論述 》統計 · 調查

提升抽樣分層效益－視覺化集群分析

本文介紹家庭收支調查運用多變量集群分析法，將全國 7 千多個村里分層，再據以抽樣，以提升樣本代表性；並以行政區域地理圖形化技術，將龐雜的分層結果轉化為地理圖形，提升檢核分層結果效率。

莊文寬、陳錫慧（行政院主計總處地方統計推展中心科長、專員）

壹、前言

抽樣調查的確度高低，取決於樣本是否與母體長得很像，要如何長得像？須先描繪出母體的各種特徵，再據以抽樣，樣本就較能與母體長得像（具樣本代表性）。

但，說得容易，要描繪出母體的特徵（統計常用的手法：分層）實非易事。要對某個母體分層（或分類），須找到分類準則及變數，這可能是一個很長的故事。話說達爾文搭乘小獵犬號航行世界，費時 5 年歷盡艱險，蒐集紀錄動植物、化石及地理現象，之後完成了有關演化觀點的巨著－「物種起源」，生物分類學以此（共同祖先原則）爲基礎，建構生物的物種分類方法。其實，自然界簡單多了，萬物都是由原子構成，質子與中子緊偎著形成原子核，外面有電子繞著原子核運轉，各種原子依大自然給予的電子數量而有不同的活性（元素週期表可概括各種元素並區分屬性），並進而相互作用產生了日、月、雲、樹及你我。

然而，人類的社會可不像「自然界的元素」這麼單純而容易分類（此處說容易，是因爲自然界講究的是「理」，所以科學家能依一定的規則去分類及預測出未知的元素及行星），除了「理性」之外，人類還多了「感性」，因而使事情變得複雜，此時更是需要以科學客觀的方式來處理。以行政院主計總處家庭收支調查而言，影響家庭收入與支出的因

素非常複雜，如何將母體適當分類以提升樣本代表性是該調查重要課題，本文將介紹其抽樣分層設計及如何運用地理空間資料精進分層效益。

貳、家庭收支調查村里分層方法及視覺化分層結果

家庭收支調查旨在瞭解全體家庭的收入與支出情形，抽樣設計係採分層二段隨機抽樣法，先從全國 7 千多個村里中抽出 20％作爲樣本村里（第 1 段抽樣單位），再從中抽出樣本戶（第 2 段抽樣單位）。

影響家庭收支水準的因素相當複雜，爲提高樣本代表性，經考量與家庭收支調查內容較有關之社經特性及資料可取得性，將全國 7 千多個村里之人口之年齡、教育程度及就業人口之產業結構等特徵結構整理分類，使之近似欲推估的母體，再據以抽樣。此龐大複雜之村里分層作業係採用多變量集群分析法，並結合村里經緯度資料將分類結果轉化爲地理空間圖形，俾利檢視分層結果。以下簡要說明家庭收支調查村里分層方法及視覺化分層結果。

一、以多變量集群分析將全國村里依其社經特性予以分層

多變量集群分析可以將大量資料依某些特性予以分群，分群後，同一群內的個體，在這些特性上較相似（群內同質性高），但群與群之間的差異較大（群間異質性高）。因此，集群分析法可以用於分類及簡化資料量。

集群分析已廣泛應用於醫學、自然科學及經濟學等領域，是多維度資料重要分析工具之一。此法也可應用於統計調查，將母體特徵結構以集群分析法整理分層，再據以執行分層抽樣，以提高樣本代表性。家庭收支調查即運用此法，先將全部 7 千多個村里依人口之年齡、教育程度及就業人口之產業結構等變數進行分層，再據以抽樣，抽出的樣本就較易與母體相似。分層方法說明如下：

（一）決定層數

分層係爲了將具有相似特性的村里歸類爲同一集群（cluster），使得群內變異小、群間變異大。由於所使用的集群分析法事先不知道集群數目，故在執行村里分層之前，首要工作是決定層數，我們以學者 Milligan 建議較佳方法 - 凝聚階層式集群分析 - 華德法來找出較佳層數。

以年齡、教育程度及就業人口之產業結構爲分層變數，利用統計軟體 SAS，計算各個副母體綜合判定指標，包括：RMSSTD（Root-Mean-Square Standard Deviation）、RSQ（R-squared）及 PSF（Pseudo-F statistic），以決定較佳層數，其所代表意義分述如下：

1. RMSSTD 所呈現的是層內的離散程度，值小代表層內同質性越高（RMSSTD 越小越好）。
2. RSQ 係指層間的變異占總變異的比率，值大代表層間異質性越高（RSQ 越大越好）。

論述 》統計 · 調查

3.PSF 爲層間變異與層內變異的比率，值大代表層間異質性越高（PSF 越大越好）。

由於所採用方法爲凝聚式，隨著觀察值的合併，集群數會逐漸減少，致群內同質性降低、群間異質性降低，而集群分析的目的在於使群內同質性大、群間異質性大，因此，如果觀察值的合併使得上述各判定指標值突然增加或減少，表示應停止合併，據此決定較佳層數。

當集群數較多時，集群內觀察值相似性也會較高，但群數多相對上較不易解讀，通常集群數以 2 ～ 6 群爲宜。爲利判別集群數爲 6 的合宜性，下面範例在決定群數時僅列 2 ～ 7 群，且橫軸須由右往左判讀。

以某個縣市資料爲例，觀察圖 1，集群數由 7 降至 5 時，RMSSTD 變化不大，但再降爲 4 時，RMSSTD 增幅較大，因 RMSSTD 越小越好，故判定群數爲 5；RSQ 在集群數爲 5 變爲 4 時有較大跳動；PSF 在集群數 5 爲最大值；綜上，據此決定層數爲 5 層。

（二）將村里分層

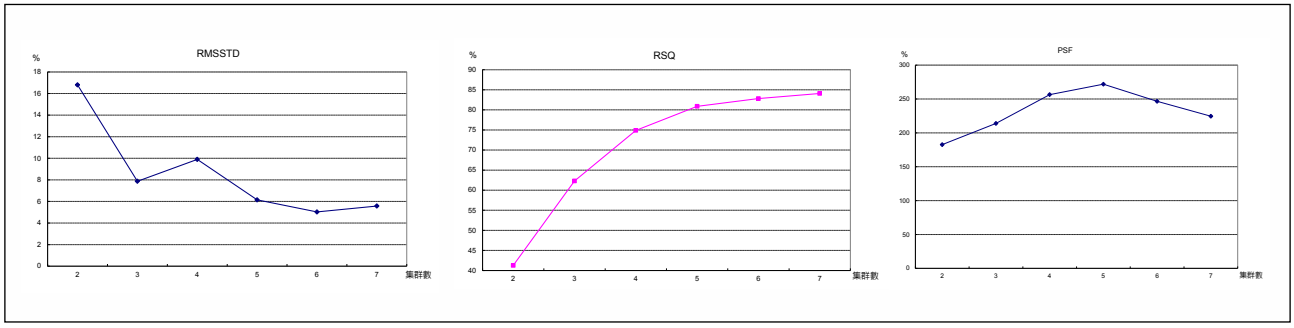
集群分析分爲階層式與非階層式集群分析，非直轄市係以各縣市爲副母體，其村里總數屬大樣本，因此採非階層式集群分析法－K-mean（Non-hierachical cluster-K-mean）進行分群；直轄市以行政區爲副母體，且其村里總數屬小樣本，因此以凝聚階層式集群分析－華德法（Hierarchical cluster-Wald's method），並以統計軟體 SAS 進行村里分層。

1. 凝聚階層式集群分析－華德法：每個觀察值均各自爲 1 群，先將兩個最靠近的群（合併後各群之組內變異總和最小者）合併成一個集群，重複此步驟，直到所有觀察值合併爲一個集群。舉 5 個村里之凝聚過程爲例：

步驟 1：每個村里均各自爲 1 群，此時有 5 群。

步驟 2：整併爲 4 群時之組內變異總和如表 1，

圖 1　決定層數之綜合判定指標



資料來源：作者自行繪製。

因 AB 合併爲 1 群時，4 群之組內變異總和最小，故 4 個集群爲 {A,B},{C},{D},{E}。

步驟 3：整併爲 3 群時之組內變異總和如表 2，因 DE 合併爲 1 群時，3 群之組內變異總和爲最小，故 3 個集群爲 {A,B},{C}{D,E}。

步驟 4：整併爲 2 群時之組內變異總和如表 3，因 CDE 合併爲 1 群時，2 群之組內變異總和爲最小，故 2 個集群爲 {A,B},{C,D,E}。

步驟 5：將 2 個集群合併爲 1 個集群之組內變異總和如表 4。

表 1　整併為 4 群之組內變異總和

	A	B	C	D	E
A					
B	1				
C	8.5	14.5			
D	6.5	12.5	4		
E	16	25	4.5	2.5	

資料來源：作者自行整理。

表 2　整併為 3 群之組內變異總和

	AB	C	D	E
AB				
C	16			
D	13.33	5		
E	28	5.5	3.5	

註：{A,B},{C},{D},{E} 整併為 3 群時，各種組合中各群組內變異之和。例：若整併為 {A,B},{C,D},{E}，此 3 群之總內變異總和為 1+4+0=5。
資料來源：作者自行整理。

表 3　整併為 2 群之組內變異總和

	AB	C	DE
AB			
C	18.5		
DE	31.75	8.33	

資料來源：作者自行整理。

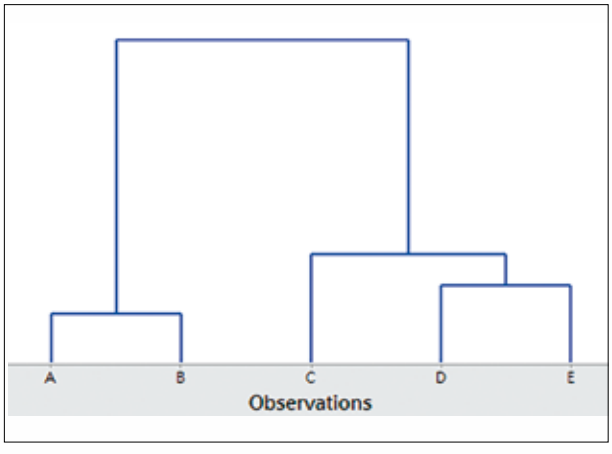
表 4　整併為 1 群之組內變異總和

	AB	CDE
AB		
CDE	38	

資料來源：作者自行整理。

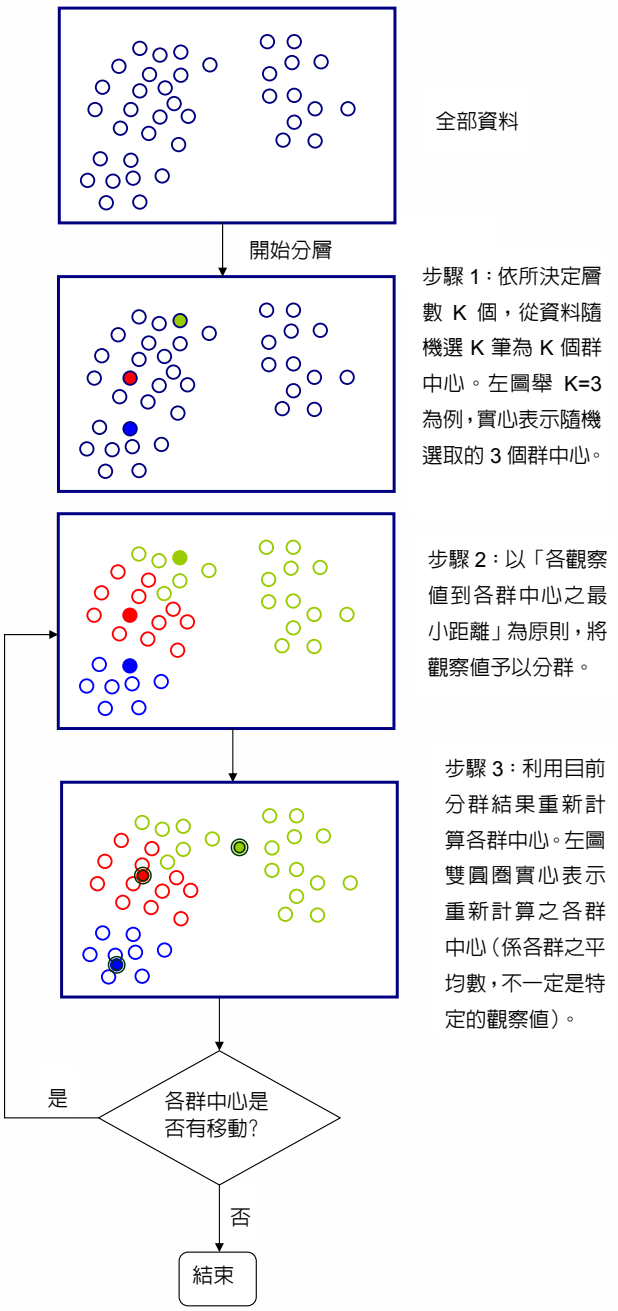
上述步驟 2 ～ 5 茲以樹狀圖（圖 2）呈現集群凝聚過程。

圖 2　集群過程之樹狀圖



資料來源：作者自行繪製。

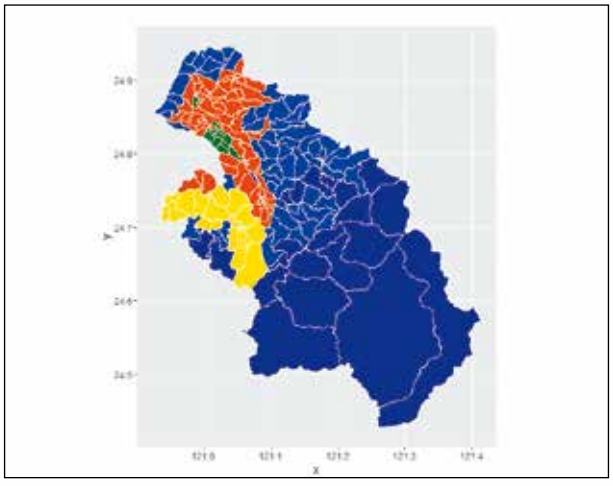
2. 非階層式集群分析法－K-mean：係找出 K 群中心，使得所有資料點到其對應 K 群之中心距離和是最小的，而求得群中心的解法係用疊代方式，步驟如下：



二、視覺化集群分析－以地圖呈現村里分層結果

全部村里經由前述集群分析法予以分層後，由於資料共有 7 千多筆，不易檢核分層結果，爰利用內政部國土資訊系統村里經緯度資料，以統計軟體 R 及 SAS 撰寫程式，並將村里分層結果轉化為地理空間圖形（圖 3），據以協助複核分層結果。

圖 3　以地圖呈現分層結果



資料來源：作者自行繪製。

參、結論

抽樣調查依據統計理論，經由將抽樣母體分層及機率抽樣法，可以抽出一組樣本，代替曠日費時的普查，大幅降低成本及提升時效。家庭收支調查的抽樣分層運用客觀的統計技術，發揮分層效益；並與時俱進，透過地理空間圖形技術，將大量的分層結果轉化為地理圖形，大幅提升檢核分層結果效率。期待科技的發展，協助統計調查作業更為精進。❖