輔仁管理評論 (專刊)
中華民國 91 年 9 月，第九卷第三期，141-162

資料採礦時非均質資料之轉換－以綜合變異數分析為例

江志民*・翁頌舜・梁德馨*****

*台灣大學農藝所
**輔仁大學資訊管理學系
***輔仁大學統計資訊學系

(收稿日期：91 年 4 月 8 日；第一次修正：91 年 6 月 20 日；
接受刊登日期：91 年 7 月 5 日)

摘要

在整個資料採礦 (Data Mining) 的過程中，許多專家認為整套資料採礦的進行有 80%的時間與精力是花費在資料的前置作業階段，其中包含資料的淨化及格式轉換甚或表格的連結。本文即對資料的轉換問題－合併不同次實驗之資料該如何對資料進行合併與轉換做一探討。

本研究以台南區五個馬鈴薯品種塊莖收量為資料，試驗資料分兩年期與兩地區，進行合併分析時先作各獨立試驗之誤差變異數均質性測驗。原始試驗資料經巴氏變異數同質性測驗後，結果不具有同質性，不合乎合併分析的條件。故本研究以同年期不同地區間收量之相關係數作為加權數，得各獨立試驗之試驗誤差變異數傾於同質性；然後再合併各年期及各地區之收量進行分析，結果合併分析之品種間的差異性與各年期獨立試驗的情形有一致的趨勢。因此以不同年期或地區之相關係數作為異質資料之修正加權數，似乎是可行的。

關鍵詞彙：綜合變異數分析，誤差同質性檢定，資料轉換

壹・緒論

在此電腦科技與資料庫技術快速進步的年代，無論是企業或政府機構均擁有更大量的資料，但對於經營、管理及行銷等方面的決策，卻仍然徬徨無助。探究其原因，主要在於無法有效利用所得資料，從中萃取重要資訊。因此，在這資訊爆炸的時代，要如何利用資訊技術來管理及分析所擁有的資料，使其成為有用的資訊，並作為管理單位在進行決策時的參考依據，已成為各決策機構面臨的一大挑戰。

資料採礦 (Data Mining) 是目前資料庫應用的領域中，一項相當熱門的研究議題，其主要目的是從資料庫 (Database)、資料倉儲 (Data Warehouse)、或其他資訊儲存體中的大量資料裡，將有價值的隱藏知識發掘出來的過程，而這些讓人感興趣的知識常使用以樣式 (Patterns)、關聯性 (Associations)、變化

(Changes)、不規則且重要的架構 (Anomalies and Significant Structures) 等方式來呈現 (Han, 1999)。換言之，資料採礦著重的是資料庫的再分析，包括模式的建構或是資料樣式的決定，而其主要目的是用以發現資料庫擁有者先前關心卻未曾知悉的有價值資訊 (Hand, 1998)。有關資料採礦的應用則可概分為分類問題 (Classification)、趨勢分析 (Trend analysis)、分群模式 (Clustering)、關聯分析 (Association) 以及順序型樣 (Sequence Pattern) 等五大類型 (Berry and Linoff, 1997; Fu, 1997)。

正式進行資料採礦前，對於資料的合併、轉換等處理是相當重要的。例如在農業試驗中，為總結特定時間或空間內作物育種或栽培改良之研究，研究者常在該特定空間範圍內的數個地區或在特定時間範圍內在連續數年中進行重複試驗，或綜合以上二種情況下重複進行同一種實驗。當進行資料分析時，則先合併不同次實驗之資料，並以綜合變異數分析估計各參數項目在各地區或各年期平均效應，及進行平均值差異的顯著性測驗。但若不同次試驗的資料之誤差均方不具同質性，則以上分析結果的可靠性不佳；必須對資料作轉換處理，如此分析結果才具意義。

本研究以台南地區五個馬鈴薯品種塊莖收量為資料，以期比較其收量是否有差異。首先以巴氏測驗法 (Bartlett's test) 作原始資料之試驗誤差同質性檢定 (test of homogeneity of error)，結果為不同質，須將資料經適當轉換，再進行綜合變異數分析 (combined analysis of variance)。在此試驗中，各獨立試驗皆為一隨機完全區集設計 (randomized complete block design)，每一處理內有相等的重複次數。轉換結果以相關係數矩陣之加權效果較佳。

貳．文獻回顧

一、資料採礦 (Data Mining)

根據 Cabena et al. (1997) 的定義：資料採礦是將先前不知道，有效的資訊從大型資料庫抽出的過程，並且將萃取出的有用資訊提供給主管做決定性的決策。Berry and Linoff (1997) 則認為資料採礦就是針對大量的資料，利用自動化或半自動的方式進行分析，以尋找出有意義的關係或法則。

而在 Fayyad et al. (1996) 的論文中，則嚴格定義了資料採礦與知識發現 (knowledge discovery in database, KDD) 的不同。基本上，Fayyad et al. (1996) 認為知識發現的整個過程是從理解所要應用的領域開始，經過資料的選取、處

理後、再進行資料轉換以及資料採礦，最後經過探勘結果的解釋與分析後成為有用的知識。這些程序是一種循環的關係，也是一種不斷重複的步驟。換言之，知識發現是一種不間斷的程序，而資料採礦是其中的一個重要步驟。此外，Fayyad et al. (1996) 對資料採礦的定義是依據使用者需求，自資料庫中選擇合適資料，加以處理、轉換，探勘至評估的一連串步驟，其目的在於尋找真實世界運行時隱含於其內的運作現象，並用以輔助解決現實之問題。

根據 Glymour 等人的研究，提出一個進行資料採礦的參考步驟如下：

1.理解資料與進行的工作

2.獲取相關知識與技術 (Acquisition)

3.整合與查核資料 (integration and checking)

4.去除錯誤或不一致的資料 (Data cleaning)

5.發展模式與假設 (Model and hypothesis development)

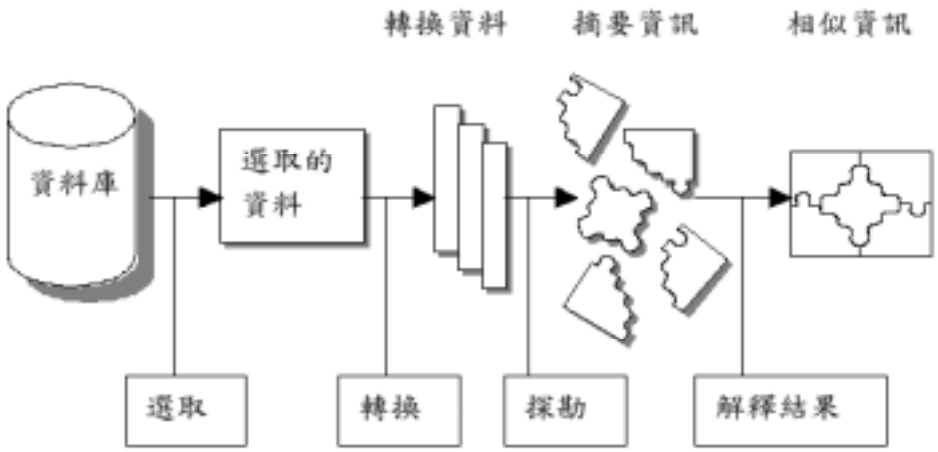
6.實際資料採礦工作

7.測試與檢核所採礦的資料 (Testing and verification)

8.解釋與使用資料 (Interpretation and use)

在進行資料採礦時，首要步驟在選取輸入資料，並指定要進行探勘和分析的對象；之後，再進行資料轉換的工作來降低資料量；待資料轉換完成後，使用者便可執行探勘功能，亦即使用如分類、趨勢分析、關連等資料採礦的相關方法，從轉換後的資料庫中挖掘存在的多種特徵及資訊；最後，再將資料採礦的結果利用文字及圖形呈現。有關資料採礦時的主要流程可以參考圖一 (IBM, 1998)：

根據上述說明，我們可以瞭解，資料採礦主要是從資料或資料庫中，運用相關的分析技術發掘出新的、未知的樣式或規則，並且透過資料採礦的應用，發掘出超越歸納範圍外的資料間關係型態 (Chung and Gray, 1999)。



資料來源：IBM (1998)

圖一　資料採礦流程

二、非均值資料之轉換

在試驗單位未能符合變異數分析的前提時，可嘗試採用適當的轉換法，以期獲得較為可靠的測驗效果。如何選擇適當的轉換法，視各處理均值與變異數關係而定。在下列三種情形下，將資料轉換，並無法改善統計測驗的效果。

1.各處理均值相近，而各處理變異數相異。

2.各處理均值相異，而與各變異數無關。

3.各處理均值相異，而各處理變異相近。

當資料無適當轉換法可用，則應考慮採用無母數統計法。常用的資料轉換法有三種：

(一)開方根轉換 (square root transformation)

試驗資料中各處理均值與變異數成比例關係時宜採用開方根轉換。如X為觀測值，則其開方根$Y=\sqrt{X}$，以轉換後的資料進行綜合變異數分析及處理均值間差異顯著實驗。

(二)對數轉換 (logarithmic transformation)

若各處理均值與其標準差成比例關係時宜採用對數轉換。如X為觀測值，則對數轉換後觀測值為$Y=\log_e X$，若 X 有 0 出現時，當以$Y=\log(X+1)$

或$Y = \log(X + C)$轉換。

(三)角度轉換 (angular transformation)

當試驗資料各處理均值與變異數成比例，且其為二項分布，常用角度轉換。如X為觀測值，則角度轉換後觀測值為$Y = \sin^{-1}\sqrt{X}$ 。

(四)倒數轉換 (reciprocal transformation)

各處理均值的平方與其標準差成比例，則宜採用倒數轉換。如X為觀測值，其轉換式為$Y = 1/X$。

參．研究方法

一、研究資料

本研究以台南地區 1985 及 1986 年馬鈴薯品種產量作比較試驗，有 Lemhi、Norgold、Russet Burbank、White Rose、Cardinal 等五個品種，分六個區集，年期為 1985、1986 年，每年兩期作，其中 1985 年第二期作因遭遇霜害而產量有明顯低落之情形，見表一。

二、研究步驟

首先將實驗所得資料以 Bartlett's test 對資料進行機差均質性試驗，如具有均質性，則進行合併分析；反之，則將原始資料開根號，取對數、倒數後加以分析。

三、研究方法

(一)誤差變方均質性測驗

綜合變異數分析的前提為：(1)機差成分之變異分布為常態性(Normality)。(2)各獨立試驗之機差均方為同質的 (Homogeneous)，可用巴氏測驗法來測。(3)各獨立試驗中品種間變異無顯著差異。

巴氏之卡方測驗法：

$$B = [N \times \log_e(S^2) - \sum_{i=1}^{k} v_i \log_e(S_i^2)]/C \qquad \textbf{(1)}$$

式中 $N=\sum_{i=1}^{k} v_i$

$C=1+\frac{1}{3(k-1)}[\sum\frac{1}{v_i}-\frac{1}{N}]$

$k=4$ (兩年兩期作)

$v_i=20$ S_i之自由度

S_i：各獨立試驗誤差均方，(k 個獨立試驗) $i=1,...,4$

$S=\sum_{i=1}^{k} v_i S_i^{\,2} / N$ 共同均方

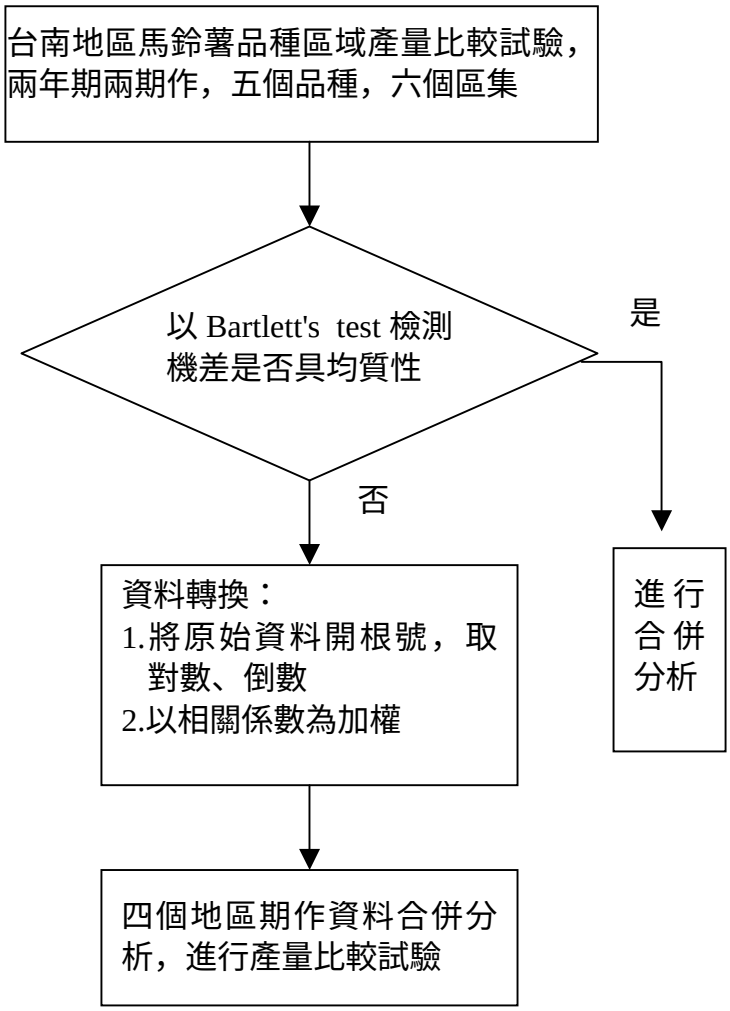
B $x^2{}_{\alpha,(k-1)}$

若 $B < x^2{}_{\alpha,(k-1)}$ ，則接受 H_0 之假設，即表示各獨立試驗機差均方是同質的。

表一 馬鈴薯實驗隨機完全區集設計試驗產量記錄表

年度	期數	品種	區集					
			(1)	(2)	(3)	(4)	(5)	(6)
1985	Ⅰ	Lemhi	16.7000	13.4000	19.0000	20.6000	25.9000	20.5000
		Norgold	15.0000	21.2000	14.3000	15.2000	17.1000	8.9000
		Russet Burbank	19.3000	15.1000	14.7000	13.6000	12.4000	13.7000
		White Rose	17.5000	19.2000	23.4000	29.3000	26.8000	18.3000
		Cardinal	27.3000	29.6000	18.8000	24.5000	22.3000	18.2000
	Ⅱ	Lemhi	9.5100	8.2800	8.0200	8.1600	9.0400	7.3000
		Norgold	6.2200	6.5600	5.5000	5.6000	7.3800	4.7800
		Russet Burbank	8.0000	8.1800	6.4000	6.6600	5.3000	7.6400
		White Rose	14.2400	13.3200	14.0000	12.0800	10.8800	12.6000
		Cardinal	13.9600	12.7800	11.3600	12.4200	11.7800	8.0000
1986	Ⅰ	Lemhi	18.5000	18.3000	19.8000	21.0000	17.2000	17.5000
		Norgold	12.2000	11.1000	11.3000	11.5000	11.2000	12.1000
		Russet Burbank	15.8000	11.7000	14.3000	10.6000	13.7000	10.9000
		White Rose	21.5000	24.2000	21.3000	17.8000	18.2000	20.2000
		Cardinal	5.5000	24.3000	23.5000	18.7000	20.4000	20.3000
	Ⅱ	Lemhi	9.2000	17.2000	10.7000	10.8000	12.5000	16.1000
		Norgold	17.8000	15.1000	17.0000	17.0000	17.7000	17.5000
		Russet Burbank	6.8000	12.6000	12.6000	12.6000	11.9000	10.8000
		White Rose	23.2000	21.0000	21.0000	21.8000	21.3000	24.2000
		Cardinal	19.2000	2.5000	22.5000	20.3000	21.2000	19.3000

Bartlett's test statistic = 5.44 < $\chi^2_{5,.05}$ ∴not significant for α ≦ .05



圖一　研究步驟

(二)綜合變異數分析法

1.合併一次之綜合變異數分析

若採用隨機完全區集設計之單因子試驗，其變值成分數學模式為：

$$x_{ijk} = \mu + \phi_i + B_{(i)j} + \tau_k + (\phi\tau)_{ik} + \varepsilon_{ijk} \qquad \textbf{(2)}$$

式中

x_{ijk} 為第 i 合併項，第 j 區集，第 k 參試項目之變值。

μ 為該樣品資料之族群均值

ϕ_i 為合併項之效應成份

$B_{(i)j}$ 為區集效應成份

τ_k 為參試項目效應成份

$\phi\tau$ 為合併項目與參試項目之交感效應成份

ε_{ijk} 為試驗誤差

2.合併二次之綜合變異數分析

若採用隨機完全區集設計，其變值成分數學模式為：

$$x_{hijk} = \mu + \alpha_h + \phi_i + B_{(hi)j} + \tau_k + (\alpha\tau)_{hk} + (\phi\tau)_{ik} + (\alpha\phi\tau)_{hik} + \varepsilon_{hijk} \qquad \textbf{(3)}$$

式中

μ 為由樣品所由來之族群均值

α_h 為第一合併項之效應

ϕ_i 為第二合併項之效應

$B_{(hi)j}$ 為區集效應

τ_k 為參試項之效應

$(\alpha\tau)$ 為第一合併項與參試項之交感效應

$(\phi\tau)$ 為第二合併項與參試項之交感效應

$(\alpha\phi\tau)$ 為第一、二合併項與參試項之交感效應

ε_{hijk} 為試驗誤差

肆．實証研究

一、誤差同質性檢定

首先以巴氏測驗法 (Bartlett's test) 對各期馬鈴薯進行獨立試驗之誤差變異數均質性檢驗，結果見表二、表三、表四、表五。

表二　1985 年馬鈴薯第一期作之變異數分析表 (ANOVA) 及顯著性檢定

SOURCE	DF	SS	MS	F
VARIETY	4	378.2053	94.5513	5.653654
BLOCK	5	86.0080	17.2016	1.028562
ERROR	20	334.4787	16.7239	
TOTAL	29	798.6920		

** CV=21.4559%

VARIETY		MEAN	DIFFERENCE
Cardinal	(5)	23.4500	
White Rose	(4)	22.4167	1.0333
Lemhi	(1)	19.3500	4.1000 3.0667
Norgold	(2)	15.2833	8.1667 A 7.1333 A 4.0667
Russet Burbank	(3)	14.8000	8.6500 A 7.6167 A 4.5500 0.4833

*‘A’ represent different significantly

表三　1985 年馬鈴薯第二期作之變異數分析表 (ANOVA) 及顯著性檢定

SOURCE	DF	SS	MS	F
VARIETY	4	211.5057	52.8764	41.126793
BLOCK	5	16.2941	3.2588	2.534680
ERROR	20	25.7139	1.2857	
TOTAL	29	253.5136		

**CV=12.3271%

VARIETY		MEAN	DIFFERENCE
Cardinal	(4)	12.8533	
White Rose	(5)	11.7167	1.1367
Lemhi	(1)	8.3850	4.4683 A 3.3317 A
Norgold	(3)	7.0300	5.8233 A 4.6867 A 1.3550
Russet Burbank	(2)	6.0067	6.8467 A 5.7100 A 2.3783 A 1.0233

*‘A’ represent different significantly

表四　1986 年馬鈴薯第一期作之變異數分析表 (ANOVA) 及顯著性檢定

SOURCE	DF	SS	MS	F
VARIETY	4	530.2580	132.5645	43.362828
BLOCK	5	36.1147	7.2229	2.362675
ERROR	20	61.1420	3.0571	
TOTAL	29	627.5147		

**CV=10.1931%

VARIETY		MEAN	DIFFERENCE							
Cardina	(5)	22.1167								
White Rose	(4)	20.5333	1.5833							
Lemhi	(1)	18.7167	3.4000	A	1.8167					
Norgold	(3)	12.8333	9.2833	A	7.7000	A	5.8833	A		
Russet Burbank	(2)	1.5667	10.5500	A	8.9667	A	7.1500	A	1.2667	
*'A' represent different significantly										

表五　1986 年馬鈴薯第二期作之變異數分析表 (ANOVA) 及顯著性檢定

SOURCE	DF	SS	MS	F
VARIETY	4	553.6153	138.4038	29.104358
BLOCK	5	18.9497	3.7899	0.796969
ERROR	20	95.1087	4.7554	
TOTAL	29	667.6737		

**CV=13.2780%

VARIETY		MEAN	DIFFERENCE							
White Rose	(4)	22.1833								
Cardinal	(5)	20.0333	2.1500							
Norgold	(2)	16.4333	5.7500	A	3.6000	A				
Lemhi	(1)	12.7500	9.4333	A	7.2833	A	3.6833	A		
Russet Burbank	(3)	10.7167	11.4667	A	9.3167	A	5.7167	A	2.0333	
*'A' represent different significantly										

Bartlett's homogeneity test：

B=33.5972

P(11.341)=0.01　P(7.815)=0.05　P(9.837)=0.02

由表二至表五，四個變異數分析表，可看出兩年期兩期作之機差為非均質性，故資料不能作合併分析。

二、資料轉換

當試驗資料未能符合綜合變異數分析的前提時，就嘗試適當的轉換法，以期獲得較為可靠的試驗結果。本研究嘗試以開方根轉換、對數轉換及倒數轉換等轉換法，對資料進行轉換，以獲得較可信的結果。

(一)開方根轉換 (Square root transformation)

如X為觀測值，則其開方根$Y=\sqrt{X}$，以轉換後的資料進行綜合變異數分析及處理均值間差異顯著實驗。

表七　1985 年馬鈴薯第一期作開方根後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	5.0219	1.2555	5.842236
BLOCK	5	1.1289	0.2258	1.050625
ERROR	20	4.2979	0.2149	
TOTAL	29	10.4486		

**CV=10.7166%

表八　1985 年馬鈴薯第二期作開方根後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	5.7181	1.4295	40.652257
BLOCK	5	0.4361	0.0872	2.480367
ERROR	20	0.7033	0.0352	
TOTAL	29	6.8576		

** CV=6.2613%

表九　1986 年馬鈴薯第一期作開方根後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	8.1182	2.0295	48.570476
BLOCK	5	0.4905	0.0981	2.347878
ERROR	20	0.8357	0.0418	
TOTAL	29	9.4444		

** CV=4.9815%

表十　1986 年馬鈴薯第二期作開方根後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	8.9048	2.2262	25.928787
BLOCK	5	0.4314	0.0863	1.004969
ERROR	20	1.7172	0.0859	
TOTAL	29	11.0534		

** CV=7.3129%

Bartlett　B=21.0771

P(11.341)=0.01　P(7.815)=0.05　P(9.837)=0.02

由表七至十，四個變異數分析表，可看出兩年期兩期作之機差為非均質性，故資料不能作合併分析。

(二)對數轉換 (Logarithmic transformation)

設原觀測值為 X，對數轉換後觀測值為$Y = \log_e X$。將資料對數轉換後的變異數分析表見表十一至表十四。

表十一　1985 年馬鈴薯第一期作對數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	1.0986	0.2747	5.822287
BLOCK	5	0.2519	0.0504	1.067857
ERROR	20	0.2519	0.0504	1.067857
TOTAL	29	2.2940		

** CV=7.4631%

表十二　1985 年馬鈴薯第二期作對數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	2.5533	0.6383	37.779556
BLOCK	5	0.1966	0.0393	2.327372
ERROR	20	0.3379	0.0169	
TOTAL	29	3.0878		

** CV=5.9949%

表十三　1986 年馬鈴薯第一期作對數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	2.0305	0.5076	51.774984
BLOCK	5	0.1112	0.0222	2.267723
ERROR	20	0.1961	0.0098	
TOTAL	29	2.3377		

** CV=3.5307%

表十四　1986 年馬鈴薯第二期作對數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	2.3735	0.5934	22.208429
BLOCK	5	0.1638	0.0328	1.226242
ERROR	20	0.5344	0.0267	
TOTAL	29	3.0717		

** CV=5.9409%

Bartlett　B=12.7354

P(11.341)=0.01　P(7.815)=0.05　P(9.837)=0.02

由表十一至表十四，四個變異數分析表，可看出兩年期兩期作之機差為非均質性，故資料不能作合併分析。

(三)倒數轉換 (reciprocal transformation)

設原觀測值為 X，倒數轉換之轉換式為 $Y = 1/X$。將資料倒數轉換後之變異數分析結果，見表十五至表十八。

表十五　1985 年馬鈴薯第一期作倒數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	0.0036	0.0009	5.055236
BLOCK	5	0.0010	0.0002	1.083186
ERROR	20	0.0036	0.0002	
TOTAL	29	0.0082		

**CV=23.6124%

表十六　1985 年馬鈴薯第二期作倒數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	0.0352	0.0088	27.756965
BLOCK	5	0.0029	0.0006	1.840458
ERROR	20	0.0063	0.0003	
TOTAL	29	0.0445		

** CV=14.7942%

表十七　1986 年馬鈴薯第一期作倒數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	0.0084	0.0021	49.927660
BLOCK	5	0.0004	0.0001	1.947435
ERROR	20	0.0008	0.0000	
TOTAL	29	0.0097		

** CV=10.3159%

表十八　1986 年馬鈴薯第二期作倒數轉換後之變異數分析表

SOURCE	DF	SS	MS	F
VARIETY	4	0.0118	0.0030	14.605807
BLOCK	5	0.0016	0.0003	1.591058
ERROR	20	0.0041	0.0002	
TOTAL	29	0.0175		

** CV=21.1120%

Bartlett　B=17.3597

P(11.341)=0.01　P(7.815)=0.05　P(9.837)=0.02

由表十五至表十八，四個變異數分析表，可看出兩年期兩期作之機差為非均質性，故資料不能作合併分析。

表七至表十八為我們將原始資料開方根，取對數及倒數後所得之變異數分析表，由此可看出其機差均為非均質性，故我們使用 Cox-Box 所建議之資料轉換並不能解決機差均方異質性的問題故資料仍不能作合併分析。

三、以相關係數加權

本研究資料之原始資料的誤差變異數以及資料經轉換的誤差變異數經巴氏變異數同質性測驗後皆不同質, 故而尋求以同年期不同地區間收量之相關係數 (correlation coefficient) 作為加權數 (weight), 以修正試驗誤差變異數之計算式。

由於作物生長受環境因子, 田間管理及施肥方式之差異而有所改變, 然這些因子對統計分析者而言為未知 (不確定) 之因子, 且其影響程度複雜, 本文試從相關係數來代表不同年同期作中未知因子表現之情形, 用相關係數為一經驗法則, 乃以不同年度同一期作各區集間相關係數作為加權之依據, 原因乃環境因子之改變, 管理施肥方式之差異, 常可由相關係數表現出來。

1985 年第一期作與 1986 年第一期作各區集之相關矩陣為：

$$\mathbf{WW1}=\begin{bmatrix} 0.7910 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4651 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8081 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7371 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7037 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7909 \end{bmatrix}$$

1985 年第二期作與 1986 年第二期作各區集之相關矩陣為：

$$\mathbf{WW2}=\begin{bmatrix} 0.5976 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.7702 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8562 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.6999 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7807 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6190 \end{bmatrix}$$

Y741：1985 年第一期作馬鈴薯產量資料矩陣。

Y742：1985 年第二期作馬鈴薯產量資料矩陣。

Y751：1986 年第一期作馬鈴薯產量資料矩陣。

Y752：1986 年第二期作馬鈴薯產量資料矩陣。

表十九　經轉換後的馬鈴薯產量資料 (Y741ㄨWW1,Y742ㄨINV (WW2))

年度	期數	品種	區集					
			(1)	(2)	(3)	(4)	(5)	(6)
1985	II	Lemhi	15.9136	10.7504	9.3669	11.6588	11.5793	11.7932
		Norgold	10.4082	8.5172	6.4237	8.0011	9.4530	7.7221
		Russet Burbank	13.3868	10.6206	7.4748	9.5156	6.7887	12.3424
		White Rose	23.8286	17.2942	16.3513	17.2596	13.9362	20.3554
		Cardinal	23.3601	16.5930	13.2679	17.7453	15.0890	12.9240
	I	Lemhi	13.2097	6.2323	15.3539	15.1842	18.2258	16.2134
		Norgold	11.8650	9.8601	11.5558	11.2039	12.0332	7.0390
		Russet Burbank	15.2663	7.0230	11.8791	10.0245	8.7258	10.8353
		White Rose	13.8425	8.9299	18.9095	21.5970	18.8592	14.4735
		Cardinal	21.5943	13.7670	15.1922	18.0589	15.6925	14.3944

Note：INV (WW2) means the inverse matrix of WW2

表二十　1985 年馬鈴薯第一期作經相關係數加權轉換後之變異數分析表，Y741*WW1

SOURCE	DF	SS	MS	F
VARIETY	4	194.9210	48.7302	6.364062
BLOCK	5	139.5517	27.9103	3.645030
ERROR	20	153.1420	7.6571	
TOTAL	29	487.6147		

** CV=20.3946%

表二十一　1985 年馬鈴薯第二期作經相關係數加權轉換後之變異數分析表，Y742*INV (WW2)

SOURCE	DF	SS	MS	F
VARIETY	4	420.8383	105.2096	33.708398
BLOCK	5	138.9878	27.7976	8.906141
ERROR	20	62.4234	3.1212	
TOTAL	29	622.2494		

** CV=13.5995%

表二十二　1986 年馬鈴薯第一期作經相關係數加權轉換後之變異數分析表，Y751

SOURCE	DF	SS	MS	F
VARIETY	4	530.2580	132.5645	43.362828
BLOCK	5	36.1147	7.2229	2.362675
ERROR	20	61.1420	3.0571	
TOTAL	29	627.5147		

**CV=10.1931%

表二十三　1986 年馬鈴薯第二期作經相關係數加權轉換後之變異數分析表，Y752

SOURCE	DF	SS	MS	F
VARIETY	4	553.6153	138.4038	29.104358
BLOCK	5	18.9497	3.7899	0.796969
ERROR	20	95.1087	4.7554	
TOTAL	29	667.6737		

**CV=13.2780%

Bartlett's homogeneity test:

Bartlett B=5.7775

P(11.341)=0.01　P(7.815)=0.05　P(9.837)=0.02

由表十九至表二十三之變異數分析表可知其 Bartlett B＝5.778 為不顯著，故資料之機差均方為均質性因為 1985 年第二期發生霜害故以 1／WW2 來作加權，而 1985 年第一期產量偏高而以 WW1 為加權數。1986 年之第一期作及第二期作均使用原始資料。

四、綜合變異數分析

以原始資料加權後，符合各獨立試驗之機差變異數互為同質之資料從事綜合變異數分析，結果如表二十四、表二十五，程式見附錄。

表二十四　加權後的綜合變異數分析表

SOURCE	DF	SS	F	Pr > F
MODEL	39	2446.04531217	13.96	0.0001
LOC	3	414.87206075	30.78	0.0001
BLOCK (LOC)	20	334.31477964	3.72	0.0001
TREAT	4	1397.36742804	77.75	0.0001
LOC*TREAT	12	299.49104374	5.55	0.0001
ERROR	80	359.45794482		
TOTAL	119	2805.50325700		

由表二十四中，區集效應、處理效應、及年期期作與處理之交感效應皆達到顯著水準。

表二十五　加權後的綜合變異數分析表

SOURCE	DF	SS	F	Pr > F
MODEL	39	2446.04531217	13.96	0.0001
YEAR	1	407.76487290	90.75	0.0001
LOC	1	6.79209259	1.51	0.2225
BLOCK (YEAR*LOC)	21	334.62987490	3.55	0.0001
TREAT	4	1397.36742804	77.75	0.0001
YEAR*TREAT	4	47.67015138	2.65	0.0390
LOC*TREAT	4	139.47079918	7.76	0.0001
YEAR*LOC*TREAT	4	112.35009318	6.25	0.0002
ERROR	80	359.45794482		
TOTAL	119	2805.50325700		

由表二十五中，年期、區集效應、處理效應、及年期與處理之交感效應、期作與處理之交感效應、年期與期作與處理之交感效應皆達到顯著水準。

由表二十四及表二十五，可得知處理效應顯著，即表示各馬鈴薯品種塊莖收量有顯著差異，在此以 Duncan 多變域比較法從事品種產量差異顯著性測驗：

表二十六　五種馬鈴薯的顯著差異性檢定

Duncan Grouping	Mean	N	TREAT	
A	19.222	24	(4)	White Rose
A	18.974	24	(5)	Condinal
B	14.346	24	(1)	Lemhi
C	11.899	24	(2)	Norgold
C	11.174	24	(3)	Russet Burbank

註：Means with the same letter are not significantly different.

故依表二十六之結果, White Rose 與 Cardinal 兩種馬鈴薯品種產量較其他三種為高，但沒有足夠證據顯示 White Rose 馬鈴薯品種產量較 Cardinal 多。而 Norgold 與 Russet Burbank 兩種馬鈴薯品種產量較其他三種為低, 但沒有足夠證據顯示 Norgold 與 Russet Burbank 兩種馬鈴薯品種產量有顯著差異。

由分析之結果可知合併分析之品種間之差異性與各年期各期作獨立試驗之情形有一致的趨勢。

伍．結論

由於商業環境不斷快速變遷，企業所面臨的競爭日趨劇烈，激增的市場交易也使得各企業所需儲存與處理的資料量越來越龐大，如何從龐大的資料中, 發掘出對企業有用的資訊, 進而作為企業制定行銷策略、尋找潛在顧客等決策的參考方針, 是一件相當困難但卻有價值的工作。換言之, 產業間共通的資訊僅可作為企業生存的基本需求, 實不足為企業創造競爭優勢, 是以如何從龐大、看似不相關的資料中, 找出潛藏的有價資訊, 是企業目前急需解決之重要課題。而要做到上述的結果, 首先便需有一“乾淨”的資料以做企業分析採礦之用。

在事前的處理資料上常遇到不同期的資料要做一合併的情況，若未考量其間的均質性問題便加以連結, 則可能使得合併後之資料與實際的情況有所落差, 欲解決此一問題, 本研究以台南區五個馬鈴薯品種塊莖收量為資料, 以期比較其收量是否有差異, 試驗資料分有兩年期地區, 進行合併分析時依一般慣用法先作各獨立試驗之誤差變異數均質性測驗, 本研究資料之原始資料之試驗誤差變異數以及資料經轉換 (對數, 開方, 倒數) 的試驗誤差變異數經巴氏變異數同質性測驗後, 不同質, 不合乎合併分析的條件, 故而尋求以同年期不同地區間收量之相關係數作為加權數, 以修正試驗誤差變異數之計算式, 而得各

獨立試驗之試驗誤差變異數而傾於同質性, 然後再合併各年期及各地區之收量進行分析, 結果合併分析之品種間的差異性與各年期獨立試驗的情形有一致的趨勢。因此以不同年期或地區之相關係數作為異質資料之修正加權數, 似乎是可行的。

本研究只是提供這兩個問題一個簡單、適當、且可行的解決方法，未來研究可朝著提出其他更精確、更嚴謹的解決方法研究。

參考文獻

沈明來，「試驗設計學」，台北：九州圖書文物有限公司，第二版，1999 年 3 月。

謝邦昌、沈明來、謝英雄，「常用生物統計分析法之電腦程式檔」，*科學農業*，1989 年。

彭昭英，「SAS 與統計分析」，台北：儒林圖書有限公司，初版，1989 年 9 月。

謝邦昌，「資料採礦入門及運用」，台北：資商訊息顧問股份有限公司，初版，2001 年 4 月。

謝邦昌、柯惠新、盧傳熙，「市場調查與分析技術」，台北：曉園出版社有限公司，初版，2000 年 9 月。

附錄

(一)合併一次之綜合變異數分析

```
OPTION PAGENO=1 LINESIZE=64;
DATA CAOVB;
    INFILE "B:\CAOVB.1";
    INPUT LOC $ TREAT $ BLOCK $ OBS @@;
PROC GLM;
    CLASS LOC TREAT BLOCK;
    MODEL OBS=LOC BLOCK(LOC) TREAT LOC*TREAT/SS1;
    MEANS TREAT/DUNCAN;
RUN;
```

(二)合併二次之綜合變異數分析

```
OPTION PAGENO=1 LINESIZE=64;
```

```
DATA CAOVB;
    INFILE "B:\CAOVB.2";
    INPUT YEAR $ LOC $ TREAT $ BLOCK $ OBS @@;
PROC GLM;
    CLASS YEAR LOC TREAT BLOCK;
    MODEL OBS=YEAR LOC BLOCK(YEAR LOC) TREAT
YEAR*TREAT LOC*TREAT
            YEAR*LOC*TREAT;
    MEANS TREAT/DUNCAN;
RUN;
```

Non-homogeneity Data Transformation in Data Mining - An Example for Combined Analysis of Variance

CHIH-MING CHIANG*, SUNG-SHUN WENG, TE-HSIN LIANG*****

**Institute of Agronomy, National Taiwan University*

*** Information Management Department, Fu-Jen Catholic University*

**** Department of Statistics and Information Science, Fu-Jen Catholic University*

ABSTRACT

For final conclusion of plant breeding and cultural method improvement research, the same experiment where usually replicated within the same space or same time constraint in consecutive years specified in the research. Before the combined analysis of variance where applied for the significant test of mean and the estimation of treatment effect in different zone or years, those replicated data must be checked for homogeneity. If the error variance for different experiment where not homogenous, data must be properly transformed before combining.

Keywords: combined analysis of variance, test of homogeneity of error, data transformation