Journal of the KSQC
Vol. 14, No. 1, 1986

資料취급의 집단적 방법(GMDH)을 사용한 予測의 精度에 관한 硏究

A Study on the Accuracy of the Forecasting Using Group Method of Data Handling

趙　　巖*

ABSTRACT

The purpose of this study has been finding where GMDH (Group Method of Data Handling) lies in accordance with comparing other methods and ascertaining the effectiveness of GMDH at the systems of forecasting method. Other methods used for the comparison are : multiple regression model, Brown's third exponential smoothing model. Also the study has reviewed how the expected value and equation are changed by GMDH. At the same time, the study has also reviewed various characteristics made with comparatively a few data.

In conclusion, GMDH is better than the other method in point of view fitness, high effectiveness in self-selection and self-construction of the variables.

Ⅰ. 序　論

본 硏究는 자료취급의 집단적방법(Group Method of Data Handling;GMDH)을 다른 方法과 比較하는데 따라 予測方法의 体系중에서 GMDH가 어느 位置에 있는가를 알고, GMDH의 有効性을 확인해 보기 위한것에 目的을 둔다. 比較하는 다른 方法들은, 多元回歸Model,(變數逐次選擇法도 포함), 指數平滑Model로서는 Brown의 3次指數平滑法을 사용했다. 또한 data數 20個 중에서 사용하는 個數를 제한한 경우에, GMDH에 의한 予測値 및 予測式이 어떻게 變化하는가를 관찰하고 GMDH가 비교적 적은 data로서 算出된다는 특징에 대하여 檢討해 본다.

Ⅱ. GMDH와 복잡한 SYSTEM의 予測

A. G. Ivakhnenko에 의해 개발된 GMDH는 cybernetics의 一般性 回復에 관한 답으로서 發

*東国大学校 工科大学 産業工学科

見的 自己組織化(Heuristic Self-organization) 의 原理에 기본을 두고 本質的으로 복잡한 SYSTEM 즉

a) 대단히 많은 變數와 parameter의 存在(高次元).

b) 상호의 관계가 非線形.

c) 原因과 結果, 入力과 出力의 관계를 찾아내는 것이 原因的으로나 實際的으로나 不可能.

이러한 것을 취급하는 一般的方法 이다. 이 方法의 特徵은

a) 적은 入出力 data에 복잡한 多變數, 非線形系의 予測이 可能하다는 것.

b) 多變數 이기는 하나 計算量이 적고, 종래의 確率的 予測法과 比較해서 알고리즘이 安定 하다는 것.

c) 數値的인 精度의 面에서 어느정도 복잡한 것이라도 數學的인 記述이 可能 하다는것.

등이다. 따라서 發見的 自己組織化(Heuristic Self-organization)의 SYSTEM은 多層 또는 階層構造를 가지고 있으며, 그 各層에 有用한 情報의 積分的 혹은 閾値的 自己選擇이 필요한 SYSTEM이다. 이 發見的 規範에 기본을 두고 自己選擇을 有用한 것으로 하기위해, 하나 또는 랜덤한 變數의 組合發生器가 사용되며 그 結果 各層마다 SYSTEM의 變數에 의한 記述의 복잡함이 增大한다. 入出力 data에 기본을 둔 非線形系의 問題에 관해서는 構造가 주어진 parameter 推定에 歸着되어진 경우를 제외하고는 많은 硏究가 없었다. 겨우 非parametic 한 一般的方法으로서, Volterra級數에 의한 定常確率過程에 대하여 Kolmogorov-Gabor의 多項式表現을 基礎로 한 Gabor의 Universal Nonlinear Filter에 의한 方法이 있다. 그러나 Volterra級數나 高次의 多項式에 의한 非線形系의 予測問題는 推定해야할 係數 및 그에 필요한 data의 量이 많아야 하며, 多變數에 대해서는 計算量과 더불어 行列計算의 安定性 이라는 面에서 實用으로 하기에는 곤란하다.

Ⅲ. 多項式에 의한 基本的 알고리즘

GMDH의 알고리즘 순서는 계략적으로 다음과 같다.

$x_i (i=1,2,\cdots,N_1\cdots)$; 入力變數

y ; 出力變數

入力과 出力 사이에 다음과 같은 非線形關係

$y=f(x_1, x_2, \cdots\cdots, x_N, \cdots\cdots)$ 에서

step 1 ; y에 관계하는 入力變數의 prehistory(y에 영향을 주는 期間)을 定하고, 전부 入力變數로 한다. 따라서 必要하다면 data의 normalization등의 前處理를 행한다.

step 2 ; 出力 y와 各入力變數의 相關을 취하여 相關係數가 큰것을 "有用한" 入力變數로서 남기고 相關係數가 작은것을 "有害한" 入力變數로서 버린다.

step 3 ; 기초 data를 training data와 Checking data로 나눈다. 나누는 方法은 分散이 큰것을 training으로 작은것을 checking으로 한다.

step 4 ; 入力變數 x_k($k=1, 2, \cdots, N$ 의 2個의 組合 x_i, x_j에 대하여 中間變數 Z_k($k=1, 2, \cdots N(N-1)/2$)을

$$Z_k = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j \cdots\cdots (1)^{註1)}$$

에 의해 만들어 진다.

단 (1)式 右邊의 a_i($i=0, 1, \cdots, 5$)은 training data에 기본을 둔 自乘平均誤差임.

$$\varepsilon_k^2 = \overline{(y-z_k)^2} \cdots\cdots (2)^{註2)}$$

을 最小가 되게 決定한다.

step 5 ; step 4에서 training data에 기초를 두고 決定된 係數를 사용하여 checking data을 (2)式에 의해 變換하고, (2)式의 自乘平均

註1) Kolmogorov-Gabor의 多項式

$$y = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k + \cdots\cdots$$ 에 基本을 둔것임.

註2) "———"는 Sample平均을 意味

池田三郎, 棋木義一, (1975), "GMDH と複雜な系の同定・予測, 計測と制御14-2

誤差를 變換시킨 checking data에 대해 計算한다.

step 6 ; $x_i = Z_i$, $x_j = Z_j$로 하여 step 4에 대입하여 다음의 中間變數를 얻는다.

step 7 ; Checking data에 대하여 計算된 自乘平均誤差의 最小 것이 前回의 最小인 自乘平均誤差를 넘은 경우 計算을 停止한다.

이러한 순서로서 各種의 發見的 最適化를 행함에 따라 全体로서 入出力關係의 最適이 되어진다. 以上의 알고리즘을 흐름도로 나타내면 그림 1 과 같다.

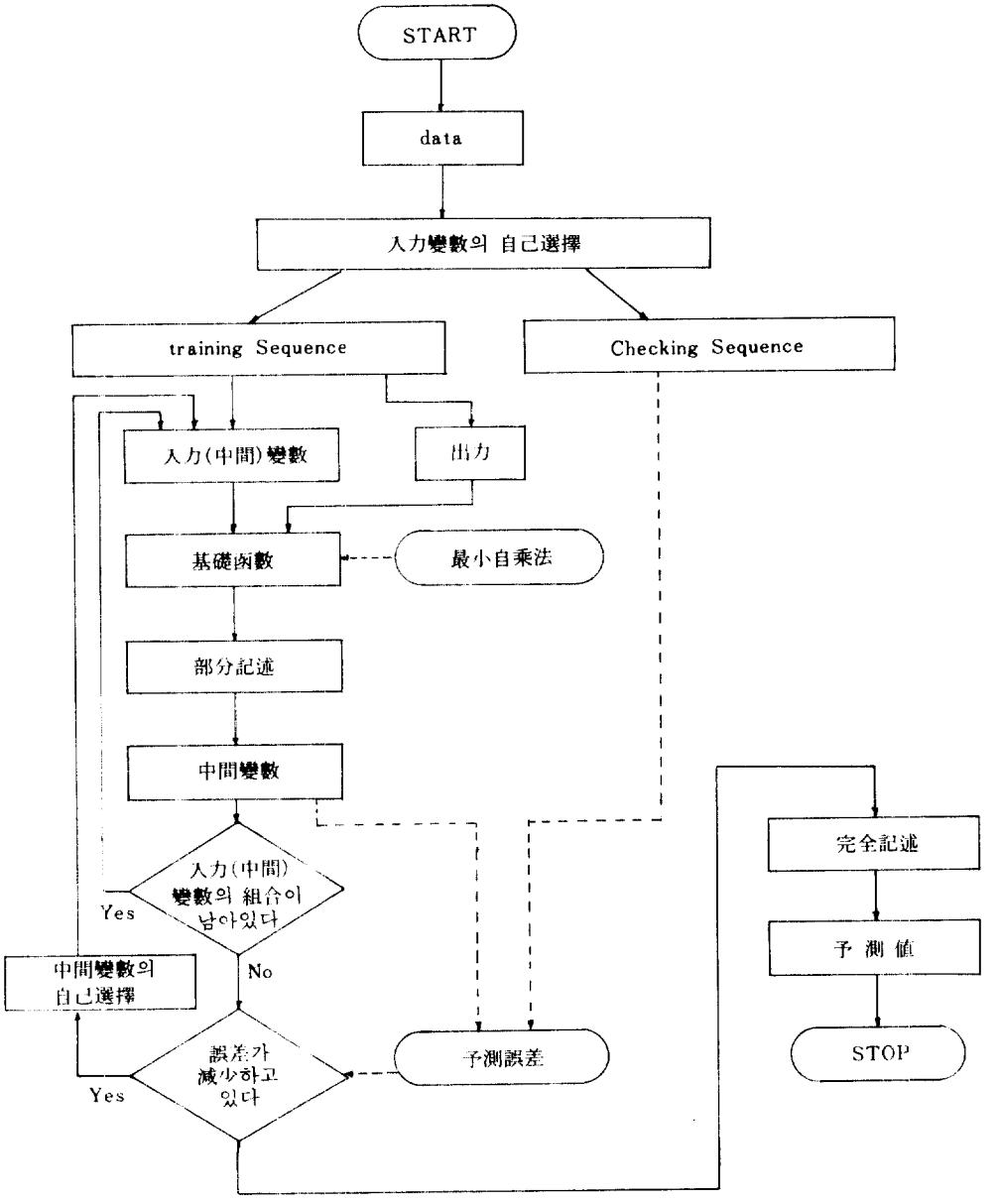


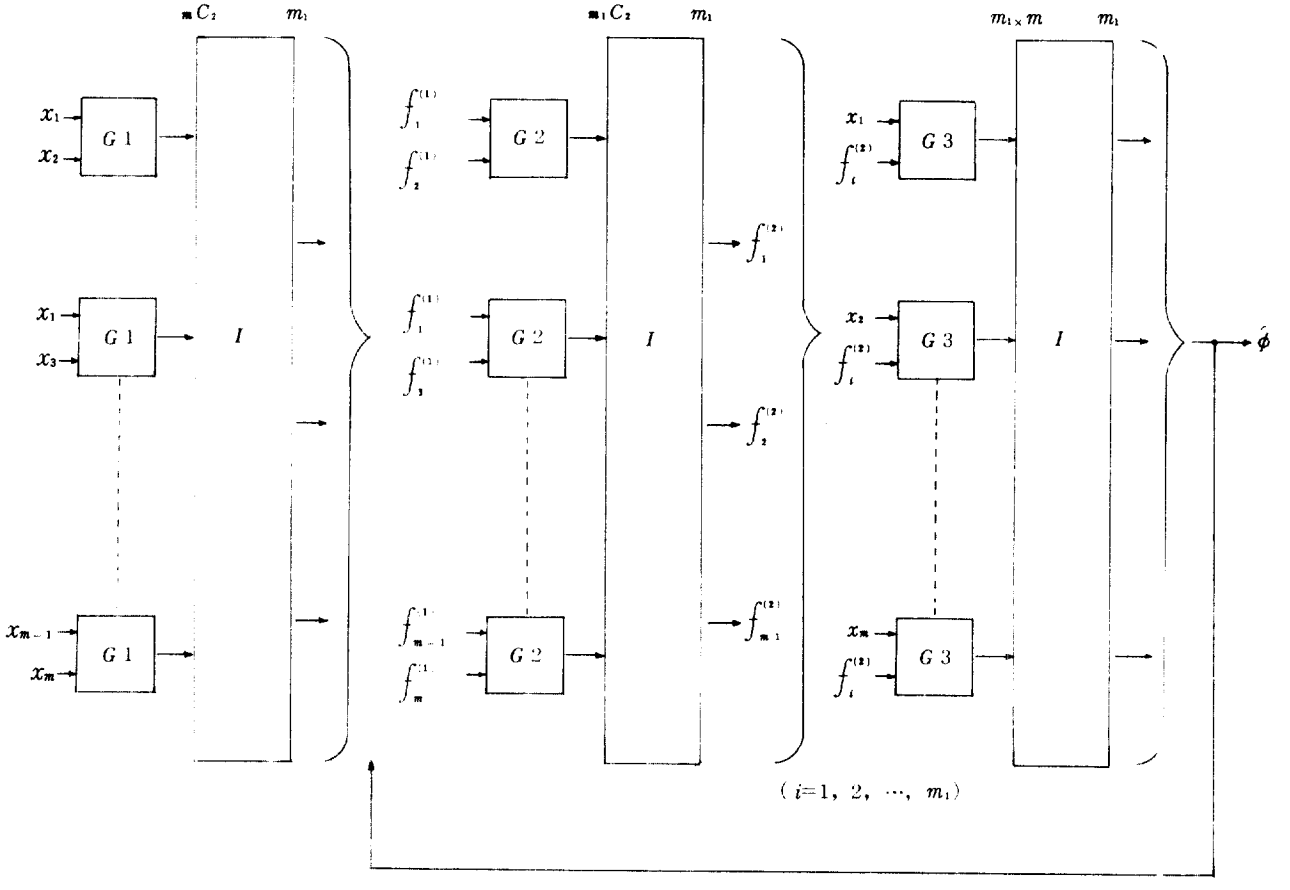
그림 1. 基本的 GMDH의 알고리즘 흐름도

Ⅳ. 中間表現式을 自己選擇하는 改良形 GMDH.

部分表現式 대신에 各選擇層에서 中間表現式을 많이 構成하여 各中間表現式에 대해 전부의 deta를 사용하여 情報量規準(AIC)을 計算하고, 이 값을 最小化하게하는 中間表現式을 自己選擇한다. 그 結果, 종래의 GMDH에서 보이는, 기초data의 分割方法에서 가져오는 Model 構造差異의 發生, Training data와 checking data에 대한 적합의 不均一性을 防止할 수 있으며 기초data에 포함되어 있는 SYSTEM構造에 관해 전부의 情報를 有效하게 使用할 수 있다. 또한 이 GMDH의 알고리즘은 部分表現式을 構成할때 最適인 部分表現式을 自己選擇 하기 위해 基本的GMDH와 比較하여 복잡한 SYSTEM 의 特徵을 간단한 Model에 의해 정확히 나타내는 完全表現式을 構成할 수 있는 性質을 가지고 있다. 이 GMDH의 Block圖를 그림 2에 나타낸다.

여기서 m : 入力變數 x의 個數

m_1 : 最適인 中間表現式의 選擇個數



I : 中間表現式의 自己選擇

$G1$, $G2$, $G3$: 中間表現式의 發生器

그림 2. 改良形 GMDH의 프로·다이아그람

(1) 第1層에 있어서의 알고리즘 ; 第1層에서는 2變數의 P次式에 대하여 AIC를 이용한 變數의 逐次選擇法을 適用하여 最適인 中間表現式을 自己選擇한다.

$$\phi = a_0 + \sum_{i=1}^{2} a_i x_i + \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} a_{i_1 i_2} x_{i_1} x_{i_2} + \cdots\cdots + \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} \sum_{i=1}^{2} \cdots\cdots \sum_{i_p=1}^{2} a_{i_1 i_2 i_3} \cdots x_{i_p} x_{i_1} x_{i_2} x_{i_3} \cdots x_{i_p} \cdots\cdots\cdots\cdots (3)$$

變數의 逐次選擇法으로서 變數增減法을 이용한다.

$$\left[\begin{array}{c|c|c} X^T X & X^T \phi & I \\ \hline \phi^T X & \phi^T \phi & O^T \end{array}\right] \cdots\cdots\cdots\cdots (4)^{註3)}$$

(2) 第2層에 있어서의 알고리즘 ;

$$\hat{\phi} = f_i^{(1)}(X),\ (i = 1, 2, \cdots\cdots, m_1) \cdots\cdots (5)$$

$f_i^{(1)}$ 와 $f_j^{(1)}$ 을 포함시킨 變數에 의해 構成된 式은

$$\hat{\phi} = f_i^{(1)}(x) + f_j^{(1)}(x) \cdots\cdots\cdots (6)$$

$$\left[\begin{array}{cc|c|cc} X_i^{(1)T} X_i^{(1)} & X_i^{(1)T} X_j^{(1)} & X_j^{(1)} \phi & I & O \\ X_j^{(1)T} X_i^{(1)} & X_j^{(1)T} X_j^{(1)} & X_j^{(1)T} \phi & O & I \\ \hline \phi^T X_i^{(1)} & \phi^T X_j^{(1)} & \phi^T \phi & O^T & O^T \end{array}\right] \cdots\cdots\cdots\cdots (7)$$

(3) 第3, 4, … 層에 있어서의 알고리즘 ; 第3, 4, … 層에서는 第2層과 같은 操作을 반복한다. 단, 多層構造의 層에서 거듭되는 경우는 다음과 같은 경우에 終了한다.

(a) AIC의 값이 대단히 적어져 層의 通過에서도 AIC의 값이 減少하지 않는 中間表現式이 얻어질 경우.

(b) AIC의 값이 그렇게 작지 않으나 m_4個의 中間表現式 構造가 前層의 中間表現式 構造와 거의 같이된 경우.

(a)에 의해 終了한 경우에는 noise의 영향을 받지않고 入出力變數間에 存在하는 非線形인 關係를 거의 完全한 形으로 抽出된다는 것을 意味한다.

以上의 알고리즘 (1), (2), (3)에 따라 기초data를 分割하지 않고 전부의 data를 사용하여 AIC을 計算하고 그 값을 最小가 되게, 各選擇層에서 中間表現式을 自己選擇하는 改良形GMDH를 구성할 수 있다.

V. Simulation 개요

i) 說明變數(入力變數) ; 다음式에 기본을 둔다. 또한 說明變數를 必要로 하는 方法에 대해서는 전부 共通으로 사용한다.

$X_1 = 10.0 + 20.0\,t$

$X_2 = 50.0 + 10.0\,t + 1.3\,t^2$

$X_3 = 2.0 + 10.0 \times (1.3)^t$

$X_4 = 500.0 - 400.0 \times (0.8)^t$

$X_5 = 100.0 + 700.0 \times (0.7)^t$

$X_6 = 800.0 / (1.0 + 20.0 e^{-1.1t})$

$X_7 = 100,\ 300, 200,\ 400,\ 100,\ 300,\ 200,$ …….

이 數値에서 誤差는 포함되지 않는 것으로 한다.

ii) 目的變數(出力變數) ; 전부의 方法에 대해 다음의 4가지 case로 한다.

Case 1 : $\phi = a_0 + a_1 x_l + a_2 x_m + a_3 x_n + \varepsilon$

Case 2 : $\phi = a_0 + a_1 x_l + a_2 x_m \cdot x_n + \varepsilon$

Case 3 : $\phi = a_0 + x_l (a_1 x_m + a_2 x_n)^2 + \varepsilon$

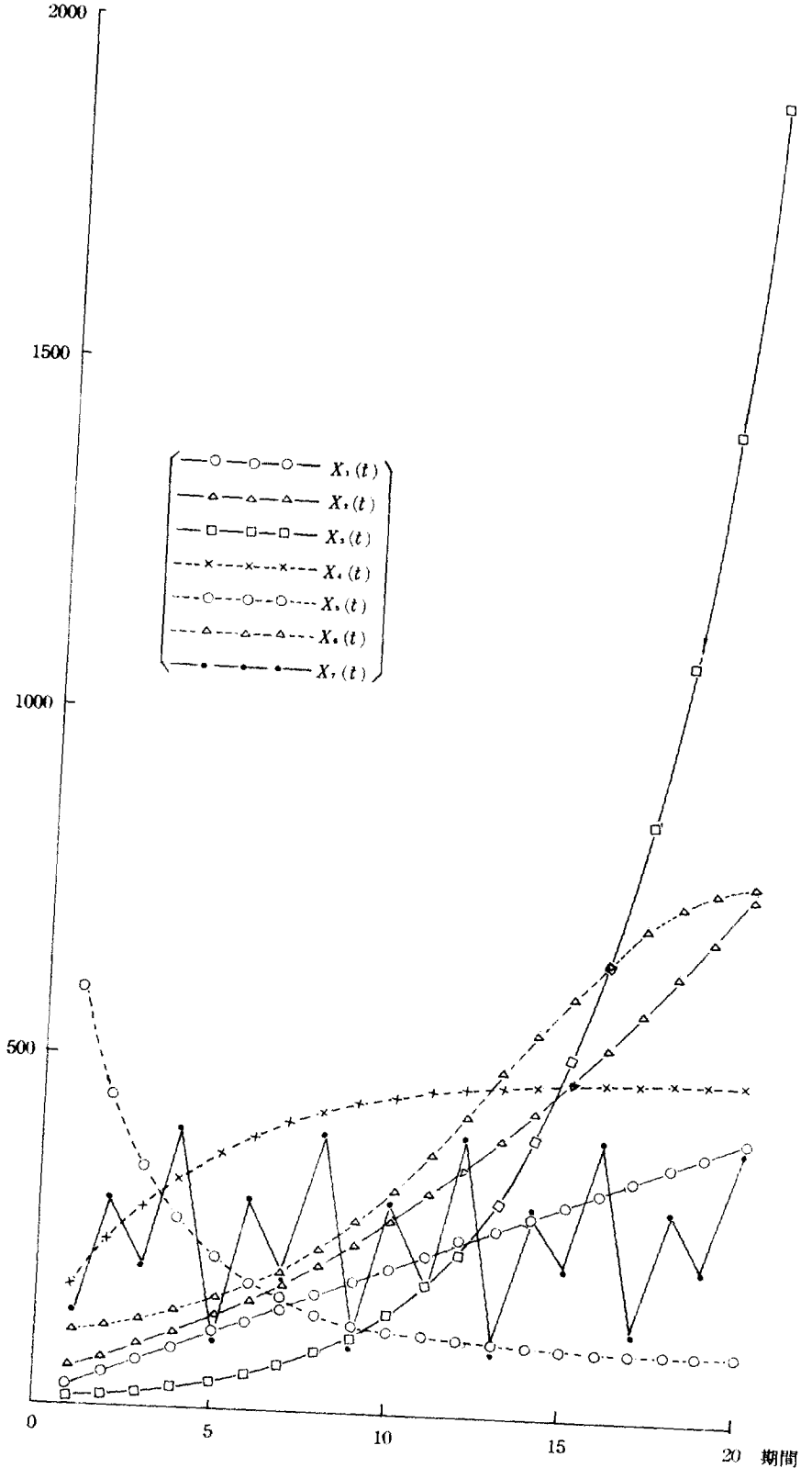
Case 4 : $\phi = a_0 + (a_1 x_l \cdot x_m + a_2 x_n)^2 + \varepsilon$

ϕ : 目的變數, a : parameter

x : 說明變數, ε : 誤差項

說明變數의 變化는 Graph 1에 나타낸다.

註3) (3)式에 対応하는 正規方程式 $(X^T X)\,A = X^T \phi$ 임 단 $A = (a_0, a_1, \cdots\cdots, a_{L-1})^T$



Graph. 1. 說明變數의 變化

表 1 . 目的變數의 構造

Case	說 明 變 數	構 造
1 − 1	X_2, X_4, X_7	$\phi = 1.0X_2 + 2.0X_4 + 3.0X_7 + 80.0 + \varepsilon$
1 − 2	X_1, X_4, X_6	$\phi = 1.0X_1 + 2.0X_4 + 3.0X_6 + 80.0 + \varepsilon$
1 − 3	X_2, X_4, X_7	$\phi = 1.0X_2 + 2.0X_4 + 3.0X_7 + 80.0 + \varepsilon$
2 − 1	X_2, X_4, X_7	$\phi = 2.0X_2 + 0.01X_4 \cdot X_7 + 200.0 + \varepsilon$
2 − 2	X_1, X_4, X_6	$\phi = 2.0X_1 + 0.01X_4 \cdot X_6 + 200.0 + \varepsilon$
2 − 3	X_2, X_4, X_7	$\phi = 2.0X_2 + 0.01X_4 \cdot X_7 + 200.0 + \varepsilon$
3 − 1	X_2, X_4, X_7	$\phi = X_7(0.01X_2 + 0.02X_4)^2 + 100.0 + \varepsilon$
3 − 2	X_1, X_4, X_6	$\phi = X_6(0.01X_1 + 0.02X_4)^2 + 100.0 + \varepsilon$
3 − 3	X_2, X_4, X_7	$\phi = X_7(0.01X_2 + 0.02X_4)^2 + 100.0 + \varepsilon$
4 − 1	X_2, X_4, X_7	$\phi = (0.001X_2 \cdot X_7 + 0.1X_4)^2 + 100.0 + \varepsilon$
4 − 2	X_1, X_4, X_6	$\phi = (0.001X_1 \cdot X_6 + 0.1X_4)^2 + 100.0 + \varepsilon$
4 − 3	X_2, X_4, X_7	$\phi = (0.001X_2 \cdot X_7 + 0.1X_4)^2 + 100.0 + \varepsilon$

(단, ε = 正規乱數)

VI. Simulation 結果 및 考察

表 2 . 各方法의 計算結果

方 法	決 定 變 數					
	Case (1−1)	Case (1−2)	Case (1−3)	Case (2−1)	Case (2−2)	Case (2−3)
G M D H	0.99654	1.00000	1.00000	0.91475	0.99999	0.98116
多元回歸 Model	0.99990	0.99999	0.99999	0.88860	0.99998	0.98040
指數平滑 Model	0.99999	0.99257	0.81596	0.47462	0.99685	0.87383
自己回歸 Model	0.48727	0.99999	0.91338	0.78366	1.00000	0.92509
	Case (3−1)	Case (3−2)	Case (3−3)	Case (4−1)	Case (4−2)	Case (4−3)
G M D H	0.90008	0.99999	0.94321	0.91634	0.99998	0.95783
多元回歸 Model	0.88370	0.99720	0.91710	0.75830	0.95950	0.83060
指數平滑 Model	0.25068	0.99895	0.78098	0.41290	0.99856	0.69354
自己回歸 Model	0.59930	0.99999	0.90260	0.64899	1.00000	0.90473

1) 指數平滑法 : case(1 − 2), (2 − 2), (3 − 2), (4 − 2)는 目的變數가 추세적인 움직임을 보여, 決定係數는 커며 잘 적합되고 있는 것으로 나타난다. 적합성이 좋은 Model은 誤差가 적으며, 적합성이 나쁜 Model에서는 實績値의 變化에 對應되지 않고 급격한 變化에도 對應되지 않는것을 알 수 있다.

2) 多元回歸 Model : Model의 構造가 非線形이면 多元回歸 Model로서는 表現되지 않는다고 할 수 있다. 이는 case 2 이후에서 보는 바와 같이 目的變數가 추세적인 움직임을 보이는것 이외에는 決定係數가 비교적 나쁜것을 볼 수 있다. 說明變數의 單相關行列은 表 3 에서 보여준다.

表3. 說明變數의 單相關 行列

說明變數	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1.0000	0.9844	0.8503	0.8807	0.7821	0.9833	0.1551
X_2		1.0000	0.9226	0.7884	0.6736	0.9920	0.1525
X_3			1.0000	0.5594	0.4464	0.8786	0.1810
X_4				1.0000	0.9799	0.7907	0.1753
X_5					1.0000	0.6705	0.1957
X_6						1.0000	0.1289
X_7							1.0000

3) GMDH : 전부의 case에서 決定係數는 0.9以上으로 되어있다.

相關係數는 0.95以上이다. Model이 線形인 case(1－1), (1－2), (1－3)의 경우를 보면 거의 完璧한 形으로 目的變數가 구축되며, 推定되고 있다. case(1－2)에 대해서도 좋은 결과가 나옴을 볼수있다. Model이 非線形인 경우는 GMDH가 다른 方法에 비해 적합성이 좋고 變數를 自己選擇하고 구축해가는 有効性이 높은 方法이라고 볼수있다. GMDH의 感度分析에서는 data數의 減少에 따른 적합성의 惡化, 構造의 變化는 별로 보이지 않으며 安定한 結果가 나타났다.

參 考 文 獻

1. N.V. Findler and B.B. McCall, (1984), "*Conceptual Framework and a Heuristic Program for the Credit-Assignment Problem.*" IEEE Transaction on SYSTEMS, MAN, AND CYBERNETICS Volume SMC-14.
2. A.G. IvaKhnenko, (1971) "*Polynomial theory of Complex Systems.*" IEEE Transaction on SYSTEMS, MAN, AND CYBERNETICS, SMC-1-4.
3. 田村坦之, 近藤 正, (1978), "モデル選擇の評價基準に予測平方和(PSS)を用いる改良形GMDH", 計測自動制御學會論文集 14－5.
4. 田村坦之, 近藤 正, (1979), "情報量規準AICを 用いて 中間表現式を 自己選擇する改良形GMDH", 計測自動制御學會論文集 15－4.
5. 円山由次郎, (1970), 需要予測と 計量經濟分析, 日本生産性本部.
6. 石渡德彌 訳(1972), I.C.I. MONOGRAPH 2, 短期予測方式, 培風館.
7. 黒田 充 訳(1972), I.C.I. MONOGRAPH 5, 非線形最適化の技法, 培風館.
8. 石渡德彌, (1979), 販賣管理, 丸善(株).
9. 日野幹雄, (1977), スペクトル解析, 朝倉書店.
10. 近藤次郎, (1980), 應用確率論, 日科技連.
11. 池田三郎, 棋木義一, (1975), "GMDHと複雜さ系の同定・予測, 計測と制御 14－2.