集群分析——整數規劃方法

張　志　强*　　姚　興　台**

摘　要

在一般的集群分析方法中，吾人所得到的分組解，通常只是一個局部最佳解，而非整體最佳解。在此，吾人考慮以整數線性規劃方法來獲得最佳之分組結果，本文中，將依照一般集群分析的方法，提出了幾個可行的整數規劃模型，加以探討分析。并在個案研究中，對一般集群分析方法及由整數規劃方法所求得之分組結果做一比較。
關鍵字：集群分析、整數規劃、SAS，LINDO

ABSTRACT

In the general methods of Cluster Analysis, we often only obtain the local optimal solution, not the global optimal solution. Here, we consider to use the Integer Linear Programming approach to obtain the global optimal solution. In this paper, we try to develop several practical Integer Programming models according to the general methods of Cluster Analysis, and give some discussion and analysis. Moreover, in a case study, we compare the results about using the general methods of Cluster Analysis and Integer Programming Approach.
Key Words：Cluster Analysis, Integer Programming, SAS, LINDO

一、前　言

集群分析（Cluster Analysis）是一種邏輯程序，它能依據客體（objects）間的相似性

* 作者爲政大統計研究所博士班研究生
** 作者爲政大統計研究所副教授

與相異性，客觀地使具有相似性的客體歸屬於同一集群（cluster）內。

集群分析所要探討的基本問題主要有三：

1. 我們如何衡量客體間的相似性？
2. 如果我們能衡量每一客體與其他客體之相似性，又如何將相似的客體分派到同一集群內呢？
3. 當集群完成後，對此集群的性質又當如何地描述呢？

其中，關於第二點，一般經常使用的方法有平均連鎖法（Average Linkage method）[8]，華德氏最小變異法（minimum variance method）[14]，K 平均數法（k－means method）[5]……等，這些方法所求出來的分組結果，通常都不是最佳的分組，只是一個局部最佳解（local optimal solution）而非真正的最佳解（global optimal solution）。如果我們只要求一個簡單的分組，則一般方法就可以應付，但是若最佳解是被要求的結果，而又無法將分組情形一一考慮的情況下，則勢必要另闢途徑。[4]

在本文中，吾人將依照一般集群分析的方法，提出了幾個可行的整數規劃模型，加以探討分析，並在個案研究中，對一般集群分析方法及由整數規劃方法所求得之分組結果做一比較。

二、本　　文

在討論整數規劃應用之前，我們首先給予一些符號的定義：

1. N：所有客體的個數。
2. M：所有集群的個數。
3. M_0：在單一集群內最大之客體個數，若無此規定，則 $M_0=N$
4. Pi：第 i 個客體之觀測值。$i=1, 2, \cdots\cdots, N$
5. Nj：第 j 個集群所含之客體個數。$j=1, 2, \cdots\cdots, M$

$$\sum_{J=1}^{M} Nj=N$$

6. C_{ij}：將第 i 個客體納入第 j 個集群所損失之資料程度，在此我們採用

$$C_{ij}=|Pi-Pj| \text{或} (Pi-Pj)^2$$

因爲其爲距離函數，故可知

$$C_{ii}=0，C_{ij}=C_{ji}$$

7.$X_{ij}=\begin{cases}1, & \text{若第 } i \text{ 個客體屬於第 } j \text{ 個集群。} \\ 0, & \text{其他。}\end{cases}$

由上述定義可知〔C_{ij}〕及〔X_{ij}〕均爲 $N \times M$ 矩陣，在此基於某些整數劃模型的需要，將〔C_{ij}〕及〔X_{ij}〕改爲 $N \times N$ 之矩陣，也就是在 N 個集群中，將有 $N-M$ 個集群其不含任何客體。

2.1 在階層集群分析（hierarchical clustering method）上之應用

本節將就階層集群分析中之完全連鎖法及華德氏最小變異法分別用 0－1 整數規劃的模型來表示。

㈠模型一：（最大組內距離極小化）

（Minimize the maximum within group distance）

Min　Z

Subject to

$$d_{ij} X_{ik} + d_{ij} X_{jk} - Z \leq d_{ij}$$

$$i = 1, 2, \ldots, N-1$$

$$j = i+1, i+2, \ldots N$$

$$k = 1, 2, \ldots, M$$

$$\sum_{k=1}^{M} X_{ik} = 1, i = 1, 2, \ldots N \quad \ldots\ldots\ldots (2.1)$$

$$\sum_{i=1}^{N} X_{ik} \geq 1, k = 1, 2, \ldots, M \quad \ldots\ldots (2.2)$$

$$X_{ij} \epsilon \{0, 1\}$$

$$Z \geq 0$$

其中（2.1）式表示每個客體只能屬於一個集群。

（2.2）式表示每個集群內至少含有一固客體。

根據模型一，我們將完全連鎖法轉化成 0－1 整數線性規劃的問題，故可用一般之整數規劃方法在滿足目的函數的情況下，求得最佳之分組解，而且它就是使用完全連鎖法在集群

數等於 M 之解。

但是，值得注意的是線性規劃問題的解題時間會隨著限制式的增加而增加，因此模型一較適用於 N 及 M 均不很大的情況下。

㈡模型二：（組內平方總和極小化）

（Minimize the total within groups sums of square）簡稱 *WGSS* 模型

1. 一維空間的情況

假設 $P_1, P_2, \cdots\cdots, P_N$ 爲 N 個客體之觀察值，如果將此 N 個觀察值按大小排列，令其爲 $q_1,\ q_2, \cdots\cdots,\ q_N,\ q_1 \leqslant q_2 \leqslant \cdots\cdots \leqslant q_N$，則將 $P_1,\ P_2, \cdots\cdots,\ P_N$，$N$ 個客體分成 M 個集群與 $q_1, q_2, \cdots\cdots, q_N$ 分成 M 個集群之意義相同。

因此將分組客體由 $P_1, P_2, \cdots\cdots, P_N$ 轉移至 $q_1, \cdots\cdots, q_N$。

爲了方便分辨每個集群，給予底下定義：

定義 2.1 ：

令第 j 個集群爲 G_j ，而且其內客體之最小值爲 q_j $j=1, 2, \cdots\cdots, N$。

定義 2.2 ：

$$令\quad X_{ij} = \begin{cases} 1, 若\ q_i\ 屬於\ G_j。\\ 0, 其他。\end{cases}$$

則由上述兩定義可知

$$(1) \qquad \sum_{i=1}^{N} \sum_{j>i}^{N} X_{ij} = 0$$

$$(2) \qquad X_{ij} \in \{\, 0\ ,\ 1\, \}$$

又依定義 2.1 可知，第 j 個集群之組內平方和爲

$$WGSS\ (\ G_j\ ) = \sum_{i=1}^{N} X_{ij}\ (q_i - \bar{q}_j)^2$$

其中

$$\bar{q}_j = \sum_{i=1}^{N} X_{ij}\, q_i \,/\, N_j$$

因此綜合以上，我們可設立一個 0－1 整數規劃的 *WGSS* 模型如下：

模型二～①

$$Min \quad \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}\,(q_i-\bar{q}_j)^2$$

$s.t.$

$$q_j=\sum_{i=1}^{N} X_{ij}\,q_i/N_j \quad j=1,2,\ldots N$$

$$\sum_{i=1}^{N} X_{ij}=N_j \qquad j=1,2,\ldots N \qquad (2.3)$$

$$\sum_{j=1}^{N} X_{ij}=M \qquad (2.4)$$

$$\sum_{i=1}^{N}\sum_{j>i}^{N} X_{ij}=0 \qquad (2.5)$$

$$X_{ij}\in\{0,1\}$$

其中（2.3）式代表第 j 個集群含有 Nj 個客體。

（2.4）式代表共有M個集群。

最後兩個限制式由定義 2.1 及 2.2 而來。

但是，仔細觀察，這是一個非線性的整數規劃問題，現考慮所謂成串性質（SP：String Prperty）將之改進爲線性的整數規劃模型。

定義 2.3 成串性質：

對於 X_{ij} 矩陣（$N\times M$）而言，若在對角線以下（below the diagonal）各元素滿足下列條件，則稱 X_{ij} 矩陣具有成串性質。

$$X_{jj}\geqslant X_{j+1\ j}\geqslant\cdots\geqslant X_{N,j} \; for\ j=1,2,\cdots\cdots,N$$

因此根據 SP 性質，我們可以知道，若 q_i， q_{i+2} 均屬於第 j 個集群 Gj，則 q_{i+1} 亦必屬於 Gj。底下以一簡單例子說明 SP 性質。

例 2.1：

若 $N=6, M=3$ 且

$G_1=\{q_1,q_2\}, G_3=\{q_3,q_4,q_5\},$

$G_6=\{q_6\}$

則 X_{ij} 矩陣就具有 SP 性質。

若

$G_1=\{q_1, q_3\}$, $G_2=\{q_2, q_4, q_5\}$,

$G_6=\{q_6\}$

則不具有 SP 性質

【定理 2.1】：

若 X_{ij} * 爲模型＝～①最佳解，則 X_{ij} * 必具有 SP 性質。

證明：

我們可以證明若 X_{ij} * 矩陣之任何行均具有 SP 性質則得證。

考慮第 j 個集群 Gj，令

$G_j{}^* = \{q_j, q_{j+1}, \cdots\cdots, q_{j+l-1}, q_{j+l}\}$

$G_j' = \{q_j, q_{j+1}, \cdots\cdots, q_{j+l-1}, q_{j+l+1}\}$

若 $WGSS(Gj') - WGSS(Gj*) \geqslant 0$，則表示不具有 SP 性質之分組必不是最佳解，故 SP 性質就爲滿足最佳解之必要條件。

$$WGSS(G_j') = \sum_{i=j}^{j+1-1} q^2_{j+l+1} - [\sum_{i=1}^{j+l-1} q_i + q_{j+l+1}]^2 / l + 1$$

$$WGSS(G_j{}^*) = \sum_{i=1}^{j+l-1} q^2_i + q^2_{j+1} - [\sum_{i=j}^{j+l-1} q_i + q_{j+l}]^2 / l + 1$$

$$\Rightarrow WGSS(G_j') - WGSS(G_j{}^*)$$

$$= (q^2_{j+l+1} - q^2_{j+l}) - [2q_{j+l+1} \cdot \sum_{i=j}^{j+l-1} q_i + q^2_{j+l+1}$$

$$-2q_{j+l} \cdot \sum_{i=j}^{j+l-1} q_i - q^2_{j+l}] / l + 1$$

$$= (q^2_{j+l+l} - q^2_{j+l}) - (q_{j+l+1} - q_{j+l})$$

$$(2 \sum_{i=j}^{j+l-1} q_i + q_{j+l+1} + q_{j+l}) / l + 1$$

$$= \frac{1}{l+1} (q_{j+l+1} - q_{j+l}) [l\, q_{j+l+1} - \sum_{i=j}^{j+l-1} q_i)$$

$$+ (l\, q_{j+l} - \sum_{i=j}^{j+l-1} q_i)]$$

已知 $q_j \le q_{j+1} \le \ldots \le q_{j+l-1} \le q_{j+l} \le q_{j+l+1}$

$$\Rightarrow \sum_{i=j}^{j+l-1} q_i \le \sum_{i=j}^{j+l-1} q_{j+l-1} = l\, q_{j+1-1} \le l\, q_{j+l} \le l\, q_{j+l+1}$$

$\therefore$

$$WGSS\,(G_j') - WGSS\,(G_j^*)$$

$$= \frac{1}{l+1}\,(q_{j+l+1} - q_{j+l})\,[(l\, q_{j+l+1} - \sum_{i=j}^{j+l-1} q_i)$$

$$+ (l\, q_{j+l} - \sum_{i=j}^{j+l-1} q_i)] \ge 0$$

因此，若 X_i * 爲最佳解，則 X_{ij} * 矩陣之任何行均具有 SP 性質，故 X_{ij} * 矩陣具有 SP 性質。

接著我們利用 SP 性質來轉換非線性整數規劃模型爲線性整數規劃模型。

【定理 2.2】

若 X_{ij} 矩陣滿足 SP 性質，且 $i>j$，

$$C_{ij} = q_i^2 - (q_j + \ldots + q_i)^2 / (i - j+1)$$

$$+ (q_j + \ldots + q_{i-1})^2 / (i - j)$$

則

$$\sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij}\, C_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij}\, (q_i - \bar{q}_j)^2$$

證明：

我們舊考慮 X_{ij} 之任一行 j 即考慮 G $_j$

則

$$WGSS\,(G_j) = \sum_{l=j}^{j+Nj-1} q_l^2 - \left(\sum_{l=j}^{j+NJ-1} q_l\right)^2 / N_j$$

而

$$\begin{aligned}
C_{jj} &= 0 \\
C_{j+1,j} &= q_{j+1}^2 - (q_j + q_{j+1})^2 / 2 + q_j^2 \\
&= q_{j+1}^2 + q_j^2 - (q_j + q_{j+1})^2 / 2 \\
C_{j+2,j} &= q_{j+2}^2 - (q_j + q_{j+1} + q_{j+2})^2 / 3 + (q_j + q_{j+1})^2 / 2 \\
&\vdots \\
C_{j+Nj-1,j} &= q_{j+Nj-1}^2 - (q_j + q_{j+1} + \ldots + q_{j+Nj-1})^2 / N_j \\
&\quad + (q_i + \ldots + q_{j+Nj-2})^2 / N_{j-1}
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
\sum_{i=1}^{j+NJ-1} C_{ij} &= C_{jj} + C_{j+1,j} + \ldots + C_{j+Nj-1,j} \\
&= q_j^2 + q_{j+1}^2 + \ldots + q_{j+Nj-1}^2 \\
&\quad - (q_j + q_{j+1} + \ldots + q_{j+Nj-1})^2 / N_j \\
&= WGSS\,(G_j)
\end{aligned}$$

$$\therefore \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij} C_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij} (q_i - \bar{q}_j)^2$$

根據以上的論述，我們可以得到一個 0－1 線性整數規劃的 *WGSS* 模型。

模型二～②：

$$Min \quad \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij} C_{ij}$$

s. t.

$$\sum_{i=1}^{N} \sum_{j>i}^{N} X_{ij} = 0$$

$$\sum_{j=1}^{N} X_{ij} = 1 \qquad i = 1, 2, \ldots N$$

$$\sum_{j=1}^{N} X_{jj} = M$$

$$X_{ij} \geq X_{j+1,j} \geq \ldots \quad X_{Nj} \quad j = 1, 2, \ldots N$$

$$X_{ij} \in \{0, 1\} \qquad V_{i,j}$$

其中 C_{ij} 爲定理 2.2 之型式。

最後爲了減少限制式之個數以節省解題時間，給予以下兩題改進：

改進 1：

對於限制式（2.5）可以知道 X_{ij} 矩陣之上三角部分元素均爲 *0*，因此將 C_{ij} 矩陣之上三角部分設定爲極大之正數，如此爲了滿足目的函數的要求，則勢必

$\sum_{i=1}^{N} \sum_{j>i}^{N} X_{ij}=0$，此時限制式（2.5）就可以消去。

改進 2：

若 $M_0 < N$ 爲已知，則每個集群內之客體數不能超過 M_0 個，對 X_{ij} 矩陣而言，第 $N-M_0$ 之後的行不考慮，對於 $N-M_0$ 之前的行，則由於受到 $X_{ij} \geqslant X_{j+l,j} \geqslant \cdots\cdots \geqslant X_{Nj}$ 故有底下的限制式：

$$\sum_{l=0}^{N-M_0-j} \sum_{j=1}^{N-M_0} X_{N-l,j} = 0 \tag{2.6}$$

同樣的，我們亦可以利用改進 *1*，使這些 X_{ij} 所對應的 C_{ij} 爲極大正數，則在 $M_0 < N$ 下，（2.6）式亦可消去。

2.2 多維空間情況下

在多維空間下，我們想知道 SP 性質是否依然有效，但是底下例子説明了 SP 性質在多維空間的情況下，並不成立。

例 2.2

$N=4,\ M=2$，四個觀察值分別爲（1,2）（2,3）（1,1）及（1,0）

則

$$d_{12}=2,\ d_{13}=1.0,\ d_{14}=4$$
$$d_{23}=5.0,\ d_{24}=10,\ d_{34}=1.0$$

對 X_{ij} 矩陣第一行而言，$d_{13}<d_{12}<d_{14}$，則根據 SP 性質 $X_{13}\geqslant X_{21}\geqslant X_{41}$，但經計算後知，其最佳分組爲〔（1,2），（2,3）〕，〔（1,2），（1,0）〕

$\Rightarrow X_{31}=0,\ X_{21}=1$，不滿足 SP 性質，因此 SP 性質在多維空間時，並不成立。

如此一來，在多維空間下，我們將無法獲得一個線性整數規劃模型，由於非線性模型在解題時，會比線性模型來得困難，因此，在此不多討論。

2.3 在非階層集群方法上之應用
（nonhierarchical clustering method）

在一般非階層集群方法中，均希望同一集群內之客體與其集群之種子點（Seed Point）距離越小越好，因此考慮底下的模型：

模型三：

$$Min \qquad Z=\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}\,d_{ij}$$

s. t.

$$\sum_{j=1}^{N} X_{ij}=1 \qquad foi\ i=1,2,\ldots N \tag{2.7}$$

$$\sum_{j=1}^{N} X_{jj}=M \tag{2.8}$$

$$X_{ij}\leq X_{jj} \quad for \quad \begin{array}{l} i=1,2,\ldots N \\ j=1,2,\ldots N \end{array}$$

(2.9)

$$X_{ij} \in \{0, 1\}$$

其中

d_{ij} 表示第 i 個客體與第 j 個客體之距離

限制式（2.7）表示每個客體只能屬於一個集群

限制式（2.8）表示共有 M 個集群

$$X_{jj} = \begin{cases} 1, & \text{若第 } j \text{ 個客體為集群 } G_j \text{ 之種子點。} \\ 0, & \text{其他。} \end{cases}$$

$$X_{ij} = \begin{cases} 1, & \text{若第 } i \text{ 客體屬於 } G_j \text{。} \\ 0, & \text{其他。} \end{cases}$$

因此由 X_{jj} 及 X_{ij} 之定義，可以得到限制式（2.9），但是，我們也發現到（2.9）式之限制式個數為 $N \times N$，在 N 很大時，將會增加解題的時間，所以做底下的改進：

$$X_{ij} \leq X_{jj} \qquad for \quad \begin{matrix} i = 1, 2, \ldots N \\ j = 1, 2, \ldots N \end{matrix}$$

$$\sum_{i=1}^{N} X_{ij} \leq \sum_{i=1}^{N} X_{jj} = NX_{jj}$$

$$\sum_{i=1}^{N} X_{ij} \leq NX_{jj} \ for\ j = 1, 2, \ldots N$$

此時限制式個數由 $N \times N$ 減少為N個。

則模型三經改進後成底下型式：

$$Min \quad Z = \sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij}\, d_{ij}$$

s.t.

$$\sum_{j=1}^{N} X_{ij} = 1 \quad for\ i = 1, 2, \ldots N$$

$$\sum_{j=1}^{N} X_{jj} = M$$

$$\sum_{i=1}^{N} X_{ij} \leq NX_{jj} \quad for\ j = 1, 2, \ldots N$$

$$X_{ij} \in \{0, 1\}$$

三、個案研究

依教育部頒定之「國民中學學生編班實施要點」內容可知，國民中學一年級學生依其入學智力測驗成績採常態分班，自二年級開始，則依據前一年之學習成就，參酌性向趨勢及興趣，對部分學科採學科能力分組，即將智能性向與該學科學習成就相近之學生，予以適當的分組及教學。

因此我們考慮學科成績爲變數，利用集群分析方法及整數規劃方法來求得較佳之分組結果，以做爲編組及教學上之參考。

在此利用 SAS（Statistical Analysis System）[12] 有關集群分析部分之套裝軟體來求得一般集群方法之分組結果，另外利用教育部電算中心，教學研究資訊服務站之套裝程式 LINDO（Linear Interactive and Discrete Optimizer）[3]來求得線性整規劃方法之分組結果，並對兩者作一比較。

我們以基隆市正濱國中二年級某班七十七學年度之學科成績爲分析資料，資料如表 3.1

其中　國文（$x1$）　英文（$x2$）　數學（$x3$）
　　　物理（$x4$）　化學（$x5$）　歷史（$x6$）
　　　地理（$x7$）

表 3.1

OBS	NAME	X1	X2	X3	X4	X5	X6	X7
1	OB111	65	88	83	50	67	90	75
2	OB112	52	84	87	79	77	91	75
3	OB113	55	68	44	82	68	90	66
4	OB114	60	87	74	55	74	97	77
5	OB115	44	57	74	66	70	89	70
6	OB116	62	76	60	61	68	92	75
7	OB117	46	73	50	53	64	82	62
8	OB118	56	65	51	74	57	87	66
9	OB119	38	47	42	46	54	84	62
10	OB120	38	50	56	60	57	85	69
11	OB121	38	76	68	45	61	79	60
12	OB122	17	42	43	72	47	71	60
13	OB123	29	67	47	60	56	66	51
14	OB124	26	81	71	70	57	79	50
15	OB125	17	60	37	50	40	62	48
16	OB126	37	60	50	60	46	71	41
17	OB127	67	75	50	57	61	92	37
18	OB128	74	97	57	46	69	88	90
19	OB129	88	94	60	67	70	89	84
20	OB130	85	98	81	82	74	80	80
21	OB131	64	99	63	78	68	87	84
22	OB132	61	79	92	85	68	86	64
23	OB133	50	73	69	65	66	76	80
24	OB134	75	88	75	68	67	84	60
25	OB135	66	87	61	53	61	82	72
26	OB136	56	90	70	74	66	81	70
27	OB137	38	70	48	46	56	77	60
28	OB138	56	83	66	75	54	77	60
29	OB139	82	88	75	84	78	89	64
30	OB140	89	100	97	85	80	89	90

以數學科爲例，希望將學生予以適當之能力分組，首先以華德氏最小變異法與模型二（ *WGSS* 模型，一維空間）之線性整數規劃法做分組，並比較兩者之結果。

表 3.2 爲華德氏最小變異法所得整個集群分組之過程，首先（OB134, OB139）合併爲同一集群，此時集群數 $M=29$，再將（OB116, OB129）合併爲同一集群，$M=28$……如此，直至 30 位學生均合併爲同一集群爲止。

表 3.2

NUMBER OF CLUSTERS	CLUSTERS	JOINED	FREQUENCY OF NEW CLUSTER
29	OB134	OB139	2
28	OB116	OB129	2
27	OB126	OB127	2
26	OB114	OB115	2
25	OB117	CL27	3
24	OB123	OB137	2
23	OB119	OB122	2
22	OB124	OB136	2
21	OB121	OB133	2
20	OB120	OB128	2
19	CL28	OB135	3
18	CL25	OB118	4
17	CL26	CL29	4
16	OB113	CL23	3
15	OB111	OB130	2
14	CL21	CL22	4
13	OB131	OB138	2
12	CL18	CL24	6
11	OB112	OB132	2
10	CL19	CL20	5
9	CL16	OB125	4
8	CL14	CL13	6
7	CL11	OB140	3
6	CL15	CL17	6
5	CL9	CL12	10
4	CL10	CL8	11
3	CL6	CL7	9
2	CL5	CL4	21
1	CL3	CL2	30

底下考慮 M＝3, 4, 5, 華德氏最小變異法之分組結果爲：

M＝3

集　群	內　　容	組內平方和
1	(OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127, OB118)	187.60
2	(OB120, OB128, OB129, OB116, OB135, OB131, OB138, OB121, OB133, OB136, OB124)	284.18
3	(OB114, OB115, OB134, OB139, OB130, OB111, OB112, OB132, OB140)	578.00

組內平方總和　＝　1,049.78

集　群	平均數	標準差
1	46.20	4.33
2	63.73	5.08
3	82.00	8.01

M＝4：

集　群	內　　容	組內平方和
1	(OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127，OB118)	187.60
2	(OB120, OB128, OB129, OB116, OB135, OB131, OB138, OB121, OB133, OB136, OB124)	284.18
3	(OB114, OB115, OB134, OB139, OB130, OB111)	78.00
4	(OB112, OB132, OB140)	50.00

組內平方總和　＝　599.78

集　群	平均數	標準差
1	46.20	4.33
2	63.73	5.08
3	77.00	3.61
4	92.00	4.08

M＝5：

集　群	內　　　　容	組內平方和
1	（OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127, OB118）	187.60
2	（OB120, OB128, OB129, OB116,OB135）	18.80
3	（OB131, OB138, OB121, OB133, OB136, OB124）	42.83
4	（OB114, OB115, OB134, OB139, OB130, OB111）	78.00
5	（OB112, OB132, OB140）	50.00

組內平方總和　＝　377.23

集　群	平均數	標準差
1	46.20	4.33
2	58.80	1.94
3	67.83	2.67
4	77.00	3.61
5	92.00	4.08

另外以 LINDO 來解模型二可得到以下的結果：

集群數 M	內　　容	平均數	組內平方和
M=3	1.（OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127, OB118, OB120, OB128）	47.92	364.92
	2.（OB129, OB116, OB135, OB131, OB138, OB121, OB133, OB136, OB124, OB114, OB115, OB134, OB139）	68.15	389.69
	3.（OB130, OB111, OB112, OB132, OB140）	88.00	172.00
M=4	1.（OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127, OB118）	46.20	187.60
	2.（OB120, OB128, OB129, OB116, OB135, OB131）	59.50	33.50
	3.（OB138, OB121, OB133, OB136, OB124, OB114, OB115, OB134, OB139）	71.33	88.00
	4.（OB130, OB111, OB112, OB132, OB140）	88.00	172.00
M=5	1.（OB125, OB119, OB122, OB113）	41.50	29.00
	2.（OB137, OB117, OB126, OB127, OB118, OB123）	49.33	11.33
	3.（OB120, OB128, OB129, OB116, OB135, OB131）	59.50	33.50
	4.（OB138, OB121, OB133, OB136, OB124, OB114, OB115, OB134, OB139）	71.33	88.00
	5.（OB130, OB111, OB112, OB132, OB140）	88.00	172.00

最後將兩者的組內平方總和做一比較：

表 3.3

集群數	組內平方總和	
	華德氏最小變異法	模型二目的函數值
M=3	1049.78	926.61
M=4	599.78	481.1
M=5	377.23	333.83

因此由表 3.3可知，以整數規劃所求得之分組結果其組內平方總和在 $M=3, 4, 5$ 時均比華德氏最小變異法所得之組內平方總和爲小，故其分組結果較佳。

接著，我們再比較以 SAS 套裝軟體使用非階層集群方法及模型三所求得的結果。

考慮集群數 $M=5$ 的情況：

表 3.4 爲由 *SAS* 所得之結果：

表 3.5 爲集群平均數及標準差。

表 3.4

NAME	X3	CLUSTER	DISTANCE
OB111	33	4	8.33333
OB112	87	2	5.00000
OB113	44	3	2.50000
OB114	74	4	0.66667
OB115	74	4	0.66667
OB116	60	5	1.37500
OB117	30	1	0.66667
OB118	51	1	1.66667
OB119	42	3	0.50000
OB120	56	5	5.37500
OB121	68	5	6.62500
OB122	43	3	1.50000
OB123	47	1	2.33333
OB124	71	4	3.66667
OB125	37	3	4.50000
OB126	50	1	0.66667
OB127	50	1	0.66667
OB128	57	5	4.37500

OB129	60	5	1.37500
OB130	81	4	6.33333
OB131	63	5	1.62500
OB132	92	2	0.00000
OB133	69	4	5.66667
OB134	75	4	0.33333
OB135	61	5	0.37500
OB136	70	4	4.66667
OB137	48	1	1.33333
OB138	66	5	4.62500
OB139	75	4	0.33333
OB140	97	2	5.00000

表 3.5

CLUSTER MEANS

CLUSTER	X3
1	49.3333
2	92.0000
3	41.5000
4	74.6667
5	61.3750

CLUSTER STANDARD DEVIATIONS

CLUSTER	X3
1	1.50555
2	5.00000
3	3.10913
4	4.71699
5	4.13824

其中表 3.4 之 DISTANCE 表示每個客體與其所屬集群平均數的距離，其距離總和爲 82.76。

另外由模型三所求得之分組結果與由模型二所求得之分組結果相同，此時其目標函數值 Z＝75.0。

值得注意的是，Z 值雖然也代表組內距離總和，但它是每個客體與其所屬集群之中位數（亦爲客體）之距離總和。

因此，由於兩者要求之標準不一樣，此時無法做比較，但若是將模型三所求得之分組結果計算出每個客體與其所屬集群平均數的距離總和，兩者就可以做比較，比較結果如下：

M＝5	非階層集群方法	整數規劃法
組內距離總和	82.76	79.67

故知，此時由整數規劃法所求得之分組結果，在集群數等於 5 時，分組較佳。

同理，我們也可以比較在集群數 M＝3, 4，或其他的情況。

綜合以上，我們可以知道，由整數規劃法所計算出之分組結果通常比一般之集群分析方法所得之結果來得好。

上述是以數學科為例，同樣的，我們可以用相同的方法來得到英文、化學，……等學科之分組結果。

由於教育部並未規定學科能力分組的方法，因此各學校通常是將學生該科成績按高低排列，然後按所需組數將學生均分，而這種方法所得之分組結果必不如由一般集群分析方法及整數規劃法所得之分組結果；以比較組內平方總和而言，考慮將學生均分三組或五組（按成績高低），其結果如下：

M＝3：

集群	內容	平均數	組內平方和
1	（OB125, OB119, OB122, OB113, OB123, OB137, OB117, OB126, OB127, OB118）	46.20	187.60
2	（OB120, OB128, OB129, OB116, OB135, OB131, OB138, OB121, OB133, OB136）	63.00	226.00
3	（OB124, OB114, OB115, OB134, OB139, OB130, OB111, OB112, OB132, OB140）	80.90	686.90

組內平方總和 ＝ 1100.50

M＝5：

集群	內容	平均數	組內平方和
1	（OB125, OB119, OB122, OB113, OB123, OB137）	43.50	77.50
2	（OB117, OB126, OB127, OB118, OB120, OB128）	52.33	53.33
3	（OB129, OB116, OB135, OB131, OB138, OB121）	63.00	56.00
4	（OB133, OB136, OB124, OB114, OB115, OB134）	72.17	30.83
5	（OB139, OB130, OB111, OB112, OB132, OB140）	85.83	312.83

組內平方總和 ＝ 530.49

將其組內平方總和分別與華德氏最小變異法及整數規劃法所得之組內平方總和比較可知，均分方式之組內平方總和高出甚多，因此其分組結果亦較為不好！

另外，若受限於學校設備，師資必須以均分方式來分組，則產生另一缺點即相同成績的學生可能分在不同的組裏，如當 $M=5$ 時，OB134 及 OB139 兩學生成績均為 75，但是 OB 134 在第 4 集群，而 OB139 在第 5 集群內，此種現象在整數規劃法或華德氏最小變異法之分組結果中不可能出現。

在此我們是以一班學生成績為分析資料，實際上我們可以推演到以若干班為一單元，做相同的研究分析，即可先考慮以一般集群分析方法對學生做學科能力分組，若成本許可下（考慮執行程式所佔用之 CPu 時間），以整數規劃法來求得更佳之分組結果，如此則由於各組組內學生程度更趨於一致，組間的差異加大，將可使得教師在選擇教材及教學方式上獲得很大的幫助，另一方面，由於同組學生程度一致，亦可提高學習效果。

四、結論與建議

統計與作業研究之間，有絕大部分是相關聯的* ，在此，我們結合了整數規劃與集群分析的概念，並且以整數規劃法來彌補一般集群分析方法無法求得最佳解的缺陷。

值得一提的是，我們是否就因此而放棄了一般集群分析的方法呢？答案當然是否定的，由個案研究中可以發現，一般集群分析方法方法之結果雖然不如整數規劃法所得之結果，但是兩者相差有限，另外當客體數目衆多或要求之集群數目較多時，有些整數規劃模型就不盡適用，並且執行程式所佔 CPU 時間，也會增長，如此成本負擔增加。因此在不要求最佳分組解時，一般集群分析方法還是很值得運用。

此外，在整篇論文中，只深入探討了幾個線性整數規劃的模型，如果是非線性或其他問題又當如何解決呢？事實上，在作業研究中，並不僅有整數規劃法才能解決集群分析的問題，其他如動態規劃法，次梯度最佳化法（Subgradient optimization method）都是很好的方法，而如何在這些方法中，找尋一個較佳的方法來解決集群分析的問題，亦是一個值得深入探討的課題。

至於個案研究中，雖然只討論一個班級及一維空間下，各種分組方法的比較及結果，但是在考慮多個班級或多維空間時，這些方法依舊適用，另外，對於整數規劃方法所得之集群結果，未深入探討其顯著性的問題，盼爾後能繼續予以研究。

* 註：請見 [7]P.8

參考書目

[1] 林邦傑著，「集群分析及其應用」，教育與心理研究 4 期，台北，國立政治大學出版，民國七十年，頁 31～57。

[2] 黃俊英著，「多變量分析」，三版，中國經濟企業研究所出版，民國七十五年。

[3] 教育部電算中心，「LINDO 使用手册」，教育部電子計算機中心出版，民國七十六年。

[4] Arthanari, T. S., and Dodge, Y., "Mathematical Programming in Statistics", John Wiley, N. Y. 1981

[5] Everitt, B. S., *Cluster Analysis*, Halstead Press, London, 1974

[6] Fisher, W. D., "On Grouping for Maximal Homogeneity", *J. American Stat. Association*, 53, P. 789, 1958

[7] Hiller, F. S. and Liberman, G. L., *Introduction to Operations Research,* 4thed., Holdenday, Inc., San Francisco, 1986

[8] Johnson, R. A. and Wichern, D. W., *Applied Multivariate Statistical Analysis,* Prentice-Hall, Inc., Englewood Cliffs, 1982

[9] Rao, M. R., "Cluster Analysis and Mathematical Programming", *J. Am. Stat. Asso.*, 66, P. 622, 1971

[10] Rustagi, T. S., *Optimizing Methods in Statistics*, Academic Press, London, 1971

[11] Spath, H., *Cluster Analysis Algorithm*, Halsted Press, Chichester, 1980

[12] *SAS User' s Guide：Statistics*, SAS Institute, N. C., 1982 edition

[13] Vinod, H. D., "Integer Programming and Theory of Grouping", *J. Am. Stat. Asso.*, P. 506, 1969

[14] Ward, J. H., Jr., "Hierarchical Grouping to Optimize an Objective Function", *J. Am. Stat. Asso.*, P. 236, 1963