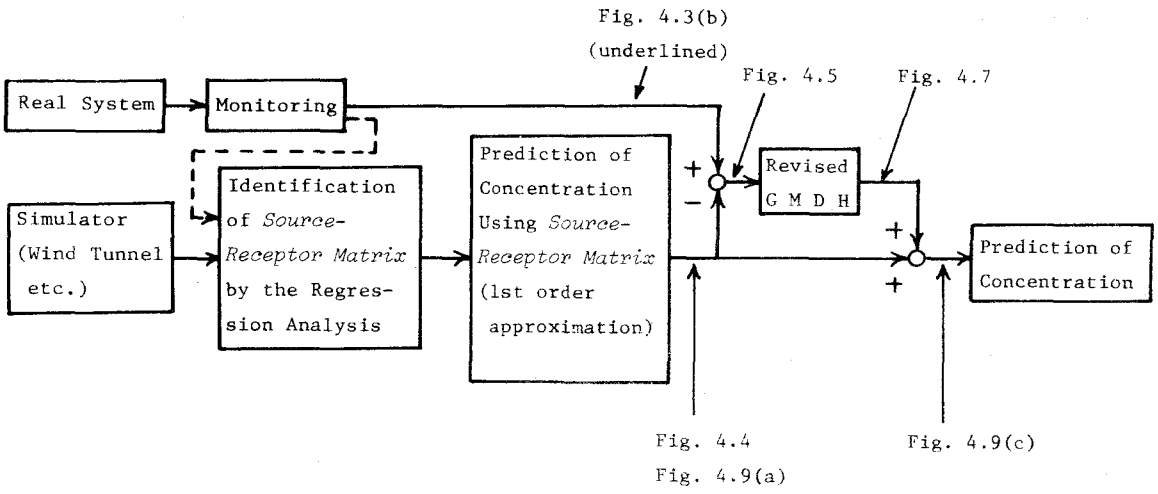


Fig. 4.6 Block diagram of the prediction system using source-receptor matrix and revised GMDH

$$x_1 = X_i^r - X_j^s,\ x_2 = Y_i^r - Y_j^s,\ x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1x_2$$

are used as input variables, five intermediate variables are selected in each layer and twenty-five points are used as the interpolation points. In the basic GMDH, the interpolation points are divided into the training data and the checking data in proportion of 17 : 8.

1) Prediction model identified by GMDH

For the air pollution sources at the coordinates (2,4), (4,2) and (4,6), the prediction models $g_1^b(X,Y)$, $g_2^b(X,Y)$ and $g_3^b(X,Y)$ are identified by the basic GMDH. Here, only $g_3^b(X,Y)$ is described as

$$g_3^b(X,Y) = z_1 = 0.095 + 0.610y_3 + 1.403y_4 + 21.567y_3y_4$$

$$- 14.867y_3^2 - 8.727y_4^2$$

$$y_3 = 0.061 + 0.132x_1 - 0.024x_4 - 0.006x_1x_4$$

$$+ 0.045x_1^2 + 0.002x_4^2$$

$$y_4 = 0.035 + 0.162x_1 + 0.191x_2 + 0.038x_1x_2$$

$$+ 0.051x_1^2 + 0.056x_2^2 .$$

The mean square error for

checking data: 0.315

training data: 0.042 .

The prediction models $g_1^r(X,Y)$, $g_2^r(X,Y)$ and $g_3^r(X,Y)$ are also identified by the revised GMDH. Here, only $g_3^r(X,Y)$ is described as follows :

$$g_3^{(1)}(X,Y) = z_2 = -0.097 + 0.575y_1 + 0.553y_5 + 0.816y_1y_5$$

$$y_5 = 0.253 + 0.207x_1 - 0.037x_1x_3$$

$$y_1 = 0.179 + 0.220x_2 - 0.006x_2x_5 + 0.063x_2^2$$

PSS/25 = 0.073, RSS/25 = 0.063

$$g_3^{(2)}(X,Y) = z_1 = -0.107 + 0.589y_3 + 0.552y_5 + 0.946y_3y_5$$

$$y_5 = 0.253 + 0.207x_1 - 0.037x_1x_3$$

$$y_3 = 0.179 + 0.220x_2 + 0.023x_1x_2 + 0.063x_2^2$$

PSS/25 = 0.074, RSS/25 = 0.066

$$g_3^{(3)}(X,Y) = v_3 = -0.103 + 1.005z_4 + 0.583z_5 - 0.290z_4^2$$

$$z_4 = -0.024 + 0.745y_5 + 0.677y_2^2$$

$$z_5 = 0.000 + 1.000y_1$$

$$y_2 = 0.029 - 0.006x_1x_4 + 0.027x_1^2 + 0.001x_4^2$$

$$y_1 = 0.179 + 0.220x_2 - 0.006x_2x_5 + 0.063x_2^2$$

$$y_5 = 0.253 + 0.207x_1 - 0.037x_1x_3$$

PSS/25 = 0.081, RSS/25 = 0.072

$$g_3^{(4)}(X,Y) = v_4 = - 0.050 + 0.679z_3 + 0.519z_5$$

$$z_3 = - 0.037 + 0.795y_5 + 0.737y_4^2$$

$$z_5 = 0.000 + 1.000y_1$$

$$y_5 = 0.253 + 0.207x_1 - 0.037x_1x_3$$

$$y_4 = 0.137 + 0.001x_4x_5 + 0.001x_4^2$$

$$y_1 = 0.179 + 0.220x_2 - 0.006x_2x_5 + 0.063x_2^2$$

PSS/25 = 0.087, RSS/25 = 0.076

Here, $g_3^{(i)}$ (i=1,2,3,4) is the i-th complete polynomial which is remained in the final layer, and the resulting prediction model $g_3^r(X,Y)$ is obtained as a weighted average of four polynomials as

$$g_3^r(X,Y) = 0.268g_3^{(1)}(X,Y) + 0.264g_3^{(2)}(X,Y) + 0.242g_3^{(3)}(X,Y)$$

$$+ 0.225g_3^{(4)}(X,Y) \ .$$

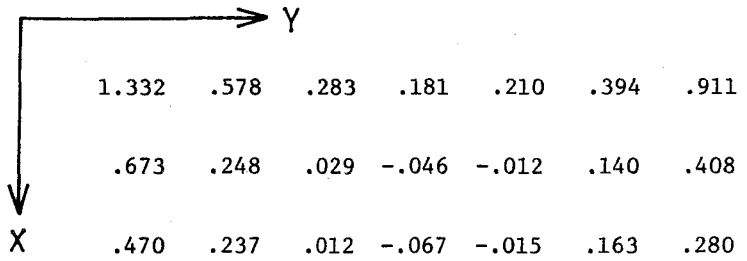
The predicted result for the residuals, which is an output of the revised GMDH, is shown in Fig. 4.7.

2) Accuracy at the interpolation points

Accuracy at the interpolation points, which is obtained from the prediction model for the air pollution source at the coordinates of (4,6), is shown in Fig. 4.8. For the basic GMDH, the changes of mean square errors for the training data and the checking data are shown in Fig. 4.8 (a). The mean square error for the training data is very small but that for the checking data is very large. This result shows that the prediction model identified by the basic GMDH is not a satisfactory model of the system. For the revised GMDH, the changes of PSS and RSS are shown in Fig. 4.8 (b). PSS and RSS are very small and coincide well at the 4-th layer. These results justify that the prediction model identified by the revised GMDH is much better than the model identified by the basic GMDH.

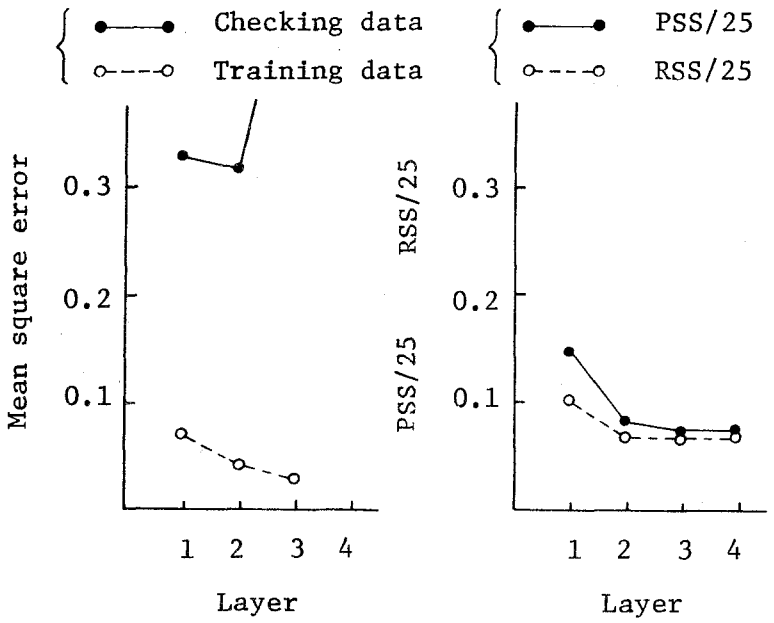
3) Accuracy at the prediction points

The large-spatial pattern of air pollution concentration predicted by the following three models are shown in Fig. 4.9.



1.332	.578	.283	.181	.210	.394	.911
.673	.248	.029	-.046	-.012	.140	.408
.470	.237	.012	-.067	-.015	.163	.280
.481	.344	.069	-.028	.051	.290	.327
.551	.456	.129	.021	.122	.420	.430
.461	.395	.063	-.033	.069	.388	.393
.098	.093	-.147	-.185	-.131	.115	.106

Fig. 4.7 Predicted values of the deviation by the revised GMDH



(a) Basic GMDH (b) Revised GMDH

Fig. 4.8 Mean square error, PSS and RSS

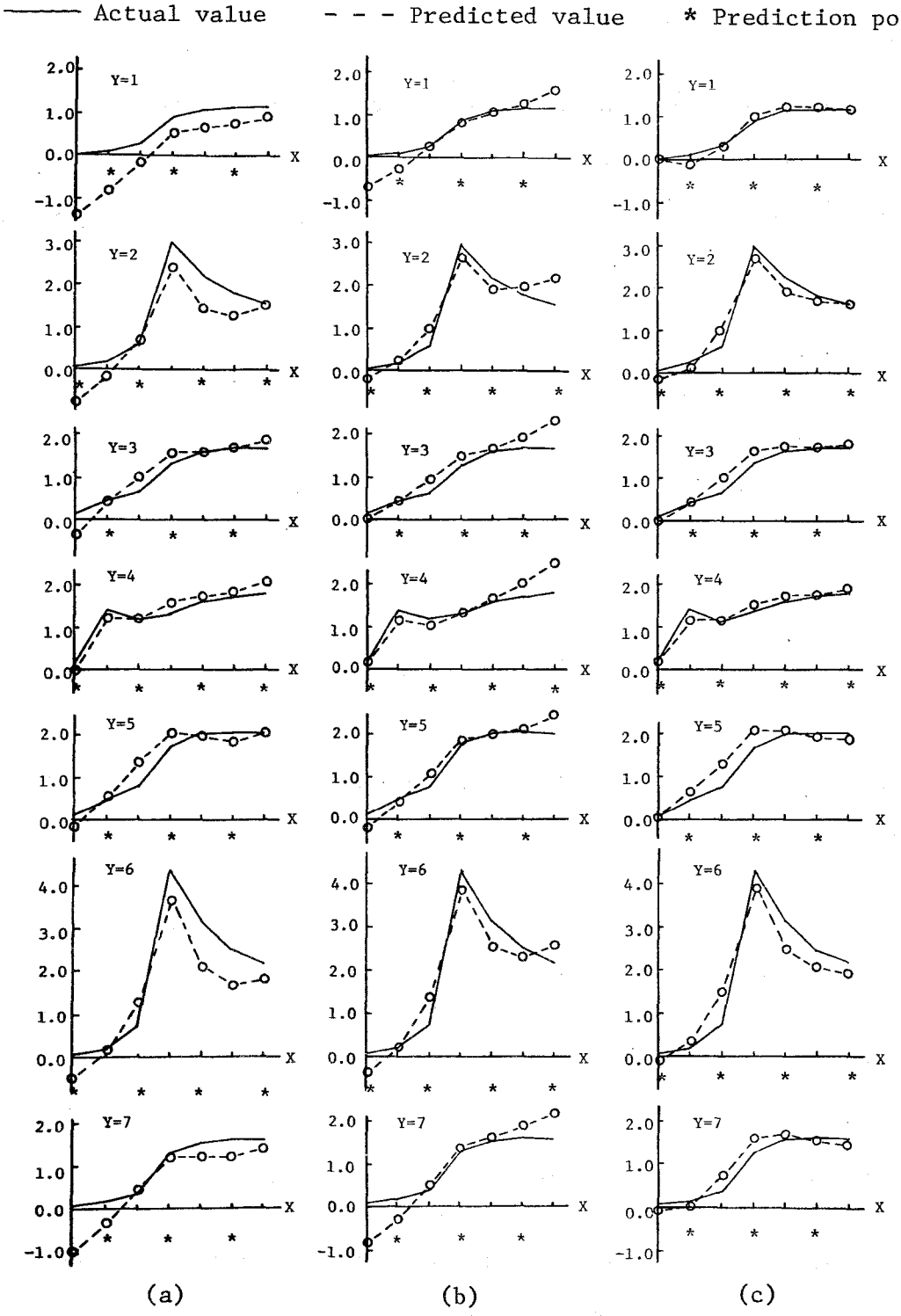


Fig. 4.9 Predicted values for three procedures

(a) Predicted values using source-receptor matrix

(b) Predicted values using source-receptor matrix and basic GMDH

(c) Predicted values using source-receptor matrix and revised GMDH

(a) Source-receptor matrix model

(b) Combined model of source-receptor matrix and the basic GMDH

(c) Combined model of source-receptor matrix and the revised GMDH

For comparing the accuracy of prediction by using these three models, a performance index defined as

$$J_k = \left(\sum_{\alpha=1}^{24} |c_\alpha - \hat{c}_{k\alpha}| \Big/ \sum_{\alpha=1}^{24} c_\alpha \right) \times 100 , \quad k=1,2,3 \tag{4.9}$$

is introduced, where $\hat{c}_{k\alpha}$ is the predicted value using the k-th model.

$$J_1 = 30.8\ \%, \quad J_2 = 23.1\ \%, \quad J_3 = 16.2\ \%$$

are obtained. The third model developed in this section gives the best performance among these three models.

4.3 Nonlinear Modeling for Short-Term Prediction of Air Pollution Concentration [11]

4.3.1 Linear and nonlinear modeling for short-term prediction

The mathematical models used for predicting air pollution concentration can be roughly classified into two groups; physical models and statistical models [3,8,9,10]. Generally, in physical models based

on the diffusion phenomena of the pollutants, the physical interpretation of the model can be easily obtained, but there exist many complex factors which cannot be incorporated in the model theoretically. On the other hand, in statistical models, the physical interpretation of the model is not clear, since the physical process is regarded as a black-box, however, the complex factors can be implicitly taken into account in the model through the measured data. As the statistical models for short-term prediction of air pollution concentration, linear models such as multiple regression models and autoregressive models have been often used [8]. However, since the phenomena in air pollution are considerably influenced by the complex weather conditions and photo-chemical reactions, linear statistical models are not sufficient to describe the phenomena.

Here, nonlinear statistical models for short-term prediction of air pollution are identified by a revised GMDH algorithm proposed in Chapter 2.

4.3.2 Nonlinear models for short-term prediction of air pollution

Here, the nonlinear statistical models for short-term prediction of air pollution concentration are constructed. We use the time series data of SO_2 concentration, wind direction and wind velocity obtained at the monitoring station in Tokushima, Japan. Suppose the time series data of these three variables which are measured every one hour have been accumulated for N days. Table 4.1 shows the structure of the data.

Since the measured data of SO_2 concentration contain a periodic phenomenon of 24 hours, the data are pre-processed in order to remove this periodic factor. Furthermore, the measured data of the wind direction and the wind velocity are transformed to the east-west component and the south-north component of the wind velocity.

The output variable of the prediction model is the SO_2 concentration at one, two and three hours in advance. The input variables of the prediction model are the time lagged values of the SO_2 concentration, the east-west component and the south-north component of the wind velocity. The number of time lagged values τ is chosen by evaluating the auto-correlation function of SO_2. In this section, we consider the following three different models to be identified by the revised GMDH.

Table 4.1 Structure of the data

Time / Day	1	2	. . .	24	
1	$C_{1,1}$	$C_{1,2}$	. . .	$C_{1,24}$	The data used for modeling
2	$C_{2,1}$	$C_{2,2}$	. . .	$C_{2,24}$	
.	.	.		.	
.	.	.		.	
.	.	.		.	
N	$C_{N,1}$	$C_{N,2}$	. . .	$C_{N,24}$	
N+1	$C_{N+1,1}$	$C_{N+1,2}$	. . .	$C_{N+1,24}$	The data to be predicted

1) Prediction model I

In this model we use only one variable, SO_2 concentration, as an input variable. Firstly, by using a revised GMDH algorithm, we identify the following model

$$\hat{x}(t+1) = f(x(t),x(t-1),\cdots,x(t-\tau)) \tag{4.10}$$

where f is a high-order polynomial, x(t) is the SO_2 concentration at time t. By using eq. (4.10) the value of the SO_2 concentration at one hour in advance is predicted. Then, the value at two hours in advance is predicted by using the same model as eq. (4.10), that is

$$\hat{x}(t+2) = f(\hat{x}(t+1),x(t),\cdots,x(t+1-\tau)) \tag{4.11}$$

where the predicted value at one hour in advance is used instead of the actual value. In the same way, the value at three hours in advance is predicted by using

$$\hat{x}(t+3) = f(\hat{x}(t+2),\hat{x}(t+1),x(t),\cdots,x(t+2-\tau)) \tag{4.12}$$

where $\hat{x}(t+1)$ and $\hat{x}(t+2)$ are the predicted values at one and two hours in advance, respectively.

2) Prediction model II

In this model also we use only one variable, SO_2 concentration, as an input variable. By using the revised GMDH algorithm, we identify the following three models

$$\hat{x}(t+1) = f_1(x(t),x(t-1),\cdots,x(t-\tau)) \tag{4.13}$$

$$\hat{x}(t+2) = f_2(x(t),x(t-1),\cdots,x(t-\tau)) \tag{4.14}$$

$$\hat{x}(t+3) = f_3(x(t),x(t-1),\cdots,x(t-\tau)) \tag{4.15}$$

The values of one, two and three hours in advance are predicted by using these models independently.

3) Prediction model III

In this model we use three variables, SO_2 concentration, the east-west component and the south-north component of the wind velocity, as input variables. By using the revised GMDH algorithm, we identify the following three models

$$\begin{aligned}\hat{x}(t+1) = g_1(&x(t),x(t-1),\cdots,x(t-\tau),\\ &v_1(t),v_1(t-1),\cdots,v_1(t-\tau),\\ &v_2(t),v_2(t-1),\cdots,v_2(t-\tau))\end{aligned} \tag{4.16}$$

$$\hat{x}(t+2) = g_2(x(t),x(t-1),\cdots,x(t-\tau),$$

$$v_1(t),v_1(t-1),\cdots,v_1(t-\tau),$$

$$v_2(t),v_2(t-1),\cdots,v_2(t-\tau)) \qquad (4.17)$$

$$\hat{x}(t+3) = g_3(x(t),x(t-1),\cdots,x(t-\tau),$$

$$v_1(t),v_1(t-1),\cdots,v_1(t-\tau),$$

$$v_2(t),v_2(t-1),\cdots,v_2(t-\tau)) \qquad (4.18)$$

where $v_1(t)$ is the east-west component of the wind velocity at time t and $v_2(t)$ is the south-north component of the wind velocity at time t. The values of one, two and three hours in advance are predicted by using these models independently.

4.3.3 Short-term prediction by the revised GMDH

The nonlinear statistical models for short-term prediction of air pollution levels are identified by the revised GMDH algorithm, and the SO_2 concentration at a few hours in advance are predicted by the identified models. The prediction accuracy obtained by the revised GMDH model is compared with those obtained by the linear models and the basic GMDH model. The prediction results of the linear models are

quoted from [8], and we applied the revised GMDH to the same measured data in Tokushima as in [8]. The time series data of air pollution were measured every one hour during the period from May to June, 1975 in Tokushima, and we use these data. The SO_2 concentration during 15 days from June 1 to June 15 are predicted, where the modeling is repeated for each day.

A. The prediction results by the revised GMDH

1) Comparison for various sample sizes used for modeling

As the sample size (N days) for modeling, we consider the following three cases.

Case 1: 5 days data (N = 5)

Case 2: 10 days data (N = 10)

Case 3: 31 days data (N = 31)

In other words the measured data during the past N days (N=5,10,31) are used for modeling, and the SO_2 levels at (N+1)th day are predicted by the identified model. Figure 4.10 shows the comparison of the prediction errors of SO_2 from the actual data at three hours in advance, where the prediction error for i-th day is evaluated under the following performance index,

$$\Delta J_i = \sum_{t=1}^{24} \{ x_i(t) - \hat{x}_i(t/t-m) \}^2 / 24 \ . \qquad (4.19)$$

The predicted values $\hat{x}_i(t/t-m)$ are computed by using the prediction

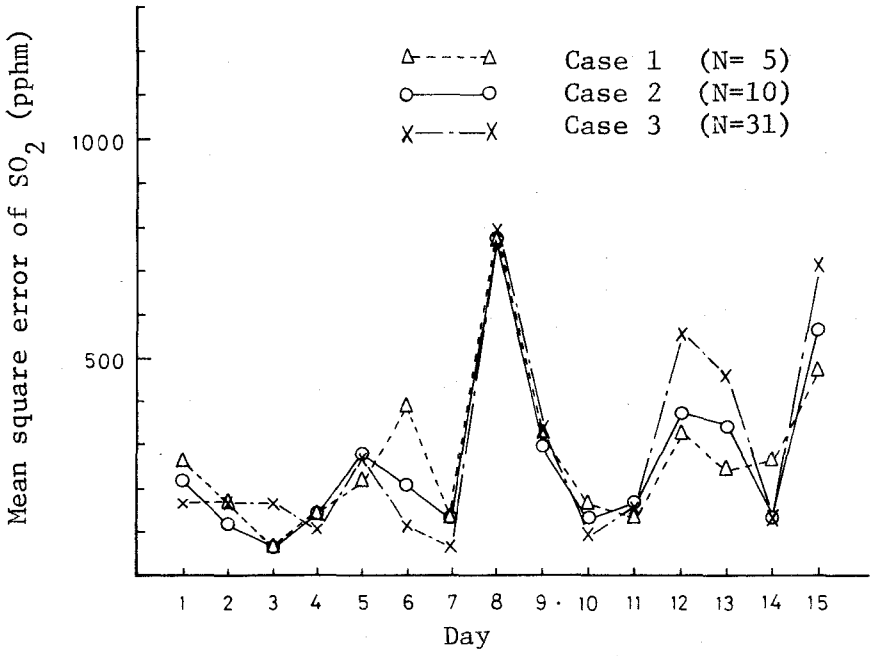


Fig. 4.10 The prediction error at three hours in advance for various sample sizes

model I (eqs. (4.10)~(4.12)). In Fig. 4.10, the prediction accuracy of Case 2 shows the same pattern of variation as that of Case 3 in most days, but that of Case 1 does not show the same pattern of variation as that of Case 2 or Case 3 in 6-th and 14-th days. Furthermore, the average prediction accuracy of Case 2 is better than that of Case 3. From these prediction results, we find that the data of 5 days are insufficient for short-term predictions in Tokushima and the data of 31 days are too many (probably because of the time-varying nature of the system). Therefore, we consider that the suitable length of data used for short-term predictions in Tokushima is about 10 days.

2) Comparison for various prediction models

Figure 4.11 shows the prediction error of SO_2 from the actual data

at three hours in advance obtained by prediction models I, II and III. Here, the prediction error for i-th day is evaluated by eq. (4.19) and the data of 10 days are used for modeling. In Fig. 4.11, the prediction error of the model II is smaller than that of the model I in most days, and the prediction error of the model III is smaller than those of other models only in a few days. From these prediction results for the SO_2 data in Tokushima, it seems that the prediction model II gives better performance than two other prediction models, and furthermore, we can not expect the improvement of the prediction accuracy by using the east-west component and the south-north component of the wind velocity as input variables. Figures 4.12 and 4.13 show the time series of the predicted values of SO_2 at one and three hours in advance, respectively, obtained by the prediction model II, and the predicted values are compared with the time series of the actual values. Table 4.2 shows

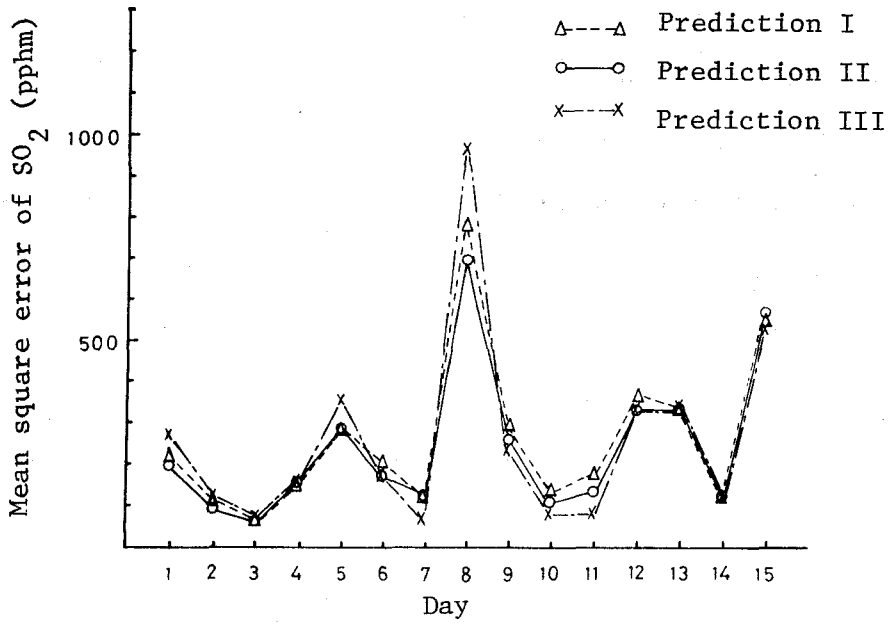


Fig. 4.11 The prediction error at three hours in advance for various prediction models

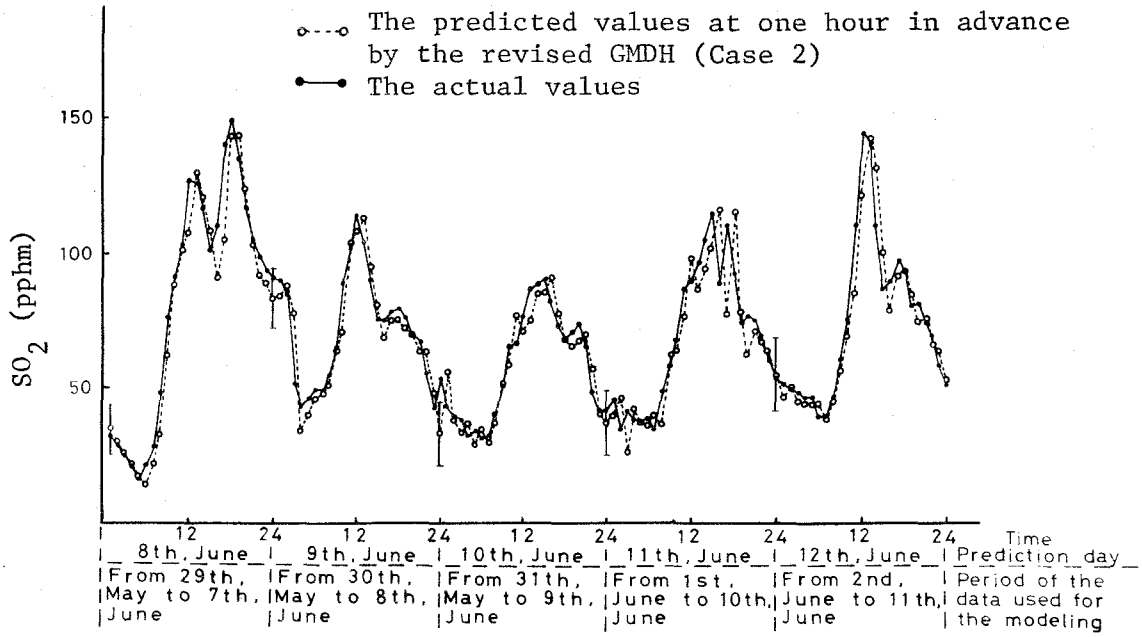


Fig. 4.12 The predicted values at one hour in advance by the revised GMDH, the confidence intervals and the actual values

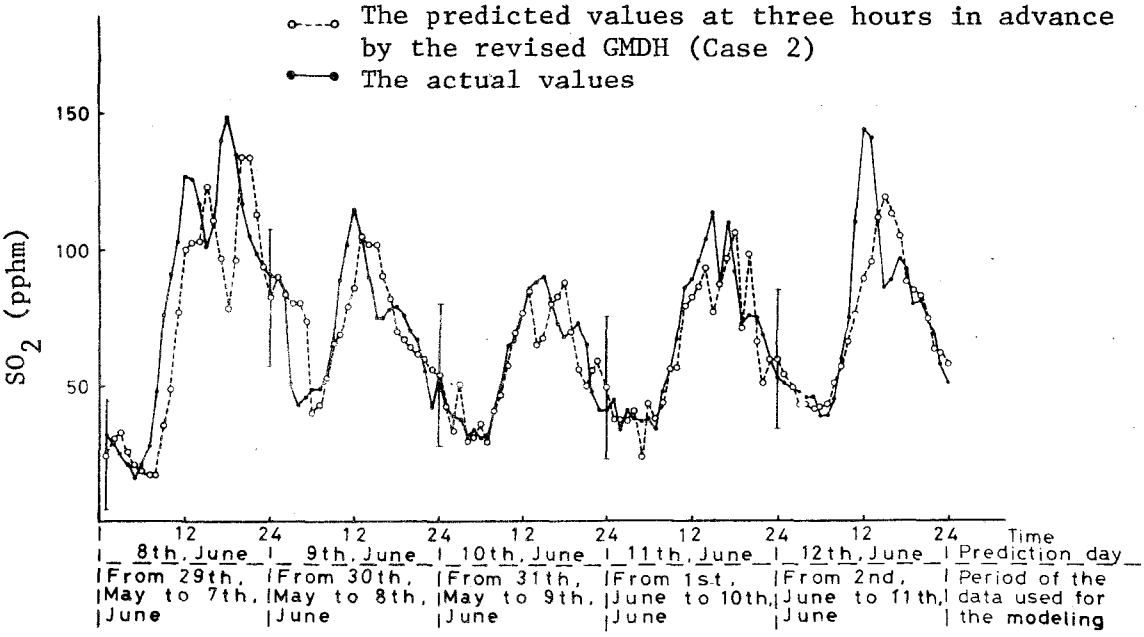


Fig. 4.13 The predicted values at three hours in advance by the revised GMDH, the confidence intervals and the actual values

Table 4.2 Input variables selected in the revised GMDH and the maximum order

Day / Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
x_2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
x_3			○		○	○	○	○	○	○	○	○	○	○	○
x_4	○		○	○	○	○	○	○	○	○	○	○	○	○	○
x_5	○	○		○		○	○		○	○	○	○	○	○	○
x_6	○		○	○		○								○	○
x_7			○					○							
Maximum Order	2	1	2	4	1	1	1	2	2	1	2	2	2	2	2

input variables selected in the prediction model at three hours in advance and the maximum order. As an example of a precise model description, the complete model for June 4 is shown as follows:

The third layer:

$$v_1 = 0.661 + 1.048z_1 - 0.018z_7^2 \qquad \text{(4-th order polynomial)}$$

$$v_2 = z_1 \qquad \text{(4-th order polynomial)}$$

$$v_3 = 0.003 + 1.288z_4 - 0.411z_7 \qquad \text{(4-th order polynomial)}$$

$$v_4 = z_2 \qquad \text{(2-nd order polynomial)}$$

$$v_5 = z_4 \qquad \text{(4-th order polynomial)}$$

$$v_6 = z_3 \qquad \text{(1-st order polynomial)}$$

The second layer:

$$z_1 = 0.063 + 0.874y_1 + 0.288y_7 + 0.053y_1y_7 - 0.042y_7^2$$

$$z_2 = - 0.005 + 0.913y_1 + 0.289y_5$$

$$z_3 = y_1$$

$$z_4 = 0.350 + 0.946y_2 + 0.046y_2y_7 - 0.056y_7^2$$

$$z_7 = y_3$$

The first layer:

$$y_1 = - 0.070 + 0.775x_1 - 0.202x_2$$

$$y_2 = - 0.078 + 0.595x_1$$

$$y_3 = - 0.067 + 0.487x_2$$

$$y_5 = 0.701 + 0.395x_4 - 0.005x_6 - 0.005x_6^2$$

$$y_7 = 0.757 + 0.364x_5 - 0.005x_6 - 0.005x_6^2$$

where $x_\ell = x(t+1-\ell)$ $(\ell=1,2,\cdots,7)$. The final model is constructed as an average of six polynomials obtained in the third layer.

B. The comparison with the prediction results obtained by the basic GMDH

Heuristics used in the basic GMDH is as follows: As the partial polynomial,

$$y_k = b_0 + b_1 x_i + b_2 x_j + b_3 x_i x_j + b_4 x_i^2 + b_5 x_j^2$$

is used, and seven variables are selected as intermediate variables in each layer. We use the following two divisions.

Division I : (Tr.) odd-numbered data
(Ch.) even-numbered data

Division II : (Tr.) data of 1~7-th days
(Ch.) data of 8~10-th days

Figure 4.14 shows the prediction error of SO_2 from the actual data at three hours in advance obtained by the basic GMDH, and the prediction error is compared with that obtained by the revised GMDH. Here, the prediction model II is used and the data of 10 days are used for modeling. We find from Fig. 4.14 that the complete polynomials constructed by the basic GMDH are very unstable in both divisions, as seen from very large prediction errors for June 5 and 8. In the basic GMDH algorithm, the structure of the partial polynomials is fixed to a predetermined description, therefore the partial polynomials obtained

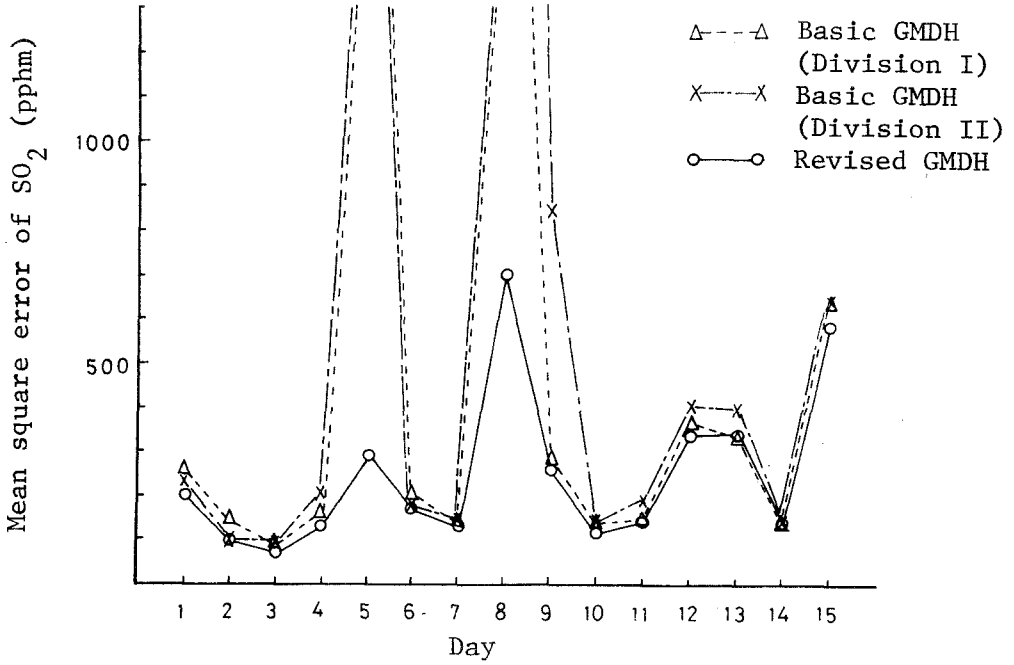


Fig. 4.14 Comparison of the prediction error at three hours in advance for the revised GMDH model and the basic GMDH model

are not the optimal regression equations. Hence, the complete polynomial is no longer an optimal regression equation, and it sometimes becomes very unstable. Table 4.3 shows the input variables selected in the basic GMDH and the maximum order. The models obtained by the basic GMDH are very complex compared with the models obtained by the revised GMDH. Since the basic GMDH needs to divide the original data into training data and the checking data, the identified results depend heavily on this division.

C. The comparison with the prediction results obtained by the linear statistical models

The prediction results obtained by the revised GMDH are compared with the results obtained by the linear statistical models such as a

Table 4.3 Input variables selected in the basic GMDH and the maximum order

(a) Division procedure I

Day / Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
x_2		○		○	○	○									
x_3						○		○		○	○	○	○		
x_4		○		○	○	○	○		○	○			○	○	○
x_5	○	○		○	○	○		○		○		○	○		
x_6			○	○	○			○		○		○	○		○
x_7	○	○		○	○	○	○	○	○	○		○	○	○	
Maximum Order	4	16	2	32	32	64	4	8	4	64	2	16	16	4	4

(b) Division procedure II

Day / Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
x_2	○	○	○	○	○	○	○	○	○						○
x_3					○	○	○	○	○					○	
x_4					○		○	○	○			○			○
x_5					○			○			○				
x_6		○	○	○				○		○					
x_7													○	○	
Maximum Order	2	4	8	4	64	4	32	16	8	2	2	2	2	4	4

regression model and an autoregressive model. Precise description of the linear statistical models can be found in [8]. Figure 4.15 shows the comparison of the prediction error incurred by the revised GMDH model with that by the linear models. We can see from Fig. 4.15 that the revised GMDH gives better performance than the linear models. The average computation time for constructing a revised GMDH model to predict the values of 24 hours is about 17 seconds, where NEAC 2200/700 of the Computation Center in Osaka University was used. The revised GMDH needs much more computation time for model building than the linear statistical model building, but the computation time is not too large for the practical use.

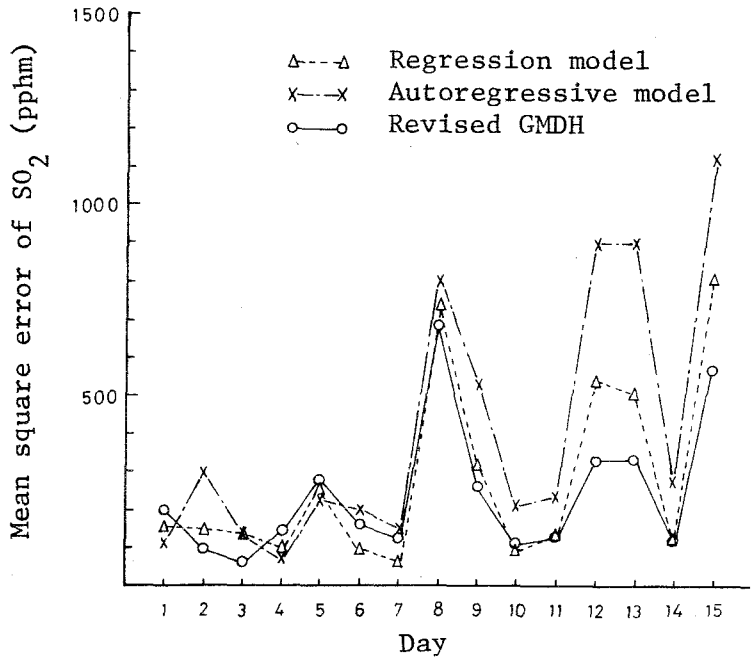


Fig. 4.15 Comparison of the prediction error at three hours in advance for the revised GMDH model and linear models

4.4 Concluding Remarks

In this Chapter, the revised GMDH algorithm developed in Chapter 2 is applied to two air pollution problems of identifying steady state and unsteady state air pollution models.

In 4.2, a method of identifying a steady state spatial pattern of air pollution concentration in a large area, is developed. By comparing three models, the effectiveness of the combined model of the source-receptor matrix and the revised GMDH is justified. A steady state (monthly or yearly average) large-spatial model developed in 4.2 would be useful for regional environmental planning and environmental impact assessment, since it could help to find

(a) Relationship between the environmental capacity and the level of pollution sources

(b) Allocation of the level of each pollution source to each polluter for regulating total amount of air pollution.

In 4.3, nonlinear statistical models for short-term prediction of air pollution concentration are identified by the revised GMDH algorithm. Comparing the prediction results of the revised GMDH model with those of the linear statistical models and the basic GMDH model, the following results are obtained.

(a) The suitable length of data used for short-term predictions in Tokushima is about 10 days.

(b) For the prediction at three hours in advance, the prediction model II gives better performance than the prediction model I. Furthermore, we cannot expect the improvement of the prediction accuracy by using

the information of wind direction and wind velocity at the concerning point only.

(c) The models obtained by the basic GMDH become very unstable in some days, however, the models obtained by the revised GMDH are always stable. Furthermore, the revised GMDH model is much simpler and gives better performance than the basic GMDH model.

(d) Although it takes longer computation time for modeling, the revised GMDH model gives better performance than the linear statistical models as well.

From these prediction results, the effectiveness of the nonlinear models obtained by the revised GMDH is justified for short-term prediction of air pollution concentration.

REFERENCES

[1] Akizuki, K. and K. Shirai: On construction of air pollution model by statistical method, Sympo. on Modeling for Prediction and Control of Air Pollution, 53-60, Kyoto University, Kyoto (June 1974)

[2] Allen, D.M.: The relationship between variable selection and data augmentation and a method for prediction, Technometrics, Vol. 16, No. 1, 125-127 (1974)

[3] Bibbero, R.J. and I.G. Young: Systems Approach to Air Pollution Control, Wiley, New York (1974)

[4] Draper, N.R. and H. Smith: Applied Regression Analysis, Wiley, New York (1966)

[5] Hino, M.: Prediction of atmospheric pollution by Kalman-Filtering, Sympo. on Modeling for Prediction and Control of Air Pollution, 77-84, Kyoto University, Kyoto (June 1974)

[6] Naito, M. and S. Otoma: On source-receptor matrix and its application, (in Japanease) Environmental Technology, Vol. 3, 545-549 (1974)

[7] Naito, M. and S. Otoma: A simple mathematical form for the urban environmental pollution management, Proc. of the International Congress on the Human Environment, 556-559, Kyoto (1975)

[8] Sawaragi, Y., T. Soeda, et al.: The predictions of air pollution levels by nonphysical models based on Kalman filtering method, Trans. ASME, Series G, Vol. 98, No. 4, 375-386 (1976)

[9] Shieh, L.J., P.K. Halpern, B.A. Clements, H.H. Wang and F.F. Abraham: Air quality diffusion model; Application to New York city, IBM J. Res. Develop., Vol. 16, 162-170 (1972)

[10] Tamura, H. and T. Kondo: Large-spatial pattern identification of air pollution by a combined model of source-receptor matrix and revised GMDH, Proc. IFAC Sympo. on Environmental Systems Planning, Design and Control, 373-380, Kyoto (Aug. 1977)

[11] Tamura, H. and T. Kondo: Nonlinear modeling for short-term prediction of air pollution concentration by a revised GMDH, Proc. International Conference on Cybernetics and Society, IEEE Syst., Man, Cybern. Society, 596-601, Tokyo and Kyoto (Nov. 1978)

CHAPTER 5 APPLICATION TO RIVER POLLUTION PROBLEM

5.1 Introduction

In the river quality system, there are many complex phenomena such as biochemical reaction, thermal behavior, sedimentation, and photosynthetic oxygen production, therefore the structure of the physical model considering the influences of these phenomena is becoming very complex [1,3]. Parameter estimation procedure of the physical model, which has been used for predicting pollution levels of the river quality, is a very complicated one.

In this Chapter, nonlinear statistical modeling of steady state river quality system is developed. The methodology used for modeling is the revised GMDH algorithm of generating optimal intermediate polynomials which is discussed in Chapter 3 [2]. By using measured data of river quality such as BOD and DO concentrations in Bormida river, Italy [3], we intend to construct two kinds of steady state models of river quality. In steady state model I, we intend to discover a suitable structure of the Bormida river by using no a priori information of the system structure. It is shown that the structure of the revised GMDH model depends on the statistical properties of the data used for modeling. Furthermore, the

prediction accuracy obtained by the revised GMDH model is compared with that obtained by the physical model which is called as Streeter-Phelps model. It is shown that the revised GMDH model gives much better performance for DO concentration compared with the physical model. In steady state model II, we intend to approximate the Bormida river system as a polynomial of input variables. But it is shown that it is difficult to approximate the DO part of the model as a polynomial of input variables, because the system structure for the DO concentration is very complex.

5.2 Modeling of the Steady State River Quality [3,4]

BOD and DO concentration have been widely accepted as the important indexes of organic river quality. The dynamic behavior of these levels is described as a generalized Streeter-Phelps model

$$\frac{\partial b}{\partial t} + V \frac{\partial b}{\partial \ell} = - \left(k_1(T) + \frac{k_3(V)}{A}\right) b \qquad (5.1.a)$$

$$\frac{\partial c}{\partial t} + V \frac{\partial c}{\partial \ell} = - k_1(T) b + \frac{k_2(T,Q)}{H(Q)} (c_s(T) - c) + \frac{k_4}{A} \qquad (5.1.b)$$

where, b is the BOD concentration (mg/ℓ), c is the DO concentration (mg/ℓ), c_s is the saturation level of DO concentration (mg/ℓ), k_1 is the deoxygenation rate (1/day), k_2 is the reoxygenation rate (m/day), k_3 is the suspended BOD sedimentation rate (m^2/day), k_4 is the photosynthetic

oxygen production rate ((mg/ℓ)(m^2/day)), t is the time (day), ℓ is the distance (km), T is the water temperature (°C), A is the cross sectional area (m^2), Q is the flow rate (10^3 m^3/day), V (= Q/A) is the average stream velocity (km/day) and H is the mean river depth (m). Here, for simplicity, it is assumed that the cross sectional area A is not varying along the river and the velocity V is constant in space and time. Then, the steady state BOD and DO concentrations satisfy the differential equations

$$\frac{db}{d\ell} = - K_1(T,Q)b \qquad (5.2.a)$$

$$\frac{dc}{d\ell} = - K_2(T,Q)b + K_3(T,Q)(c_s-c) + K_4(Q) \qquad (5.2.b)$$

where the functions K_h (h=1,2,3,4) depend upon the two independent variables Q and T, i.e.

$$K_1(T,Q) = k_1(T)/V(Q) + k_3(V(Q))/Q \qquad (5.3.a)$$

$$K_2(T,Q) = k_1(T)/V(Q) \qquad (5.3.b)$$

$$K_3(T,Q) = k_2(T,Q)/(H(Q)V(Q)) \qquad (5.3.c)$$

$$K_4(Q) = k_4/Q \quad . \qquad (5.3.d)$$

The solution to eq. (5.2) is well known and is obtained as

$$b(l,K_1,b_0) = b_0 e^{-K_1 l} \tag{5.4.a}$$

$$c(l,\ K_1,K_2,K_3,K_4,b_0,c_0) = c_s + K_4/K_3 - [c_s+(K_4/K_3)-c_0]e^{-K_3 l}$$

$$+ [K_2 b_0/(K_1-K_3)][e^{-K_1 l} - e^{-K_3 l}] \tag{5.4.b}$$

where b_0 and c_0 are BOD and DO concentrations near the discharge point, and it is assumed that there is no discharge inside of the subject range.

Data are measured for n different steady states. The i-th steady state is characterized by the flow rate Q^i and the temperature T^i. BOD and DO concentrations are measured at r points along the river as shown in Fig. 5.1. Suppose the following measured data are available.

$$(\ b_0^i,\ c_0^i\ ), \qquad (i=1,2,\cdots,n) \tag{5.5.a}$$

$$(\ b_j^i,\ c_j^i\ ), \qquad (i=1,2,\cdots,n;\ j=1,2,\cdots,r) \tag{5.5.b}$$

5.2.1 Parameter estimation of the physical model [3]

Here, the estimation method of parameters contained in eqs. (5.4.a) and (5.4.b) is introduced briefly. This method is proposed by Rinaldi, et al. [3]. The structures of functions K_h (h=1,2,3,4) contained in eqs. (5.4.a) and (5.4.b) are assumed as

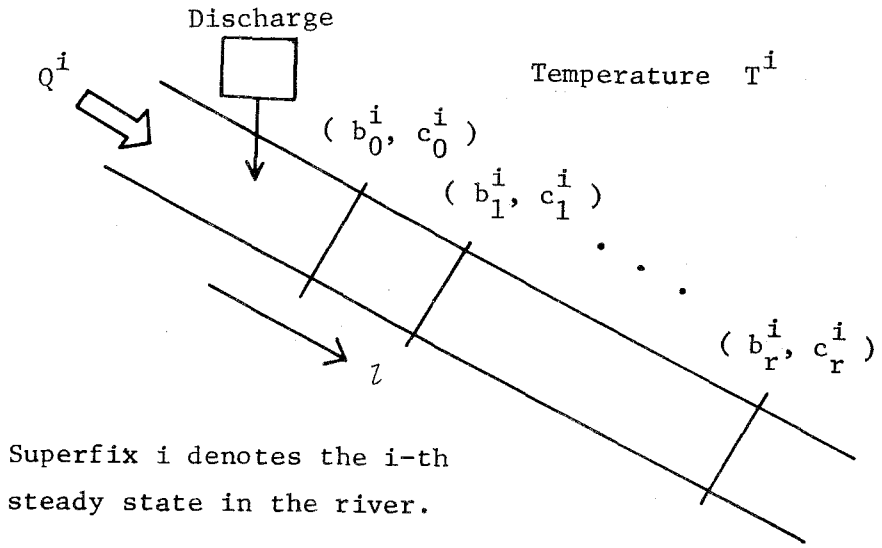


Fig. 5.1 The variables measured in a river

$$K_h = K_h(\underline{\theta}_h, T, Q), \qquad (h=1,2,3,4)$$

where $\underline{\theta}_h$ (h=1,2,3,4) denote the parameters contained in K_h. By using measured data (5.5.a) and (5.5.b), parameters $\underline{\theta}_h$ (h=1,2,3,4) are estimated so as to minimize the criterion

$$J = \sum_{i=1}^{n} J^i \qquad (5.6.a)$$

where

$$J^i = \sum_{j=1}^{r} [\lambda \varepsilon_b^{ji} + (1 - \lambda) \varepsilon_c^{ji}], \qquad 0 \leq \lambda \leq 1 \qquad (5.6.b)$$

$$\varepsilon_b^{ji} = [b(l_j, K_1^i, b_0^i) - b_j^i]^2 \qquad (5.6.c)$$

$$\varepsilon_c^{ji} = [\ c(\ell_j, K_1^i, \cdots, K_4^i, b_0^i, c_0^i) - c_j^i\ ]^2 \qquad (5.6.d)$$

and ε_b^{ji} is a square error between the measured value of BOD concentration of the i-th steady state at the j-th point and the estimated value by eq. (5.4.a). ε_c^{ji} is a square error for DO concentration, and λ is a weight for the BOD concentration. It is very difficult to estimate parameters $\underline{\theta}_h$ (h=1,2,3,4) directly so as to minimize J in eq. (5.6.a) because the dimension of $\underline{\theta}_h$ is very high. Therefore, the following procedure is used to estimate $\underline{\theta}_h$. Firstly, by using the data measured in each steady state, functions K_h^i (h=1,2,3,4; i=1,2,$\cdots$,n) are estimated so as to minimize J^i (i=1,2,$\cdots$,n). Then, by using the estimated values of K_h^i, parameters $\underline{\theta}_h$ are estimated so as to minimize

$$J' = \sum_{i=1}^{n} \sum_{h=1}^{4} (\ K_h(\underline{\theta}_h, T^i, Q^i) - K_h^i\ )^2. \qquad (5.7)$$

More precise description of this procedure can be found in [3].

5.2.2 Modeling of the steady state system by the revised GMDH [4]

Here, the steady state model of the river quality is constructed by the revised GMDH algorithm developed in Chapter 3. In this revised GMDH algorithm, optimal intermediate polynomials, which express the direct relationship between the input and output variables, are generated automatically in each selection layer so as to minimize AIC

and the complete polynomial is obtained from the optimal intermediate polynomial remained in the final layer. By using the revised GMDH algorithm, the following two steady state models are constructed.

A. Steady state model I

Steady state model in the form of eq. (5.2) is constructed. Two variables b(j+1) and c(j+1) are used as output variables and five variables b(j), c(j), Q^{-1}, $Q^{-0.5}$ and T are used as input variables. Here, it is assumed that the measuring points of BOD and DO concentrations are equally spaced along the river. The steady state model to be identified by the revised GMDH is

$$b(j+1) = f_1(b(j),c(j),Q^{-1},Q^{-0.5},T) \tag{5.8.a}$$

$$c(j+1) = f_2(b(j),c(j),Q^{-1},Q^{-0.5},T) \ . \tag{5.8.b}$$

Equation (5.8) can be transformed to

$$\frac{b(j+1) - b(j)}{\Delta l} = \frac{1}{\Delta l} \{ f_1(b(j),c(j),Q^{-1},Q^{-0.5},T) - b(j) \} \tag{5.9.a}$$

$$\frac{c(j+1) - c(j)}{\Delta l} = \frac{1}{\Delta l} \{ f_2(b(j),c(j),Q^{-1},Q^{-0.5},T) - c(j) \} \ . \tag{5.9.b}$$

In eqs. (5.9.a) and (5.9.b), if the left hand sides of the equations are approximately replaced by db/dl and dc/dl, respectively, steady state

model in the form of eq. (5.2) can be obtained.

B. Steady state model II

Steady state model in the form of eq. (5.4) is constructed. Two variables $b(\ell)$ and $c(\ell)$ are used as output variables and seven variables b_0, c_0, ℓ, ℓ^{-1}, $Q^{0.5}$, $Q^{-0.5}$ and T are used as input variables. In this case, the physical interpretation of the model constructed by the revised GMDH is not possible, because eq. (5.4) cannot be described as a physically meaningful polynomial in terms of these input variables. That is, a revised GMDH model obtained is a nonphysical model. The steady state model to be identified by the revised GMDH is

$$b(\ell) = g_1(b_0, c_0, \ell, \ell^{-1}, Q^{0.5}, Q^{-0.5}, T) \qquad (5.10.a)$$

$$c(\ell) = g_2(b_0, c_0, \ell, \ell^{-1}, Q^{0.5}, Q^{-0.5}, T) . \qquad (5.10.b)$$

For constructing this model, measuring points of BOD and DO concentrations are not necessarily equally spaced along the river.

5.3 Modeling of the Steady State Bormida River Quality [3,4]

The steady state model of the Bormida river shown in Fig. 5.2 is constructed by applying the revised GMDH algorithm to the data shown in Table 5.1 and the predicted results obtained by the revised GMDH model are compared with those obtained by the physical model estimated by

Rinaldi, et al.

The data measured in the Bormida river are used [3], where four variables, BOD concentration b, DO concentration c, flow rate Q and temperature T are measured as shown in Table 5.1. Data of BOD and DO concentrations are the daily average value and measured at six points which are located with the interval of about 10~15 km along the river. Here, the data at the fourth point is not the measured value but the value obtained by a linear interpolation. Data of the temperature are the average values obtained at six points but the measurement time is different for each steady state, and therefore it is difficult to find a significant interpretation for the data. We simply neglected the effect of the temperature variation. Fifteen steady states are measured (n=15). Among them thirteen steady states data are used for modeling and two steady states data are used for model validation.

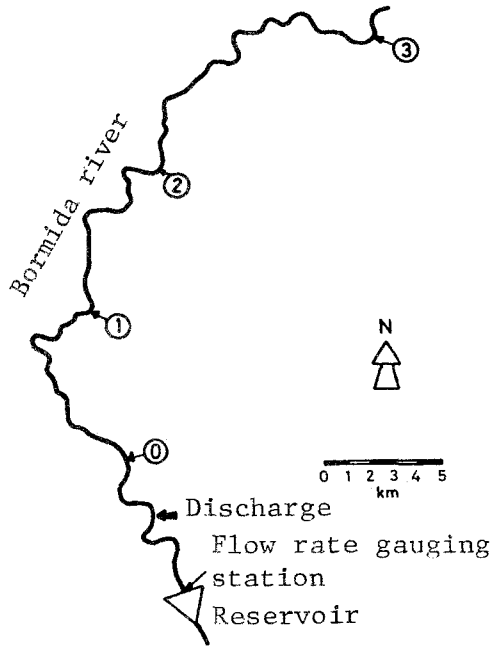


Fig. 5.2 The Bormida river and locations of measurement stations [3]

Table 5.1 The data used for modeling and model validation [3]

Station number / Distance [Km]	0 / 4.20	1 / 14.00	2 / 25.00	3 / 40.00	(4) / (54.00)	5 / 68.00	Flow rate [10^3 m^3/day]	Water Temperature[°C] Average	Range
Steady state									
1	200.0 0.0	118.0 4.5	64.0 5.5	38.0 6.5	24.0 7.8	10.0 9.0	55	17.5	4.2
2	120.0 3.0	92.0 5.5	72.0 9.0	58.0 9.5	41.0 9.5	24.0 9.5	60	9.0	5.1
3	162.0 1.0	126.0 3.0	110.0 5.0	66.0 6.5	53.0 8.5	40.0 10.5	125	0.5	5.0
4	105.0 2.0	84.0 5.0	70.0 5.5	44.0 6.0	41.0 6.8	38.0 7.5	100	19.0	3.0
5	125.0 1.5	78.0 3.5	46.0 4.5	18.0 5.5	16.0 6.3	14.0 7.0	75	18.0	3.2
6	125.0 2.0	86.0 5.0	70.0 6.0	46.0 6.0	33.0 6.3	20.0 6.5	80	17.0	3.3
7	68.0 2.0	56.0 6.0	50.0 7.0	34.0 9.5	29.0 10.8	24.0 12.0	225	5.0	2.5
8	145.0 0.0	72.0 1.2	68.0 2.2	30.0 3.6	23.0 4.7	16.0 5.8	100	25.0	3.7
9	200.0 0.0	104.0 4.0	98.0 6.0	60.0 6.0	59.0 6.5	58.0 7.0	55	10.0	8.9
10	90.0 4.0	70.0 4.0	68.0 8.0	58.0 9.0	40.0 9.0	22.0 9.0	200	1.8	3.5
11	80.0 6.0	60.0 8.0	50.0 10.0	36.0 10.5	30.0 10.8	24.0 11.0	250	3.5	2.4
12	135.0 0.5	100.0 4.0	85.0 5.0	62.0 6.0	56.0 7.0	50.0 8.0	125	11.8	2.4
13	70.0 3.0	60.0 6.0	44.0 7.0	46.0 7.5	34.0 7.8	22.0 8.0	200	16.0	2.5
14	85.0 3.0	70.0 6.0	55.0 7.0	40.0 9.0	30.0 9.3	20.0 9.5	200	11.5	5.5
15	80.0 2.5	40.0 5.0	30.0 7.0	20.0 8.5	16.0 8.8	12.0 9.0	150	16.0	6.0

5.3.1 Results of parameter estimation of the physical model [3]

Parameters of physical model are estimated by using the procedure described in 5.2.1. The data of the 1~13-th steady states are used for modeling. The structure of K_h (h=1,2,3,4) are assumed as

$$K_h(\underline{\theta}_h, Q) = \theta_{h1} Q^{\theta_{h2}} \tag{5.11}$$

where, $\underline{\theta}_h = (\theta_{h1}, \theta_{h2})$.

Functions K_h^i (h=1,2,3,4; i=1,2,$\cdots$,13) are estimated so as to minimize J^i (i=1,2,$\cdots$,13) in eq. (5.6.b) and as the result

$$K_1 \simeq K_2 , \qquad K_4 \simeq 0 \tag{5.12}$$

is obtained. This result shows that BOD and DO concentrations in the Bormida river can be described as the Streeter-Phelps model. Then parameters $\underline{\theta}_1$ and $\underline{\theta}_3$ are estimated so as to minimize J' in eq. (5.7) and

$$\frac{db}{dt} = - 0.2\ Q^{-0.43} b \tag{5.13.a}$$

$$\frac{dc}{dt} = - 0.2\ Q^{-0.43} b + 16.4\ Q^{-0.8}(c_s - c) \tag{5.13.b}$$

is obtained.

5.3.2 Results of modeling by the revised GMDH [4]

A. Steady state model I

Four variables b(j), c(j), Q^{-1} and $Q^{-0.5}$ are used as input variables. Parameters used in the revised GMDH are

$$p = 2 , \quad L_1 = 10 , \quad m_1 = 6 .$$

1) BOD model identified by the revised GMDH

BOD models identified by the revised GMDH are shown in Table 5.2. The fourth model is identified by using all the data of 15 steady states. From Table 5.2, we can see that the structure of the model is varying slightly according to the measured data used for modeling. In the revised GMDH, the structure of the model is determined by using only the measured data, and therefore the dependence of the structure of the model on the statistical characteristics of the measured data cannot be avoided. But, if sufficiently many data can be used, the dependence can be reduced. The third model

$$b(j+1) = -4.22 + 0.920b(j) + 0.000037b(j)^2 - 0.0133Q^{-0.5}b(j)^2 \tag{5.14}$$

is identified by using the measured data of 1~13-th steady states. This model can be transformed to

$$\frac{b(j+1) - b(j)}{\Delta l} = \frac{1}{\Delta l}\{ -4.22 - 0.080b(j) + 0.000037b(j)^2 - 0.0133Q^{-0.5}b(j)^2 \} . \tag{5.15}$$

Since $\Delta l \simeq 10$ km, eq. (5.15) can be approximately reduced to

$$\frac{db}{dl} = -0.422 - 0.0080b + 0.0000037b^2 - 0.00133Q^{-0.5}b^2 . \tag{5.16}$$

We can find that the second order terms of BOD concentration are

contained in eq. (5.16), and the structure of the model is a little more complex than the physical model (5.2.a). In order to verify the effectiveness of eq. (5.14), the prediction errors for the 14-th and 15-th steady states of eq. (5.14) are compared with those of the physical model (5.2.a). In eq. (5.14), the BOD concentration b(1) is predicted by using the measured data b_0, and the BOD concentrations b(j+1) for j=1~4 are obtained by using the predicted values for j=0~3. Predicted results for the 14-th and 15-th steady states are shown in Figs. 5.3 and 5.4. It can be seen that the prediction accuracy obtained by the revised GMDH model (5.14) is identical with that obtained by the physical model (5.2.a).

Table 5.2 Structures of the BOD model I

Model	Prediction points	constant	b	b^2	$b\,Q^{-0.5}$	$b^2Q^{-0.5}$
1	4 , 5	-5.84	0.960	-0.00040	*	-0.011
2	9 , 10	-2.38	1.027	-0.00070	-2.06	*
3	14 , 15	-4.22	0.920	0.00004	*	-0.013
4	0	-3.82	0.900	0.00008	*	-0.013

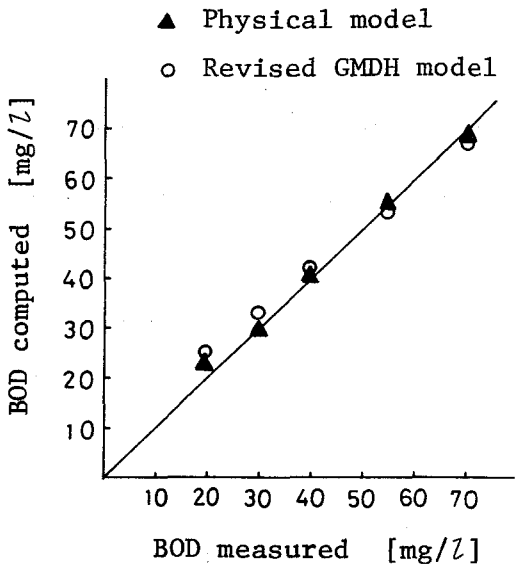


Fig. 5.3 Measured and computed values of BOD for 14-th steady state by model I-3

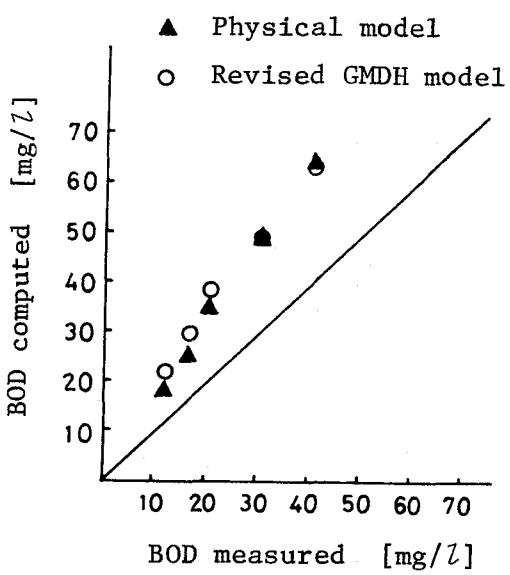


Fig. 5.4 Measured and computed values of BOD for 15-th steady state by model I-3

2) DO model identified by the revised GMDH

Identified DO model is shown in Table 5.3. The fourth model is identified by using all the data of 15 steady states. From Table 5.3, we can see that the structure of the model is varying remarkably according to measured data used for modeling. In particular, the terms concerned with the flow rate Q is remarkably varied. The reason for this is that the number of different measurement data for the flow rate are very few compared with the number of the terms contained in the model, and therefore the information contained in the input variable Q is not fully taken out from the data. The third model

$$c(j+1) = 6.72 + 0.431c(j) - 0.000203b(j)^2 + 0.00222Q^{-0.5}b(j)^2$$

$$- 46.1Q^{-0.5} + 3.91Q^{-0.5}c(j) \qquad (5.17)$$

is identified by using the measured data of 1~13-th steady states.

Table 5.3 Structures of the DO model I

Model	Prediction points	constant	c	b^2	$b^2Q^{-0.5}$	b^2Q^{-1}
1	4 , 5	2.39	0.895	0.00003	*	*
2	9 , 10	7.75	0.993	-0.00020	0.0024	*
3	14 , 15	6.72	0.431	-0.00020	0.0022	*
4	0	10.3	0.553	-0.00008	*	0.0080

Model	$Q^{-0.5}$	Q^{-1}	$c\ Q^{-0.5}$	$c\ Q^{-1}$
1	*	*	-1.19	*
2	-54.2	*	-10.4	78.6
3	-46.1	*	3.91	*
4	-118.	382.	*	18.3

This model can be transformed to

$$\frac{c(j+1) - c(j)}{\Delta \ell} = \frac{1}{\Delta \ell} \{ 6.72 - 0.569c(j) - 0.000203b(j)^2$$

$$+ 0.00222Q^{-0.5}b(j)^2 - 46.1Q^{-0.5} + 3.91Q^{-0.5}c(j) \} . \qquad (5.18)$$

Using $\Delta \ell \simeq 10$ km, eq. (5.18) can be approximately reduced to

$$\frac{dc}{d\ell} = 0.672 - 0.0569c - 0.0000203b^2 + 0.000222Q^{-0.5}b^2$$

$$- 4.61Q^{-0.5} + 0.391Q^{-0.5}c \quad . \qquad (5.19)$$

From this model, we can find that the second order terms b^2 and $Q^{-0.5}b^2$ are contained in both BOD model (5.16) and DO model (5.19). The terms $Q^{-0.5}$ and $Q^{-0.5}c$ are similar to $Q^{-0.8}$ and $Q^{-0.8}c$ contained in the physical model (5.2.b), respectively. In order to verify the effectiveness of eq. (5.17), the prediction errors for the 14-th and 15-th steady states of eq. (5.17) are compared with those of the physical model (5.2.b). In eq. (5.17), the DO concentration c(1) is predicted by using the measured data b_0 and c_0, and the DO concentration c(j+1) for j=1~4 are obtained by using the predicted values for j=0~3. Predicted results for the 14-th and 15-th steady states are shown in Figs. 5.5 and 5.6. From Fig. 5.5, it can be seen that the revised GMDH model (5.17) gives much better prediction accuracy for the 14-th steady state than that of physical model (5.2.b). From these prediction

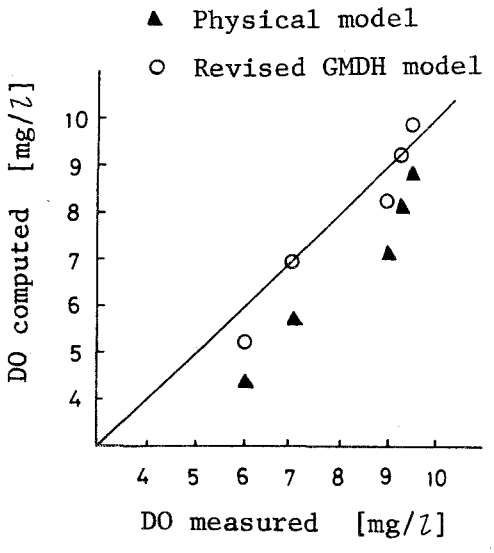


Fig. 5.5 Measured and computed values of DO for 14-th steady state by model I-3

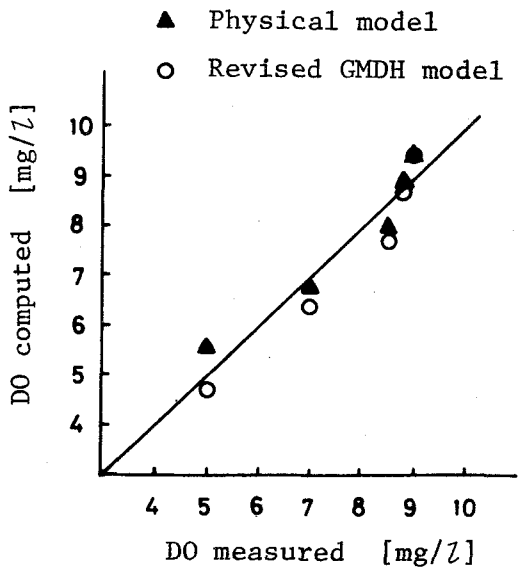


Fig. 5.6 Measured and computed values of DO for 15-th steady state by model I-3

results, we can see that the steady state model I identified by the revised GMDH algorithm is fairly reliable as the prediction model. Furthermore, the structure of the steady state model I is a little more complex than that of the physical model but they are very similar. This shows that the statistical analysis of the input and output data by the revised GMDH algorithm of using intermediate polynomials enables to give the important information concerned with the structure of the system which is very complex and completely unknown.

B. Steady state model II

Six variables of b_0, c_0, ℓ, ℓ^{-1}, $Q^{0.5}$ and $Q^{-0.5}$ are used as input variables. Parameters used in the revised GMDH are as follows.

$$p = 2 \ , \quad L_1 = 10 \ , \quad m_1 = 6$$

1) BOD model identified by the revised GMDH

By using the measured data of the 1~13-th steady states, BOD model is identified as

$$b(l) = 28.9 - 0.0268b_0 l - 0.0217l^2 + 1.523l + 0.000261b_0 l^2$$

$$+ 10.2b_0 Q^{-0.5} l^{-1} + 0.0004b_0^2 Q^{0.5} + 0.871b_0 c_0 Q^{-0.5}$$

$$- 0.000042b_0^2 c_0 Q^{0.5} \ . \qquad (5.20)$$

We can see that the structure of eq. (5.20) is more complex than the steady state model I (5.14). In order to verify the effectiveness of eq. (5.20), the prediction errors for the 14-th and 15-th steady states of eq. (5.20) are compared with those of the physical model (5.4.a). Predicted results for the 14-th and 15-th steady states are shown in Figs. 5.7 and 5.8. We can see that the revised GMDH model (5.20) has the same prediction accuracy as the physical model (5.4.a).

2) DO model identified by the revised GMDH

By using the measured data of the 1~13-th steady states, DO model is identified as

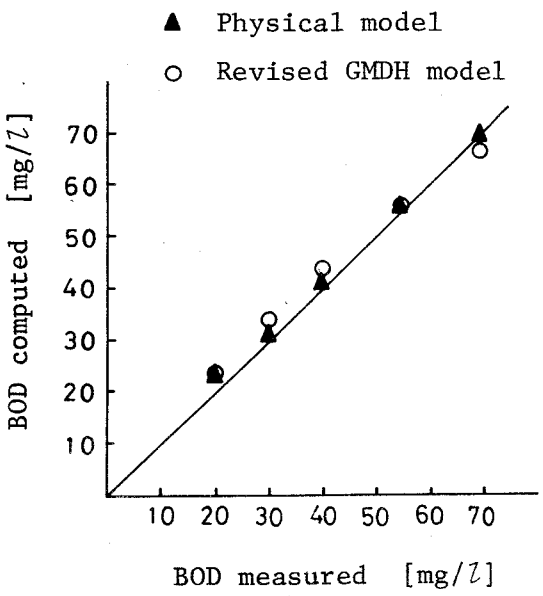


Fig. 5.7 Measured and computed values of BOD for 14-th steady state by model II

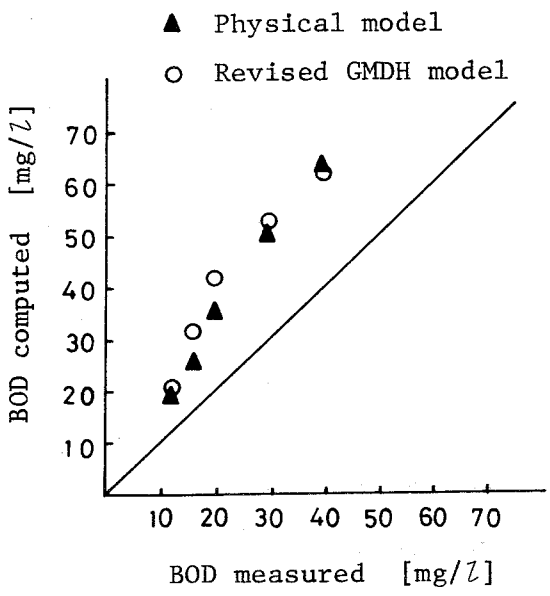


Fig. 5.8 Measured and computed values of BOD for 15-th steady state by model II

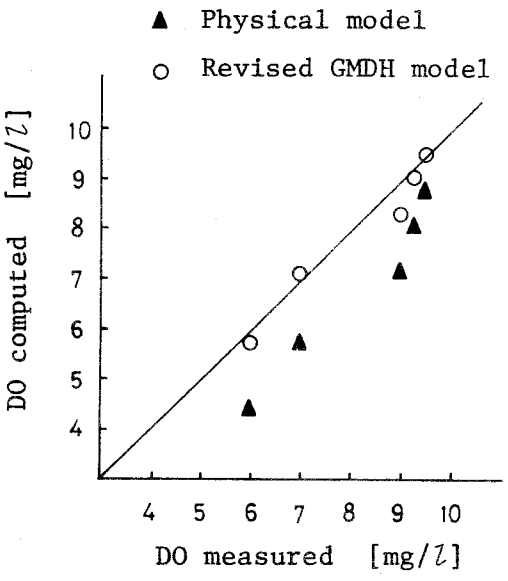


Fig. 5.9 Measured and computed values of DO for 14-th steady state by model II

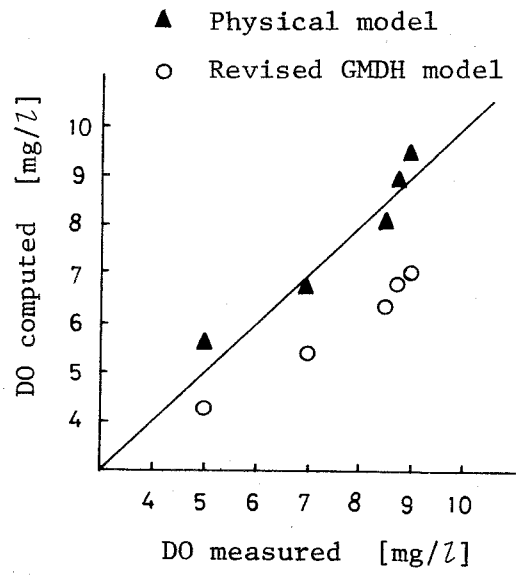


Fig. 5.10 Measured and computed values of DO for 15-th steady state by model II

$$c(l) = -34.8 + 1.74Q^{0.5} - 11.6l^{-1} - 0.00104l^2 + 189Q^{-0.5}$$

$$+ 9.26c_0Q^{-0.5} + 0.0106Q^{0.5}l - 0.000436b_0c_0l$$

$$+ 0.000004b_0l^2 + 0.000003b_0^2c_0l \ . \qquad (5.21)$$

We can see that the structure of eq. (5.21) is also more complex than the steady state model I (5.17). In order to verify the effectiveness of eq. (5.21), the prediction errors for the 14-th and 15-th steady states of eq. (5.21) are compared with those of the physical model (5.4.b). Predicted results for the 14-th and 15-th steady states are shown in Figs. 5.9 and 5.10. From Fig. 5.10, the revised GMDH model (5.21) gives worse prediction accuracy for the 15-th steady state than the physical model (5.4.b). The reason for this is that the structure of the system for the DO concentration is very complex and cannot be described as a polynomial approximation of six input variables used in steady state model II. From these prediction results, we cannot expect a good prediction accuracy for DO concentration in the steady state model II.

5.4 Concluding Remarks

In this Chapter, two kinds of steady state river quality models are constructed by applying the revised GMDH algorithm to the measured data in the Bormida river. By comparing the revised GMDH model with the

physical model estimated by Rinaldi, et al., the following results are obtained.

(a) Steady state model I identified by the revised GMDH gives the same prediction accuracy as the physical model for BOD concentration but gives better prediction accuracy than the physical model for DO concentration.

(b) In the revised GMDH models identified for the DO concentration, the steady state model I gives better prediction accuracy than the steady state model II. The reason for this is that the structure of the system for the DO concentration is very complex and cannot be described by a polynomial approximation of six input variables used in the steady state model II.

(c) The structure of the revised GMDH model is heavily dependent upon the statistical properties of the data used for modeling, because the structure of the model is determined by using only input-output data. In the case of the Bormida river, the terms of the flow rate in the revised GMDH model is particularly dependent on the data because of the lack of information contained in only a few different flow rate data.

(d) For the steady state model II identified by the revised GMDH algorithm, second order terms of BOD concentration are contained in both BOD and DO models. The other terms are similar to those of the physical model.

(c) In the physical model, the computation for estimating the parameters is quite complex, but in the revised GMDH model it is not.

From these investigations, the effectiveness of the revised GMDH algorithm is justified for constructing steady state models of river quality.

REFERENCES

[1] Beck, B.: A comparative case study of dynamic models for DO-BOD algae interaction in a freshwater river, International Institute for Applied System Analysis, Working Paper No. WP-78-16, (May 1978)

[2] Kondo, T. and H. Tamura: Revised GMDH algorithm of self-selecting optimal intermediate polynomials using AIC, (in Japanese) Trans. Soc. Instr. Control Engineers. (forthcoming)

[3] Rinaldi, S., P. Romano and R. Soncini-Sessa: Parameter estimation of a Streeter-Phelps type water pollution model, Proc. 4th IFAC Sympo. on Identification and System Parameter Estimation, Tbilisi, U.S.S.R (1976)

[4] Tamura, H. and T. Kondo: Nonlinear modeling for the steady state river quality by a revised GMDH, (in Japanese) Trans. Soc. Instr. Control Engineers. (submitted)

CHAPTER 6 CONCLUSION

In this thesis, two kinds of new revised GMDH algorithms are developed and applied them to modeling of air pollution and river pollution problems.

In Chapter 1, the fundamental concept of GMDH which is called the heuristic self-organization is described. Then, the algorithm of the basic GMDH proposed by Ivakhnenko is shown, and the advantages, disadvantages and heuristics involved in the basic GMDH are discussed. Then, the improvements, which have been made on the basic GMDH algorithm, are briefly surveyed, and the motivation to this thesis research is clarified.

In Chapter 2, a revised GMDH algorithm of generating optimal partial polynomials under the prediction error criterion is developed in which we do not require to divide the available data into two groups; the training data and the checking data. In this algorithm, all the data can be used not only as the training data but as the checking data, that is, the prediction error such as PSS and AIC calculated from all the data is used as a criterion for selecting intermediate variables and for stopping the multilayered computations. Therefore, the

identified results do not depend on the heuristics of dividing the data into two groups. Furthermore, the revised GMDH developed in Chapter 2 generates optimal partial polynomials automatically in each selection layer. The revised GMDH, therefore, has much better flexibility than that of the basic GMDH in constructing a complete polynomial. The revised GMDH algorithm is applied to a simple illustrative example and compared with the results obtained by the basic GMDH algorithm. Many advantages of the revised GMDH algorithm compared with the basic GMDH algorithm are clarified.

In Chapter 3, a revised GMDH algorithm of generating optimal intermediate polynomials under the prediction error criterion is developed. This revised GMDH algorithm generates optimal intermediate polynomials in each selection layer, which express the direct relationship between the input and output variables, so as to minimize the prediction error criterion evaluated by using all the data. Therefore, physically meaningful structures can be identified when the characteristics of the system are well reflected in the data. The revised GMDH algorithm is applied to the input-output data observed in a simple kinetic system, and we tried to discover the Newton's second law of motion. The result obtained is compared with that obtained by the revised GMDH of using partial polynomials. The effectiveness of the revised GMDH algorithm of using intermediate polynomials for identifying physically meaningful structure between the input and output variables is justified.

In Capter 4, the revised GMDH algorithm developed in Chapter 2 is applied to two kinds of air pollution problems, the steady state

modeling and unsteady state modeling of air pollution. In steady state modeling, a method of identifying a steady state spatial pattern of air pollution concentration in a large area, is developed. By comparing three models, the effectiveness of the combined model of the source-receptor matrix and the revised GMDH is justified. The combined model of this kind would be useful for regional environmental planning and environmental impact assessment. In unsteady state modeling, nonlinear statistical models for short-term prediction of air pollution concentration are developed. By comparing the prediction results of the revised GMDH model with those of the linear statistical models and the basic GMDH model, the following results are obtained.

(a) Suitable length of data used for short-term predictions in Tokushima is about 10 days.

(b) We cannot expect the improvement of the prediction accuracy by using the information of wind direction and wind velocity at the concerning point only.

(c) The revised GMDH model is very stable and simple, and furthermore it gives better performance than the basic GMDH model and the linear statistical models.

From these prediction results, the effectiveness of the nonlinear models obtained by the revised GMDH is justified for short-term prediction of air pollution concentration.

In Chapter 5, nonlinear models for steady state river quality is developed by the revised GMDH proposed in Chapter 3. By comparing the revised GMDH model with the physical model developed by Rinaldi, et al.,

the following results are obtained.

(a) Steady state model identified by the revised GMDH algorithm gives better prediction accuracy for DO concentration compared with the physical model.

(b) In the GMDH model, second order terms of BOD concentration are appeared both in BOD and DO models, while in the physical model only linear terms are taken into account. The linear terms in the GMDH model appeared are similar to those in the physical model.

(c) The structure of the revised GMDH model depends on the statistical properties of the data used for modeling, therefore, it is necessary that the characteristics of the concerning system are well reflected in the data used for modeling.

(d) In the physical model, the computation for estimating the parameters is quite complex, but the computation for obtaining the revised GMDH model is fairly simple.

From these results, the effectiveness of the revised GMDH algorithm is justified for constructing steady state river quality models.

The advantages of the revised GMDH algorithms developed in this thesis are now clarified both from the methodological point of view and from the practical point of view. Finally, it should be noted that we need further researches to develop

(a) Multivariate GMDH for identifying nonlinear multi-input multi-output systems

(b) On-line recursive GMDH for updating the model whenever new time series data are obtained in the concerning nonlinear dynamical system.