輔仁管理評論 (專刊)
中華民國 91 年 9 月，第九卷第三期，163-180

資料採礦資料缺值插補之變異數分析

翁頌舜* ・ 梁德馨**

*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 3 月 20 日；第一次修正：91 年 6 月 13 日；
接受刊登日期：91 年 7 月 6 日)

摘要

隨著資訊科技的進步以及電子化時代的來臨，現今企業所面對的是一個與以往截然不同的競爭環境。在資訊科技的推波助瀾下，不僅企業競爭的強度與速度倍數於以往，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。在這種情況下，企業的焦點已從以往的資料蒐集與整理，轉變成如何有效的利用資料庫來進行資訊的獲取。換言之，企業如何因應外界的競爭，能快速且有效的從資料庫中取得有用的資訊，並反應市場或消費者的需求，已成為各企業急於解決的重要議題之一。

在整個資料採礦 (Data Mining) 的過程中，牽涉了大量的準備工作與規劃過程，事實上許多專家皆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與格式轉換甚或表格的連結。在資料蒐集上常因為人為或不可抗拒因素造成資料缺值，此時如不加以處理，將因為不完整的資料庫而導致資料採礦的結果失真。因此本文對資料淨化過程常遇到的問題－缺失資料的處理方面做一探討，並以農業試驗中的馬鈴薯品種塊莖收量資料來加以說明。

關鍵詞彙：缺值，EM 遞迴

壹・緒論

由於企業競爭的全球化與白熱化，以及資訊科技與管理技術的一日千里，對全球各地的產業均帶來了巨大的衝擊。企業的競爭優勢已不再單純的建構在豐厚的資金或是龐大的土地上，而是取決於企業是否能善用所擁有的資訊，使其成為企業競爭的利器。而在電腦科技與資料庫技術快速進步的今日，企業雖較以往擁有了更大量的資料，然而對於經營或是行銷等方面的決策，卻仍然徬徨無助。深究其原因，我們不難發現癥結仍在於資訊的利用不足。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，並作為企業在進行決策時的參考依據，已成為現代企業所必須重視的課題之一。

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)。

而在做資料採礦之前，資料的“淨化 (Cleaning)”是相當重要的。所謂資料淨化 (Data Cleaning), 包含了資料整理與處理資料中不符定義的數值 (例如缺值或年齡為負值等)，採用淨化後的資料再依專業知識做合理性的判斷 (判斷是否在淨化後樣本會偏離母體的情況)，再以此資料做分析，才能得出更嚴謹及正確的結果。

在進行資料蒐集時，常因為人為因素或其它不可抗拒的原因造成資料蒐集不完整，使測量變項產生缺失值。而在農業試驗中，常因天災蟲禍而導致資料蒐集的不完整，利用不完整資料逕行試驗設計之變異數分析，將會得到不可靠之結果。 Allan & Wishart (1930) 首先提出單一缺值估計法，插補單一不完整資料之觀念； Yates (1933) 將其方法改良後，利用解聯立方程式技巧，推廣至數個缺值之插補; Laird & Rubin (1977) 利用最大概似估計發展出‘EM’遞迴計算缺值方法。本研究採用‘EM’遞迴缺值估算法，以先估算缺值部份再進行一般變異數分析的方式來比較台南地區五個馬鈴薯品種塊莖收量是否有差異。但若在兩年期六地區的資料中有一個或數個試驗單位之資料缺失 (missing data) 則無完整資料做變異數分析，必須以所得的資料估算缺值，然後再以一般的變異數分析法 (analysis of variance) 進行資料分析。

貳．文獻回顧

一、資料採礦 (Data Mining)

為因應商業環境的快速變遷以及資訊科技的發展，許多企業開始引進資訊技術，期望透過資訊科技的應用為企業帶來競爭的優勢。但隨著時間的累積，各企業的資料儲存量也隨之增加，在這許多的龐大資料群當中，常隱藏著許多超出我們直覺可以想像的有用資訊，雖然我們通常會使用傳統的資料查詢

和統計功能來進行資料的解析。但儘管如此，似乎也並不容易找到這些未知的資訊特徵以及關係。因此，如何透過特定的程序以及方法，幫助我們從大量的資料中萃取出以往未知的有用知識與資訊，以提供企業決策者進行決策的參考依據，是相當重要也刻不容緩的事。而資料採礦是目前資料庫應用領域中相當熱門的也重要的一個技術，其目的就是從大量資料中，尋找出事前未知、且有效可以付諸行動的規則或知識 (Cabena et al., 1997; Fayyad et al., 1996; Chen et al., 1996)。

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)

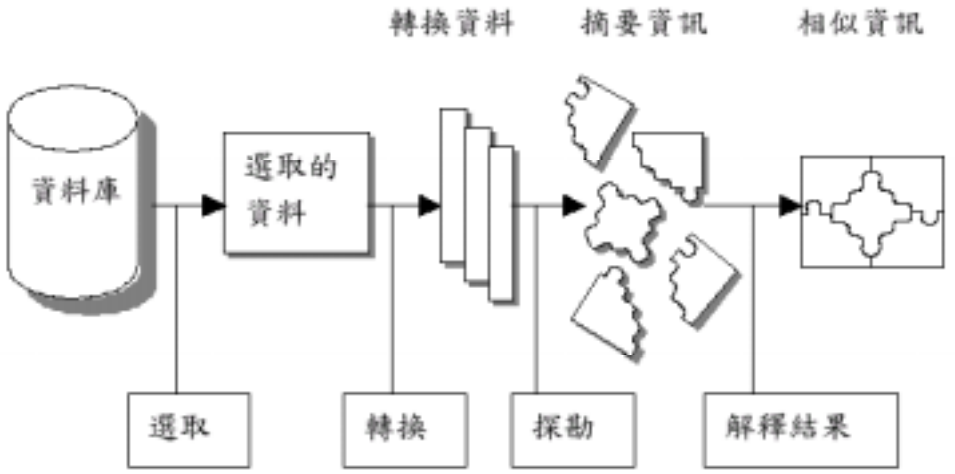
6.實際資料採礦工作

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)

因此，我們可以瞭解在進行資料採礦時，其主要運作流程是先選取輸入

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)：



資料來源：IBM (1998)

圖一　資料採礦流程

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)。

二、資料缺值之處理

在許多情況下，小量的缺失值是可以容忍的。但是如果缺值的比例超過了 10%，就可能出現嚴重的問題。處理缺失值主要有四種方法：

(一)用一個樣本統計量的值去代替缺失值

缺失值可以用一個樣本統計量去代替，最典型的作法是使用變數的平均值。如此，由於該變數的平均值會保持不變，那麼其他的統計量例如標準差和相關係數等也不會受很大的影響。以市場調查的情況為例，若一個被訪者沒有回答其收入，那麼就用整個樣本的平均收入、或用該被訪者所在的子樣本 (比

方說是屬於社會地位比較高的那個階層) 的平均收入去代替。不過從邏輯上來說, 這樣做是有問題的, 因為被訪者如果回答了該問題的話, 其答案可能是高於或低於該平均值的。

(二)用從一個統計模型計算出來的值去代替缺失值。

另一種缺失值的處理方法就是利用某些統計模型計算所得比較合理的值來代替, 例如利用迴歸模型、判別分析模型等等。比方說, “ 產品的使用程度 ” 可能與 “ 家庭規模 ” 和 “ 家庭收入 ” 有聯繫, 利用回答了這三個問答題的被訪者的資料, 可能構造出一個迴歸方程式。對於某個沒有回答 “ 產品的使用程度 ” 的被訪者, 只要其 “ 家庭規模 ” 和 “ 家庭收入 ” 是知道的, 就可以通過這個迴歸方程式計算出其 “ 產品的使用程度 ”。又如在選舉預測中, 如果問到下次選舉中會投誰的票時, 有許多被訪者常常會給出 “ 還沒有決定 ” 的回答。如果只是簡單地刪除這一部份的回答 (有時可能高達 30%左右), 那麼肯定會引起嚴重的預測偏差。處理這一問題的統計方法之一是尋找一個判別函數, 使其能夠區別那些已經決定投票選 A (假定只有兩個候選人 A 和 B) 的群體和已經決定選 B 的群體。這個函數可能由一些獨立變數來解釋, 比如被訪者的社會地位、職業、黨派、教育程度、生活形態等等。假定某位說 “ 還沒有決定 ” 的被訪者給出了上述變數的答案, 那麼就可能通過計算將他 (她) 劃入 “ 已經決定選 A ” 或 “ 已經決定選 B ” 的群體中。這樣, 選舉預測的成功率就會大大地提高。

(三)將有缺失值的資料整個刪除 (list-wise)

將有缺失值的個案整個刪除 (list-wise) 的方法, 結果可能會導致很小的樣本, 因為很多資料 (包含多個變數) 都會有一些變數是缺少的。刪除大量資料並不是所希望的, 因為資料的蒐集是需要大量時間、人力和經費。而且, 有缺失的資料與完整的資料之間可能會有顯著的差異, 若真是如此, 則整個刪除的 List-wise 方法會導致嚴重偏差的結果。

(四)將有缺失值的個案保留, 僅在相對應的分析中作必要的排除

將有缺失值的個案保留, 僅在相對應的分析中作必要的排除 (paire-wise) 的方法, 會使分析中不同的計算將根據不同的樣本量進行, 這也有可能導致不適宜的結果。但是在實際執行中這種方法常被研究人員所採用, 因此如果能滿足以下三個條件, 這種方法是妥當的：

1.樣本量很大

2.缺失值很少

3.變數之間不是高度相關的

以上介紹了四種處理資料缺值的方法，而不同的缺失值處理方法可能產生不同的結果，特別是當缺失值的出現不是隨機及變數之間存在高度相關的情況。因此，應當使缺失資料保持在最低的水準。在選擇一種處理缺失資料的方法之前，研究人員應該仔細地考慮各種方法所可能產生的後果。

參．研究方法

一、研究資料

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

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、缺失值的估算

(一)當一個或數個實驗單位的試驗資料產生缺值

隨機完全區集設計所得觀測值之組成份有總平均值、處理效應、區集 (重複) 效應及試驗誤差等四種，其數學模式為：

$$x_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad \textbf{i=1,2,...,m} \qquad \textbf{j=1,2,...,n}$$

式中 x_{ij} 為試驗或觀察所得觀測值,

μ 為族群均值,

τ_i 為第 i 處理效應且 $\sum_{i=1}^{m} \tau_i = 0$,

β_j 為第 j 區集效應,

ε_{ij} 為第 i 處理 j 區集之試驗誤差。

利用最小平方法 (least squares method)，使誤差成份為最小，以求各成份之估式：

令 $Q = \sum^{m}\sum^{n}(x_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2$ 今以 Q 分別對 $\hat{\mu}$，$\hat{\tau}_i$，$\hat{\beta}_j$ 偏微分，得下式：

1.對 μ 估算：

$$\frac{dQ}{d\bar{\mu}} = -2\sum^{m}\sum^{n}(x_{ij} - \bar{\mu} - \bar{\tau}_i - \bar{\beta}_j) = 0$$

$$mn\bar{\mu} = \sum^{m}\sum^{n} x_{ij}$$

$$\bar{\mu} = \sum^{m}\sum^{n} x_{ij} / mn = \bar{x}..$$

$\hat{\mu}$ 為試驗資料的總平均值，以 $\bar{x}..$ 代表之。

2.對 τ_i 估算：

$$\frac{dQ}{d\hat{\tau}_1} = -2\sum^{n}(x_{ij} - \hat{\mu} - \hat{\tau}_1 - \hat{\beta}_j) = 0$$

$$n\hat{\tau}_1 = \sum^{n} x_{1j} - n\hat{\mu}$$

$$\hat{\tau}_1 = \sum^{n} x_{1j} / n - \hat{\mu}$$

$$= x_{1.} / n - \hat{\mu} = \bar{x}_{1.} - \bar{x}.. = t_1$$

同理可得第 i 處理效應之估式，以 t_i 代表之

$$\hat{\tau}_i = \bar{x}_{i.} - \bar{x}.. = t_i$$

3.對 β_j 估算：

$$\frac{dQ}{d\hat{\beta}_1} = -2\sum^{m}(x_{i1} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_1) = 0$$

$$m\hat{\beta}_1 = \sum^{m} x_{i1} - m\hat{\mu}$$

$$\hat{\beta}_1 = \sum^{m} x_{i1} / m - \hat{\mu}$$

$$= x_{.1} / m - \hat{\mu} = \bar{x}_{.1} - \bar{x}..$$

同理可得第 j 區集效應之估式，以 b_j 代表之

$$\hat{\beta}_j = \bar{x}._j - \bar{x}.. = b_j$$

所以 $x_{ij} = \mu + \tau_i + \beta_j$ 的估式為

$$\hat{x}_{ij} = \bar{x}.. + t_i + b_j$$

$$= \bar{x}.. + (\bar{x}_{i.} - \bar{x}..) + (\bar{x}._j - \bar{x}..)$$

當資料出現缺值時，以其他觀測值所估計的總平均值、處理效應及區集效應，並採用代疊的方式，再以此完整的資料估計總平均值、處理效應及區集效應，再代入原本缺值的資料，重複相同程序，直到前後兩次的估計值差異較小時停止。

(二)當資料缺值為整組區集資料時

仍以 $x_{ij} = \bar{x}.. + t_i + b_j$ 估計缺值，但此時採用其他各組區集和的平均為此組資料的區集和。

肆．實證研究

以台南地區 1985 年第 I 期作馬鈴薯品種塊莖收量為例，該獨立試驗為一隨機完全區集試驗。當缺值產生時，本研究嘗試以 SAS 統計套裝軟體估計各缺值。

一、當試驗資料有一個或數個缺值時

(一)當品種二區集五為缺值：

以總平均值、第二處理效應及第五區集效應估計該缺值，其結果為：

估計值為 17.1538

估計後 SSE 為 334.478

(二)當品種二區集五 (缺值1) 和品種四區集三 (缺值2) 為缺值：

分別以總平均值、第二處理效應及第五區集效應估計缺值(1)，總平均值、第四處理效應及第三區集效應估計缺值(2)，

其結果如下：

估計值(1)為 17.1538

估計值(2)為 20.3632

估計後 SSE 為 328.436

估計品種四區集三 (缺值 2) 較估計品種二區集五 (缺值 1)，對試驗誤差平方和 (SSE) 有較大的影響。就原始資料來看，缺值 1 的實際值和以各平均值所得出的估計值較為接近，而缺值 2 的實際值則對各平均值有較大的影響力，估計值和實際值有較大的差距，故對試驗誤差平方和影響較大。 估計方

法方面, 因採用數學上使誤差成份為最小的最小平方法, 估計缺值後的試驗誤差平方和皆不會大於原始資料之試驗誤差平方和。

二、當整組區集資料為缺值時

當區集三的整組資料均遺失時可求得

估計值 (品種 1) 為 19.42

估計值 (品種 2) 為 15.48

估計值 (品種 3) 為 14.82

估計值 (品種 4) 為 22.22

估計值 (品種 5) 為 24.38

估計後 SSE 為 312.294

此組資料的區集和和其他各組區集和的平均較為接近, 故而估計值和原來資料較為接近。

三、估計前後品種間的差異性比較

1985 年第 I 期原始資料之處理均值比較試驗, 處理項 p-value= 0.0033, 表示品種間產量有顯著差異。而由 Duncan's multiple range 檢定結果 (見表二、表三、表四、表五), 可看出品種 D 及 E 產量明顯高於品種 B 及 C 但 B 與 C 間、D 與 E 間並無明顯差異。此結果與原資料之分析相符合。

表二 五個處理 Duncan 多變域顯著差異值表

Duncan Grouping	Mean	N	TREAT
A	23.450	6	E
A	22.417	6	D
B A	19.350	6	A
B	15.283	6	B
B	14.800	6	C

**Means with the same letter are not significantly different

(一)當品種二區集五為缺值

缺值估算後之處理均值比較試驗, 處理項 $F-value = 5.37 > F_{0.05,4,19}$，表示品種間產量有顯著差異。

表三　五個處理 Duncan 多變域顯著差異值表

Duncan Grouping	Mean	N	TREAT
A	23.450	6	E
A	22.417	6	D
B A	19.350	6	A
B	15.292	6	B
B	14.800	6	C

**Means with the same letter are not significantly different

(二)當品種二區集五 (缺值1) 和品種四區集三 (缺值2) 缺值

缺值估算後之處理均值比較試驗, 處理項 $F-value=4.90>F_{0.05,4,18}$，表示品種間產量有顯著差異。

表四　五個處理 Duncan 多變域顯著差異值表

Duncan Grouping	Mean	N	TREAT
A	23.450	6	E
A	21.911	6	D
B A	19.350	6	A
B	15.318	6	B
B	14.800	6	C

**Means with the same letter are not significantly different

(三)當區集三整組資料缺值時

缺值估算後之處理均值比較試驗, 處理項 $F-value=4.97>F_{0.05,4,15}$ 表示品種間產量有顯著差異。

表五　五個處理 Duncan 多變域顯著差異值表

Duncan Grouping	Mean	N	TREAT
A	24.380	6	E
A	22.220	6	D
B A	19.420	6	A
B	15.480	6	B
B	14.820	6	C

**Means with the same letter are not significantly different

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。

本研究所提之資料缺值插補模式，主要的目的便是希望透過所提之插補模式，建構出更完整之資料庫，以降低缺失值對於研究分析所產生的各種影響。而試驗中各獨立試驗皆為一隨機完全區集設計，採用隨機完全區集設計唯一的限制是每一處理必須有相等的重複次數，如果在試驗終了，有一個或數個試驗單位之資料遺失或失敗，試驗結果沒有完整的資料以便進行變異數分析，實在可惜，故而採取補救的辦法，以所得的試驗資料估算缺值，然後再以一般的變異數分析法進行資料分析。在此以所得資料之總平均值、處理效應、區集效應估算缺值，估算方法採最小平方法估算總平均值、處理效應、區集效應。

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，利用此法以台南地區五個馬鈴薯品種塊莖收成量做一簡單分析比較。未來研究可朝著提出其他更精確、更嚴謹的解決方法研究。或著可探討如自由度的減少等，當資料遺失時所產生的問題。

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。

Allan, F.E. & Wishart, J, "A Method of Estimating the Yield of Missing plot in field Experimental Work", J. Agric Sci. 20, 1930, pp.399-406.

A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. R. Statist. Soc. B 39, 1977, pp.1-38.

Yates, F., "The Analysis of Replicated Experiments when the field Results are Incomplete", Emp. J. Exp. Agric. 1, 1933, pp.129-142.

附錄：缺值估計之電腦程式

```
OPTIONS NODATE;
DATA ARR03;
    INFILE "B:\RCBD";
    INPUT P1-P6 Q1-Q6 R1-R6 S1-S6 Z1-Z6;
    ARRAY A{5} A1-A5;
    ARRAY B{6} B1-B6;
    ARRAY C{5,6} P1-P6 Q1-Q6 R1-R6 S1-S6 Z1-Z6;
    ARRAY D{5,6} D1-D30;
    ARRAY E{5,6} E1-E30;
    ARRAY F{5,6} F1-F30;
    ARRAY V{5} V1-V5;
    ARRAY W{6} W1-W6;
    ARRAY X{5} X1-X5;
    ARRAY Y{6} Y1-Y6;
    M=5;
    N=6;
    DO I=1 TO M;
      DO J=1 TO N;
        D{I,J}=0;
        E{I,J}=0;
      END;
    END;
    L=0;
    DO I=1 TO M;
      X{I}=N;
      DO J=1 TO N;
        IF C{I,J}=. THEN X{I}=X{I}-1;
```

```
    END;
END;
DO J=1 TO N;
    Y{J}=M;
    DO I=1 TO M;
        IF C{I,J}=. THEN Y{J}=Y{J}-1;
    END;
END;
DO I=1 TO M;
    DO J=1 TO N;
        IF C{I,J}=. THEN L=L+1;
        IF C{I,J}=. THEN C{I,J}=0;
    END;
END;
DO I=1 TO M;
    V{I}=0;
    DO J=1 TO N;
        U1=U1+C{I,J};
        U2=U2+C{I,J}*C{I,J};
        V{I}=V{I}+C{I,J};
    END;
    U3=U3+(V{I}*V{I})/X{I};
END;
DO J=1 TO N;
    W{J}=0;
    DO I=1 TO M;
        W{J}=W{J}+C{I,J};
    END;
    U4=U4+(W{J}*W{J})/Y{J};
END;
T=0;
DO I=1 TO M;
    A{I}=0;
```

```
    DO J=1 TO N;
      T=T+C{I,J};
      A{I}=A{I}+C{I,J};
    END;
    A{I}=A{I}/X{I};
END;
DO J=1 TO N;
    B{J}=0;
    DO I=1 TO M;
      B{J}=B{J}+C{I,J};
    END;
    B{J}=B{J}/Y{J};
END;
T=T/(M*N-L);
DO I=1 TO M;
    A{I}=A{I}-T;
END;
DO J=1 TO N;
    B{J}=B{J}-T;
END;
DO I=1 TO M;
    DO J=1 TO N;
      IF C{I,J}=0 THEN F{I,J}=T+A{I}+B{J};
      IF C{I,J}=0 THEN C{I,J}=F{I,J};
    END;
END;
T=0;
K=0;
DO I=1 TO M;
    A{I}=0;
END;
DO J=1 TO N;
      B{J}=0;
```

```
END;
DO UNTIL (K>=L);
  K=0;
  DO I=1 TO M;
    DO J=1 TO N;
      D{I,J}=T+A{I}+B{J};
    END;
  END;
  T=0;
  DO I=1 TO M;
    A{I}=0;
    DO J=1 TO N;
      T=T+C{I,J};
      A{I}=A{I}+C{I,J};
    END;
    A{I}=A{I}/N;
  END;
  T=T/(M*N);
  DO J=1 TO N;
    B{J}=0;
    DO I=1 TO M;
      B{J}=B{J}+C{I,J};
    END;
    B{J}=B{J}/M;
  END;
  DO I=1 TO M;
    A{I}=A{I}-T;
  END;
  DO J=1 TO N;
    B{J}=B{J}-T;
  END;
  DO I=1 TO M;
    DO J=1 TO N;
```

```
            E{I,J}=T+A{I}+B{J};
        END;
    END;
    DO I=1 TO M;
        DO J=1 TO N;
            IF C{I,J}=F{I,J} THEN C{I,J}=E{I,J};
            IF C{I,J}=D{I,J} THEN C{I,J}=E{I,J};
        END;
    END;
    DO I=1 TO M;
        DO J=1 TO N;
            IF C{I,J}=E{I,J} AND ABS(D{I,J}-E{I,J})<0.1 THEN K=K+1;
        END;
    END;
END;
U1=0;
U2=0;
U3=0;
U4=0;
DO I=1 TO M;
    V{I}=0;
END;
DO I=1 TO M;
    DO J=1 TO N;
        U1=U1+C{I,J};
        U2=U2+C{I,J}*C{I,J};
        V{I}=V{I}+C{I,J};
    END;
    U3=U3+(V{I}*V{I})/N;
END;
DO J=1 TO N;
    W{J}=0;
    DO I=1 TO M;
```

```
            W{J}=W{J}+C{I,J};
          END;
          U4=U4+(W{J}*W{J})/M;
        END;
        G=(U1*U1)/(M*N);
        H1=U2-G;          /* SST */
        H2=U3-G;          /* SSt */
        H3=U4-G;          /* SSB */
        H4=H1-H2-H3;      /* SSE */
PROC PRINT;
        VAR Q5 S3 H1-H4;
RUN;
```

Imputing Missing Data in Analysis of Variance In Data Mining

SUNG-SHUN WENG*, TE-HSIN LIANG**

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

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

ABSTRACT

In agricultural experiment, it happens that the accidental causes such as natural disaster or pest diseases will make the field data incomplete. If this kind of data was analyzed by the ordinary procedures of the analysis of variance, then the results obtained might not be reliable. Allan and Wishart (1930) gave an approach to compute one missing value and constructed the concept of estimating missing data. Yates (1933) extended the approach of Allan & Wishart to several missing values by solving several equations simultaneously. Laird & Rubin presented a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as missing data-EM algorithm. This article follows'EM algorithm' to estimate missing value(s), and to be analyzed by the ordinary analysis of variance for the potato yield of 5 different varieties in Taiwan.

Keywords: missing data, EM algorithm