國立政治大學經濟研究所碩士論文

勞工職位特質分析
-多元尺度法於大資料分析之應用

The Occupational Characteristics Analysis
-The Application of Large Data Multidimensional Scaling Method

指導老師：曾正男博士、李浩仲博士

研究生：陳烽威

中華民國一百零一年八月

摘要

本文自美國人口普查局 (United States Census Bureau) 取得多達十萬筆的勞工資料，然而在如此大量的勞工資料中因維度的詛咒，所以我們無法使用傳統的資料探勘的方法分析資料，而且傳統的序述統計也無法提供一個好的分析方向，因此我們藉由Tzeng et al. (2008)所提出的分解與結合多元尺度法 (Split-and-combine Multidimensional Scaling, SC-MDS) 爲分析方法來剖析此資料。多元尺度法主要的目的有二：第一，使資料展現在空間中，並以資料點與點之間的距離表示其相關性；第二，降低資料維度避免維度的詛咒。SC-MDS提供我們在分析此大資料相關聯性時的優先順序爲年齡、學歷、性別；並結合職位資訊聯合資料庫 (Occupational Information Network)分析在此架構下不同分類的勞工在其就業的職位特質上的差異。我們發現了教育程度會影響性別間在勞工職位特質上的差異，且這些差異的數量又會隨年齡的增加而增加；教育程度在各個年齡層都對勞工職位特質產生很大的差異；最後，青年與壯年的勞工在職位特質上相較於壯年與中年勞工相似，並對以上產生相似或差異的原因提出解釋。

Abstract

A big labor data from United States Census Bureau will occur two problems. First, since the big data issue, we can not use the traditional method of data mining. Second, the descriptive statistics can not offer an explicit analysis, so we use Split-and-combine Multidimensional Scaling (SC-MDS), which is proposed by Tzeng et al. (2008) to mining this labor data. MDS has two main purposes: First, Express data similarity by the distance between each pair points in spatial configuration. Second, Reducing data dimension to aviod the curse of dimension. After SC-MDS, the big labor data can be analysed by age, education and sex. We combine this order and the Occupational Information Network data base to develope the differences in occupational characteristics. We find the following phenomenon: first, differences are increasing with ages. Second, eduction do impact labors' characteristics in every ages. Third, the youth labors are more similar in occupational characteristics than olders. Finally, we try to explain the results above.

誌謝

短短兩年的碩士生涯終於要畫下句點，最風風雨雨的莫過於碩士論文一事，雖然一路上峰迴路轉，但最後也是柳暗花明，對於這一切的經歷我感恩惜福。首先感謝家人，感謝他們對於這個兒子如此包容與支持，並提供了不畏風雨的家，讓我能無後顧之憂地在求學階段完成許多想做的事。再來是我的指導老師曾正男博士，從碩一新生開始的旁聽到後來擔任通識課助教，雖然不過兩個寒暑，但我常從曾老師身上聯想到薩依德 (Edward W. Said) 在知識分子論一書中的一句話:「知識分子是以代表藝術爲業的個人，不管那是演說、寫作、教學或上電視。而那個行業之重要在於那是大眾認可的，而且涉及奉獻與冒險，勇敢與可能受到傷害」；常能從曾老師身上感受到對社會、人群與學生的愛，以及那種從老師身上射放出來的光線，那樣的熱情感動了不只我，還有許許多多的學生。還有另一位指導教授李浩仲博士，感謝老師在百忙之中爲我的論文提供寶貴的意見。還有研究室的大家，感謝澤佑與冠慧特地花時間爲我校閱這篇文章，還有靜儒、沛承、宥柔、裕哲等，對我這麼一名突如其來的客人提供了一個溫馨且充滿歡樂的研究室。還有一些無數的朋友在這一路上的加油打氣與關懷，謝謝你們。我終於要畢業了。

目錄

圖目錄

表目錄

1 緒論

1.1 研究動機

自古以來勞工都是經濟社會的砥柱，因此無論是失業率、薪資福利甚至教育環境都是執政當局所關注的議題。勞工的組成複雜難以分類，使影響勞工行爲的因素盤根交錯，所以當今社會學家都希驥藉由政治、經濟、教育與心理各個角度去剖析、釐清其中的因果關係，進而提供執政者在施政方針的檢討與建議。

就業議題也是影響勞工行爲的一大因素，其中求職過程中勞方與資方的媒合更是一大課題。各個企業在尋找理想的勞工時，可從應徵者中評判是否符合其應徵職位特質，抑或是從各個職位的特質去尋找合適的勞工；然而，這些勞工的特質如能力、知識、技能與工作活動上又與其基本條件有關，如年齡、性別、種族、教育程度等有關；然而，不同於婚姻狀況、居住城市、產業類型等，這些基本條件是勞工在尋找工作時無法選擇或改變的部份。所以該如何從中找出影響勞工特質的項目，便能提供客觀的建議使企業運用在徵選人才，或政府在制定教育政策與勞工福利上。

早在 1940 年代開始美國人口普查局 (United States Census Bureau) 就進行系統性地統計全國勞動力普查資料；然而在資料的使用方面， 在政治、經濟領域的發展中，學者大多直接使用實證計量模型去探討其因果關係，卻在一開始的資料分群上著墨不多。本論文主要是希望藉由在心理學、資訊科學與生物資訊學上廣泛使用的分群分析 (clustering analysis)的概念，藉由 Tzeng et al. (2008) 所提出的分解與合併多元尺度法 (Split-and-Combine Multidimensional Scaling, SC-MDS) 來剖析勞工資料，進而建構一個分群分類的架構，並且以此架構結合職位資訊聯合資料庫 (Occupational Information Network)，分析影響勞工特質的因素，並對這些變因提出經濟邏輯上的解釋。

1.2 文獻回顧

隨科技的發展，記錄與儲存資料已不再是個高成本的行爲；相反地，資料的

產生速度遠遠超過科學家分析和消化的速度。沒有被分析的資料就等於一塊巨石，它不只佔據空間，更會浪費資源；但如果能夠從中挖掘出有用的資訊，則可點石成金。

因此，當前無論是學界或是產業界，都在苦思該如何著手解構龐大的資料。如金融業，信用卡的活動促銷，能不能有效且適時適地傳達給消費者便是個重要的議題，因爲如果無的地放矢，會造成企業資源的浪費，但又該如何從廣大的消費者資料中找到合適的群體，又是件令人頭痛的問題。又如半導體業，生產業越趨複雜，記錄每件產品的生產參數是件輕而易舉的事，但對於每件瑕疵品該如何從繁複的流程中抓錯，最有效率的方法絕對不是召集所有工程師坐下來一一抓錯；相反地，應該是建立一套分群體系，將各種瑕疵品的生產流程分類，以致在下次產率下降時，能快速找出問題所在。以上的例子，都需使用資料探勘 (data mining) 來挖掘資料中的資訊。

Frawley et al. (1992) 指出資料探勘就是從資料中提取出隱含的過去未知的有價值的潛在信息，但資料探勘與一般統計方法不同，因爲大多數的資料都由多條規則或模型混合而生的。如消費者的偏好可能受天氣、消費行爲、節慶活動等影響；而工廠的產率也可能因原料、設備儀器或操作人員所改變。所以如果只用單一規則或模式去解讀資料，便可能會產生處處是變因、處處都重要的無用資訊。而資料探勘就是在考量所有的規則與模式，並透過統計分析或人工智慧的技術，去規納出哪一些規則或模式是重要的，進而幫你安排解構資料的優先順序。一般來說，資料探勘的方式有神經網絡學 (neural netwoks)、基因演算法 (genetic algorithms)、統計推論 (statistical inference)與視覺化資料分析 (data visualiztion)等。然而，在記錄與儲存資料的方式改變與其成本的降低，大資料 (big data) [1] 已隨之而生，對於大資料進行探勘的步驟分別爲降低維度 (dimension reduction)、分群 (classification)、估計(estimation)、預測 (prediction)。以下我們介紹降低維度與分群的相關文獻。

[1]White (2009)定義大資料就是太過龐大且複雜的資料，讓使用者在截取、儲存、搜尋、分享、分析甚至視覺化的處理上已無法使用傳統的資料庫工具解決問題。

首先，降低維度是爲了減少資料搜索的成本；常使用的方式有：主成份分析法 (Principal Components Analysis, PCA)、亂數投影 (Random Projection)與多元尺度法 (Multidimensional Scaling, MDS)。PCA 是由 Pearson (1901) 所提出，其概念爲保留原變數間最大變異 (variance) 的部份，並透過線性組合產生新的變數，直到新變數間的變異能函蓋大部份原變數間的變異。亂數投影的概念是來自 Johnson and Lindenstrauss (1984)與 Dasgupta and Gupta (1999)。他們指出若能將向量空間中的點隨機投影到相符的子空間時，在大部份的點與點間距離與相對角度都能被保留。MDS 是由Torgerson (1952)所提出，他是使用將資料間的相似度以空間中點與點間距離表達，如此便能在低維度的空間裡分析高維度的資料。以上三個種降維方式都廣泛地使用在都市計劃、市場研究、性格分析、圖像辨識、財務工程與生物資訊上。本論文使用 MDS 的方式降低資料維度。

進十年來隨計算機的發展 MDS 不再只限於行銷學或心理學等領域，也開始應用在資料較爲龐大的財金領域。如 Groenen and Franses (2000) 使用 MDS 來分析不同股票市場間股價的時間序列資料；Dwyer and Gallagher (2004) 透過 MDS 將投資組合以視覺化的圖像分析。然而，傳統的 MDS 方法都會面臨到當原始資料增加時，讓計算的複雜度 (complexity) 急劇上升的問題；即使當今科學家擁有今非昔比的計算機，但當我們需要分析如生物基因、投資組合或是勞工資料等數據時，在傳統的 MDS 方法上都會面臨到因資料過於龐大使計算耗時甚至電腦無法計算的問題。傳統的 MDS 計算複雜度爲 $O(N^3)$，但在 1990 中期開始有科學家著手改進 MDS 的計算流程；如 Chalmers (1996) 提出線性疊代運算法 (linear iteration algorithm) 將其計算複雜度降爲 $O(N^2)$；之後 Morrison et al. (2003) 將部份非計量 MDS 的計算複雜度再降爲 $O(N\sqrt{N})$；然而以上的計算量對於動則上萬的大資料來說，仍是太大。所以 Tzeng et al. (2008)提出 SC-MDS 使在特定維度的資料能使計算量大幅降爲 $O(N)$。如此的計算量就可以用來處理分析大資料。因此，本文中我們使用 SC-MDS 進行降低維度的資料分析。

在分群方面，常使用的方式有：k-mean 演算法 (k-mean algorithm, kmeans)、期望值極大演算法 (Expectation maximization algorithm, EM) 或最臨近搜索法 (Nearest Neighbor Search, NNS)。其中 kmeans 由 Lloyd (1957) 所提出，並由 Lloyd (1982) 發表標準的演算法 [2]；k-means需要事先知道分群的數目，記錄每筆資料的距離以產生分群的邊界，並藉由計算分群中的質心是否有所改變，以決定是否分群完成。EM 是由 Dempster et al. (1977)所提出，主要是在尋找機率模型中參數的最大概似估計式使用的演算法。NNS的概念在 Knuth (1973)一書中以居民該如何搜尋離自己住家最近的郵局爲例[3]，其概念就是去記錄每個資料點間的位置，並找出離目標集合最近的資料點，進而將該點歸納於該目標集合。最後，若我們一開始便知道資料分群的數目，便可使用 kmeans 或 NNS，但若資料的分佈較爲特殊，如每分群以層狀的方式分佈在空間裡，則使用 NNS 演算法爲最好的分群方式，所以本文我們以 NNS 將 MDS 的結果分群。

最後資料探勘還有估計與預測，這兩個方式都是跟據分群的結果再做下一步的推算。但也有演算法從資料輸入後便自動幫分析者進行所有資料探勘的步驟，如支持向量機 (Support Vector Machines, SVM) 或線性辨識分析 (Linear discriminant analysis, LDA) 等。然而，本論文主要並非在勞工資料的預測，而是分群歸納勞工資料；因此我們將採用 MDS 作原始資料的降維與 NNS 作降維後資料的分群，並找出在勞工資料中隱藏的分析架構。

1.3 文章架構

本論文一共分爲五章。第一章爲緒論，介紹研究動機與文獻回顧。第二章爲 MDS 的理論介紹，包含傳統 MDS 與 SC-MDS 的理論方法和最鄰近搜索法的介紹。第三章爲使用 SC-MDS 的方式分析美國當期人口調查 (Current Population Survey, CPS) 的勞工資料，繪製出該筆資料的三維空間視圖與分群樹狀圖。第四章爲結合職位資訊聯合資料庫 (Occupational Information

[2]k-means 又被稱爲 Lloyd 演算法。

[3]NNS 在該書中稱爲郵局問題。

Network, O*NET) 與 CPS，並採用第三章所建構分析勞工資料的步驟逐一拆解職位特質資料。第五章爲結論。

2 大資料的多元尺度法與最鄰近搜索分群法

2.1 維度的詛咒

搜尋與分群是分析數據的第一步驟，在數值分析上常使用距離的概念來進行搜尋與分類資料。譬如說，假設有一家航空公司有台灣直飛香港、上海、柏林與巴黎這四條航線，該公司想知道飛機油料、工作人員和餐點飲食該如何配置，此時我們會很直覺地說這四條航線可分成兩組，香港與上海一組、柏林與巴黎一組，這是因爲在腦海中我們知道從台灣到香港、上海的飛行浬程相近，而到柏林與巴黎的距離相近；或者換句話說，如果以這四個城市各自爲圓心，搜尋其它三個城市是否在半徑2000公里內，一樣可以知道，以香港爲圓心只可搜尋到上海，而以巴黎爲圓心只可搜尋到柏林，所以這四個城市就可以分成兩組。

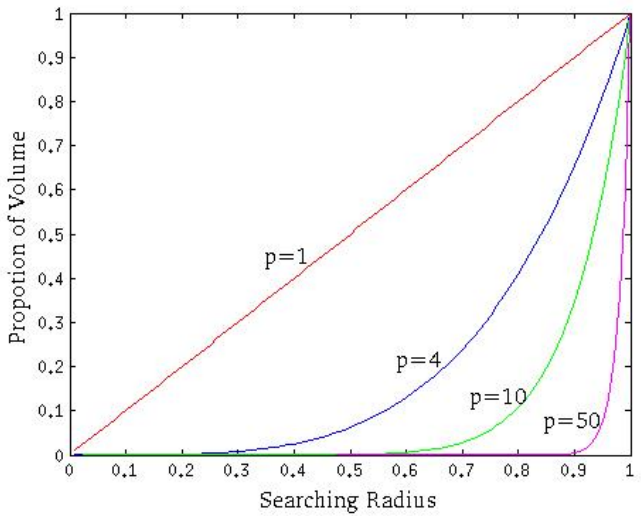


圖 2.1: 維度的詛咒

上述的例子是二維資料，由於維度較低，因此搜尋的複雜度與時間成本都較低；然而當空間維度上升時，會遇到搜尋體積與單位體積的比例呈指數倍減的問題：單位體積內，以一點爲中心且搜索半徑固定的體積會逐步隨維度的上升而佔單位體積的比例指數倍減，最後這個比例可能會趨近於零，此時研究者就需要

更多樣本點以增加搜尋體積，但如此搜索資料的複雜度與時間成本便會隨之增加，這個現象叫作維度的阻咒(curse of dimensionality)。如圖(2.1) 所示，當我們設定搜尋半徑爲 0.7 時，對應維度爲 $p = 4$ 時，這搜尋體積約佔單位體積的 20%；然而，一旦維度上升到 $p = 10$ 時，同樣的半徑搜尋體積卻只佔單位體積不到 5%，甚至在維度爲 50 時，比例就趨近於零了。這樣的例子告訴我們，即使將搜索半徑定在單位長度的 70%，一旦維度上升到 50 這樣的搜索範圍仍是缺乏精準度的。

Bellman (1957)提出維度的詛咒，該文指出我們能使用 100 個點將單位長度畫分成每點距離不超過 0.01 的區間；但當維度增加到 10 時，相鄰不超過 0.01 的單位超正方體，則需要 10^{20} 個點；平均畫分區間的採樣點隨維度增加而成指數增加，如此便造成了維度的詛咒。維度的詛咒在採樣 (sampling) 、組合數學 (combinatorics) 、機器學習 (machine learning) 與數據探勘 (data mining) 都有所討論，雖然各家研究的問題不同，但共同的特色都是在資料維度提高時，空間體積提高得太快，使數據變得稀疏，進而降低分析的效率與精確性。爲避免維度的詛咒，最直接的就是透過降低維度 (dimension reduction) ，使資料不要展現在高維度的空間。本文則使用 MDS 爲降低維度的方法，下節我們介紹 MDS 的基本理論與延申方法。

2.1.1 多元尺度法的意義

MDS 是由 Torgerson (1952)所提出，主要是延伸 Young and Householder (2003)後所提出的一套分析心理行爲學中刺激體與事件之間關聯性的方法；簡單來說，MDS 是一種 N 個主體 (subject) 根據 p 個準則 (criterion) 評估 M 個客體 (objects) 的統計方法。以購買咖啡的行爲爲例，如果我們想知道購買星巴克的消費者有那幾種類型。此時我們以星巴克爲客體，選擇 N 個消費者爲主體，再請他們各自依據 p 個準則，如價格、口味、品牌或包裝等，來給予分數，因爲每個消費者是同時依據這 p 個準則給予分數，所以我們不能只依據某一個準則，如價格，來給予分群，而是需要考量全部的準則。此時 MDS 就是一個很好

的工具讓我們發掘這 N 個消費者，在這 p 個準則下的相似性與差異性。

MDS 的精神在於將每筆資料間的相關性轉換成各筆資料間的距離，它可以協助研究者在歐氏空間 (Euclidean space) 中發掘出隱藏在原始資料背後的結構圖形 (configuration)；換句話說， MDS 是將主體以空間上的點來代表，而主體各自之間的關係則透過點與點的距離來描述。同上面的例子，如果我們要採用一千個準則來分類一萬名購買星巴克的消費者，這一萬筆一千維度的資料會面臨到以下兩個問題：第一，維度的詛咒；因爲資料維度已經高達一千維，直接使用原始資料來作相關性的搜索與分類需付出很高的計算代價，而且如此稀疏的資料也會降低分析的精準度。第二，無法單一採用少數幾個準則來分類；因爲消費者是同時依據各項準則來評分，所以如果我們只單看少數幾個準則，如價格，那會喪失很多原始資料中的訊息。解決這類高維度資料問題， MDS 就是一個很好的方法。

MDS 提供了一個能保持原始資料的相似性來描述資料；它使相關性越高的資料，距離越近；且若能將維度降至三維以下，更可以藉由視覺化的方式來觀察資料。此外，對於一樣能以圖像化分析資料的 PCA ， MDS 的優點在不需要線性的假設，但缺點在不太容易對向度作命名；也就是 MDS 能保留了資料的相似性卻無法說明分群的相似處在那，所以我們需藉助最鄰近搜索分群法，將分群後的資料取出來，比對各項準則後才能找出產生分群的關鍵因素。

2.1.2 多元尺度法的理論架構

MDS 又可依研究者所獲得的資料分爲計量多元尺度法 (Metric Multidimensional Scaling,計量 MDS) 和非計量多元尺度法 (Non-metric Multidimensional Scaling, 非計量 MDS)。計量 MDS 由 Torgerson (1952) 所提出，該輸入資料需是連續性尺度，如距離尺度 (interval scale)、比率尺度 (ratio scale) 等；而非計量 MDS 由 Shepard (1962) 與 Kruskal (1964) 所提出，該輸入資料爲間斷尺度，如順序尺度 (ordinal scale)、名辭尺度 (nominal scale) 等。

由 MDS 使用距離來描述每筆資料間相關性的特質，會有兩個問題需要解決；第一，距離矩陣該如何求出？第二，在滿足距離矩陣後，該使用哪個向量空間展現主體的座標？以下我們分別介紹計量 MDS 與非計量 MDS 的理論架構與計算流程。

計量多元尺度法

假設資料型態是 n 個主體，且每個主體有 p 個準則，此時我們可以將資料型態以矩陣 $\mathbf{X}_{p\times n}$，而該矩陣的每一列皆爲主體之準則，表示爲 $x_i = (x_{1i},\ x_{2i,\cdots,x_{pi}})$，視爲 p 維歐氏空間中的一點。爲了確保 $\mathbf{X}$ 在MDS後具有唯一性且各主體的相對位置對資料分析並無影響，所以我們先將每筆資料移到質心，$\sum_{i=1}^{n} x_i = 0$。接著，定義 $\mathbf{D}$ 爲 $\mathbf{X}$ 的距離平方矩陣，$\mathbf{D} = [d_{ij}]_{n\times n} = (x_i - x_j)^T (x_i - x_j)$，由距離平方矩陣可以得到以下關係:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} d_{ij} =& \frac{1}{n}\left(\sum_{i=1}^{n} x_i^T x_i - x_j^T \sum_{i=1}^{n} x_i - x_j^T \sum_{i=1}^{n} x_i + n x_j^T x_j\right)\\ =& \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i + x_j^T x_j \end{aligned} \tag{2.1a}$$

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} d_{ij} =& \frac{1}{n}\left(\sum_{j=1}^{n} x_j^T x_j + x_i^T \sum_{j=1}^{n} x_j + x_i^T \sum_{j=1}^{n} x_j + n x_i^T x_i\right)\\ =& \frac{1}{n}\sum_{j=1}^{n} x_j^T x_j + x_i^T x_i \end{aligned} \tag{2.1b}$$

$$\begin{aligned}\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij} =& \frac{1}{n^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} x_i^T x_i + n x_j^T x_j\right)\\ =& \frac{1}{n}\left(\sum_{i=1}^{n} x_i^T x_i + \sum_{j=1}^{n} x_j^T x_j\right)\end{aligned} \tag{2.1c}$$

爲了取得每個 x_i 的相對關係，我們定義 $\mathbf{B}$ 爲 $\mathbf{X}$ 的內積矩陣，$\mathbf{B} = \mathbf{X}^T\mathbf{X}$；在 Young and Householder (2003) 中提到，使在歐幾里得空間裡實數集合中各點的相互距離相等 ($d_{ij} = d_{ji}$) 之充份且必要條件爲 $\mathbf{B} = \mathbf{X}^T\mathbf{X}$爲半正定矩陣 (semi-positive definition matrix)；且這集合中的各點在歐氏轉換 (Euclidean

transformation) 後唯一。結合式(2.1a)到式(2.1c)可得到:

$$\begin{aligned} b_{ij} =& x_i^T x_j = -\frac{1}{2}\left(d_{ij} - \frac{1}{n}\sum_{i=1}^{n} d_{ij} - \frac{1}{n}\sum_{j=1}^{n} d_{ij} + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}\right) \\ =& -\frac{1}{2}\left(d_{ij} - d_{i\cdot} - d_{\cdot j} + d_{\cdot\cdot}\right) \end{aligned}$$

由上式，矩陣 $\mathbf{B}$ 可表示爲:

$$\mathbf{B} = \mathbf{X}^T\mathbf{X} = -\frac{1}{2}\left(\mathbf{D} - \bar{\mathbf{D}}_r - \bar{\mathbf{D}}_c + \bar{\mathbf{D}}_g\right) \tag{2.2}$$

式(2.2)中， $\bar{\mathbf{D}}_r = [d_{i\cdot}]$ 爲 $\mathbf{D}$ 的行平均； $\bar{\mathbf{D}}_c = [d_{\cdot j}]$ 爲 $\mathbf{D}$ 的列平均； $\bar{\mathbf{D}}_g = [d_{\cdot\cdot}]$ 爲 $\mathbf{D}$ 的群平均。式(2.2)又稱爲 double centering。透過 $\mathbf{B}$ 正半定矩陣的性質與式(2.2)，$\mathbf{B}$可表示爲:

$$\mathbf{B} = \mathbf{HPH} \tag{2.3}$$

式(2.3)中， $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$，且 $\mathbf{P} = [p_{ij}]_{n\times n} = -\frac{1}{2}\mathbf{D}$。式(2.2)與式(2.3)說明了，只需將針對距離平方矩陣 $\mathbf{D}$ 作double centering後再乘上 $-\frac{1}{2}$，便可得到內積矩陣 $\mathbf{B}$。

接下來，我們還需要找到適當的空間去展現各筆資料的空間座標。透過奇異值分解 (Singular Value Decomposition, SVD) 與 $\mathbf{B}$ 爲對稱矩陣的性質，可以將 $\mathbf{B}$ 拆解爲:

$$\mathbf{B} = \mathbf{Z}\Sigma\mathbf{Z}^T \tag{2.4}$$

式(2.4)中， Σ 將特徵值 (eigenvalue) 由大到小在對角線上排列的對角矩陣，$\mathbf{Z}$ 爲正交矩陣；因此，由式(2.2)與式(2.4)可以得到 $\mathbf{X} = \Sigma^{\frac{1}{2}}\mathbf{Z}^T$。此時，爲了使降維的同時不影響 MDS 展現原始資料的相關性，我們採用類似主成份分析法(PCA)的方式，將相對較小的特徵值捨去，只保留 $N - r$ 個特徵值。最後，我們可以得到 MDS 後資料的座標爲:

$$\tilde{\mathbf{X}} = \Sigma_r^{\frac{1}{2}}\mathbf{Z}_r^T\ , \quad r < p \tag{2.5}$$

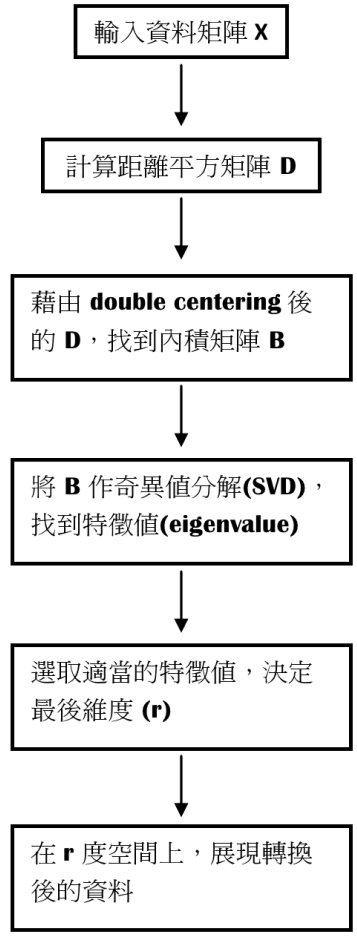


圖 2.2: 計量多元尺度法流程圖

式(2.5)爲將 p 維度的資料使用 MDS 降至 r 維而不失原始資料內相關性的結果。這裡必須要強調的是，能將多維度的資料降至三維以下，是爲了協助研究者能以圖像化的方式分析資料，但並不是每筆資料都可以降至三維以下而不失相關性，因此降低維度的程度得視特徵值的相對大小而定。

非計量多元尺度法

在上一節的計量 MDS ，是分析連續性尺度的資料，但在實務上，收集到的資料多爲間斷尺度，如年齡、性別、喜好程度等，此時我們就無法從資料中

直接算出距離矩陣，使計量多元尺度法無法運作。Shepard (1962) 與 Kruskal (1964) 提出解決間斷尺度的方法，稱爲非計量 MDS。

非計量 MDS 依然保有計量 MDS 的精神，只是將觀察所得到的間斷尺度資料間的關係與計算後的距離有單調 (monotonic) 關係；換句話說，在最後展現在低維度空間中各點的相對距離需與原始資料次序關係一致。舉例來說，如表(2.1)，有三個主體 A、B、C 分別對星巴克店內裝璜的滿意度爲很滿意、尚可、不滿意。我們知道，喜好程度滿意與尚可相較於滿意與不滿意接近，所以在最後展現的空間上，A 與 B 應該較接近，A 與 C 應該較遠。這裡我們要強調的是：非計量 MDS 中，因爲次序關係是由相對距離的遠近所表達，所以絕對的距離是沒有代表任何意義的。

主體	間斷尺度資料	可能距離
A、B	很滿意、尚可	0.5
A、C	很滿意、不滿意	1.5
B、C	尚可、不滿意	0.6

表 2.1: 次序資料與可能距離間的單調關係示意圖

此外，滿足主體 A、B 之間的距離最近，A、C 之間的距離最遠的條件可能有無限多種；換言之，N 個主體，有 $C_2^N = \frac{N(N-1)}{2}$ 個次序關係。如果我們想將 N 個點展現在 r 維度的空間裡，則須要 $N \times r$ 個座標值才能表示，但當次序的關係比座標值的數目相對增加時，次序關係便會限制各個座標的相對位置，如此便能求出唯一解。以下我們介紹非計量 MDS 的計算過程。

□ 非計量 MDS 的計算過程:

1. 計算各成對主體間相似程度的排列順序。

2. 選擇一個 r 維度的空間配置。

3. 計算各點間的歐氏距離。

$$d_{ij} = \left[\sum_{k=1}^{q} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} \tag{2.6}$$

4. 計算一致不等式 $\hat{d}_{ij}$。

   $\hat{d}_{ij}$為在 r 維空間中 i, j 兩點距離的單調函數 (monotonic function)。因為相對距離須與原始資料的次序關係一致，所以當有兩組或兩組以上的相對距離與其次序關係一致時，$\hat{d}_{ij} = d_{ij}$；一旦有不一致的情形出現，則$\hat{d}_{ij}$為其平均值。

5. 計算壓力係數 (Standardizes Stress)。

   Kruskal (1964)提出壓力係數的觀念與指標，目的在衡量 q 度空間中各點相對距離與原資料次序關係一致度的指標。計算方式如下：

$$\text{Stress} = \frac{\sum_{i \neq j}^{r} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i \neq j}^{q} d_{ij}^2} \tag{2.7}$$

6. 使用最陡斜坡法來重新估算空間配置。

7. 重新計算壓力係數，若得到比原先更小的壓力係數，則回到步驟 3或停止。

更詳細的非計量MDS計算過程，請參考 Cox and Cox (2001)。

分解-結合多元尺度法

Tzeng et al. (2008) 提出了分解-結合多元尺度法 (Split-and-Combine Multidimensional Scaling, SC-MDS) 解決 MDS 在分析資料數目龐大且資料維度較小的問題。SCMDS 的概念是：要得到大資料在 MDS 後的空間配置，並不需要資料內所有兩兩之間的相互距離。舉例來說，如果我們想將一個點很精確地加到一個平面上，只需要知道此點與三個任意但非共線性的點之間的距離，而不需此點與所有平面上的點之間的距離；換言之，如果我們想要將一個點加到一個 q 維

空間上，只需要知道此點與 $q+1$ 個點之間的相互距離即可。如此，當我們欲分析的資料型態是 p 維度的N 筆資料且 $N \gg p$ 時，這樣的概念就能大量地簡化傳統 MDS 在計算上的複雜度。

SC-MDS 的精神是運用以上的概念將大資料拆解成數個相互重疊的子集合，在得到各個子集合分別進行 MDS 之後的空間配置後，再使用重疊的部分將每個子集合結合起來。以下我們將介紹在 SC-MDS 中，分解與結合這兩個步驟。

分解:

假設原始資料爲 p 維 N 筆 $(N,p \in \mathbb{N},\ N \gg p)$。我們將資料分解成兩群 $\mathbf{X}_1$ 與 $\mathbf{X}_2$ 且兩者重疊爲 $\mathbf{Y}$ ($\mathbf{X}_1 \cap \mathbf{X}_2 = \mathbf{Y} \neq \phi$)。兩群各自作完 MDS 後的空間配置爲 $\mathbf{X}_1'$ 與 $\mathbf{X}_2'$，$\mathbf{Y}_1'$ 與 $\mathbf{Y}_2'$ 爲重疊的部份在 $\mathbf{X}_1'$ 與 $\mathbf{X}_2'$ 裡的空間座標。

結合:

由 Young and Householder (2003) 可知，這兩個空間配置在歐氏轉換後一定一致。因此，我們希望找到一個仿射矩陣 (affine mapping) $\mathbf{U}(\cdot)+\mathbf{b}$ 讓每一個 $x_{1j}' \in \mathbf{Y}_1$ 可以在 $\mathbf{Y}_2$ 找到一個對應的點滿足 $x_{1j}' = \mathbf{U}x_{2j}' + \mathbf{b}$。將 $\mathbf{Y}_1'$ 與 $\mathbf{Y}_2'$ 平移到各自的質心後作 $\mathbf{QR}$ 分解，則可以得到:

$$\begin{aligned}\left(\mathbf{Y}_1 - \bar{\mathbf{Y}}_1\mathbf{1}^{\mathrm{T}}\right) &= \mathbf{Q}_1\mathbf{R}_1 \\ \left(\mathbf{Y}_2 - \bar{\mathbf{Y}}_2\mathbf{1}^{\mathrm{T}}\right) &= \mathbf{Q}_2\mathbf{R}_2\end{aligned}$$

利用上三角矩陣 $\mathbf{R}_1$ 與 $\mathbf{R}_2$ 相等，整理上式我們可以得到:

$$\mathbf{Y}_1 = \mathbf{Q}_1\mathbf{Q}_2^{\mathrm{T}}\mathbf{Y}_2 - \mathbf{Q}_1\mathbf{Q}_2^{\mathrm{T}}\bar{Y}_2\mathbf{1}^{\mathrm{T}} + \bar{\mathbf{Y}}_1\mathbf{1}^{\mathrm{T}} \tag{2.8}$$

由式(2.8)可知， 此仿射矩陣爲$\mathbf{U} = \mathbf{Q}_1\mathbf{Q}_2^{\mathrm{T}}$ 且 $\mathbf{b} = \mathbf{Q}_1\mathbf{Q}_2^{\mathrm{T}}\bar{Y}_2\mathbf{1}^{\mathrm{T}} + \bar{\mathbf{Y}}_1\mathbf{1}^{\mathrm{T}}$。

SC-MDS 的計算量

傳統的 MDS 的計算量爲 $O(N^3)$。在 SC-MDS 中，假設將資料分成 K 個子集合，每個子集合有 $N_g = \alpha \times (p+1)$ 筆資料，而N_I 每兩兩個子集合重疊的

部份；則可以得到 $KN_g - (K-1)N_I = N$ 或是 $K = \frac{N-N_I}{N_g-N_I}$。每個子集合各自運算 MDS 的運算量爲 $O(N_g^3)$ 而重疊部份的 **QR** 分解運算量爲 $O(N_I^3)$，所以我們可以得到全部 SC-MDS 的運算量爲:

$$\begin{aligned} & KO\left(N_g^3\right) + (K-1)\,O\left(N_I^3\right) \\ & \quad = \frac{N-p-1}{(\alpha-1)(p+1)} O\left(\alpha^3(p+1)^3\right) + \frac{N-\alpha(p+1)}{(\alpha-1)(p+1)} O\left((p+1)^3\right) \\ & \quad \sim O\left((p+1)^3 N\right) \end{aligned} \tag{2.9}$$

由式(2.9)可知，當 $N \gg (p+1)$ 時，SC-MDS 的計算量爲收斂到 $O(N)$，相較於 Morrison et al. (2003) 所改進的非計量 MDS 的計算量 $O(N\sqrt{N})$ 還少，所以在下一章我們將使用 Tzeng et al. (2008) 所提出的 SC-MDS 來進行以下勞工資料的分析。

2.2 最鄰近搜索法

在透過 MDS 將原始資料降到低維度的空間座標後，我們還需要作資料的分群；以下介紹最鄰近搜索法 (Nearest Neighbor Search, NNS)。

NNS 的概念相當簡單，就是在 r 維度的空間裡的一個點集合 S 中，搜索離一個給定的目標點集合 Q 最靠近的點 s；找到 s 後，將 s 加入 Q 集合中，如此重複疊代運算，便能將原始的點集合 S 作分群。以下一章的勞工資料爲例，我們觀察三維空間裡點分佈的情形後，判斷資料應該分成三群 (R, B, G)，再給定各群一些起始點，在計算全部資料點到這些起始點的距離後，記錄下離各啓始點最短的距離與位置，再將最短距離的點合併到該起始點中，如此重複地運算，直到全部資料都完成分群爲止。

3 美國當期人口調查的多元尺度分析

3.1 資料收集與整理

我們自 ceprDATA 取得 2010 年美國當期人口調查 (Current Population Survey, CPS) 的勞工資料; CPS 是由美國人口普查局 (United States Census Bureau) 執行的統計調查資料，其內容包含:年齡、性別、種族、教育程度、婚姻狀況、撫養小孩人數、是否是市民身份、出生地、居住洲名、勞動力狀況、就業狀況、產業、職位、領取薪資方式等多種基本資料。

選取年齡、性別、種族、教育程度、居住州名、出生國家這六筆資料，並將年齡分爲: 青年 $(16-30$ 歲)、壯年 $(31-44$ 歲)、中年 $(45-64$ 歲)與老年 (65 歲以上)，共四種級距，並且將教育程度區分爲: 高中或高中以下 (以下簡稱高中程度)、大學或技職學校 (以下簡稱大學程度)、碩士或專業學院 (以下簡稱碩士程度)、與博士，共四個等級 [4]。此外，有四種種族 [5]，五十一個州名及一百四十九個國家。在刪除產業與職位有遺失的資料後，得到 $106,711$ 筆勞工資料，每筆資料分爲以上六大類共 213 個欄位，所以這是一筆$106,711 \times 213$ 的大資料，而且符合 SC-MDS 對資料的型態的設定 $(N >> p)$。

3.2 SC-MDS 的三維視圖

將這 $106,711 \times 213$的大資料矩陣作 SC-MDS 後得到圖 (3.1)的三維空間配置圖。可以很明顯地由圖 (3.1)中觀察出，這十萬多名勞工大致可分爲四大群，每群又分三組，每組又分兩小組。

[4]CPS 的教育程度一共有 16 種分類，我們將 Less than 1st grade、1st-11th grade、 12th grade-no diploma、 HS graduate GED 分類於高中或高中以下； Some college but no degree 、 Associate degree-occupational/vocational 、 Associate degree-academic program、 Bachelor's degree 分類於大學或技職學校; Master's degree ， Professional school 分類於碩士或專業學院; 而 Doctorate 分類於博士

[5]CPS 的種族分有白人(white)、黑人(black)、西班牙裔美國人(hispanic)與其他(other)四種。

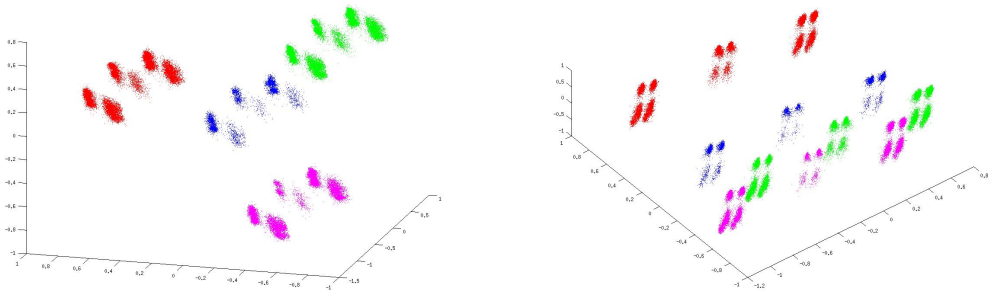
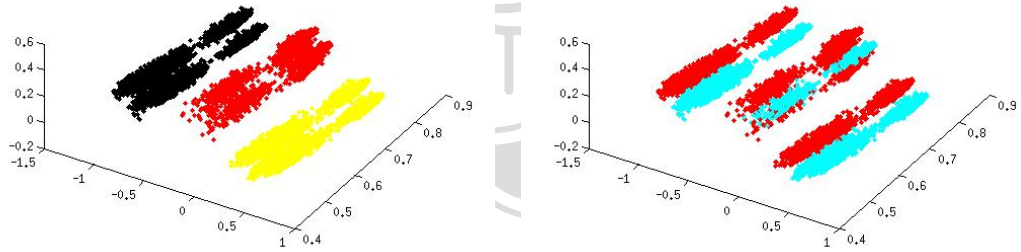


圖 3.1: 將 CPS 資料作 SC-MDS 後得到的三維空間配置圖。
各顏色所代表的是：紫色爲青年、綠色爲壯年、紅色爲中年和藍色爲老年的勞工。



(a) 在 45 – 64 歲中，不同教育程度在分群中的相對位置。黑色爲高中或高中以下，紅色爲碩士與博士程度，黃色爲大學程度。

(b) 在 45 – 64 歲中，男性與女性在分群中的相對位置。紅色爲男性，青色爲女性。

圖 3.2: NNS 分群後，教育程度與性別在分群中的相對位置

圖(3.2)表示透過 NNS 的分群後，我們可以得到圖(3.1)中的四大群分別爲：青年 (16 – 30 歲)、壯年 (31 – 44 歲)、中年 (45 – 64 歲)與老年 (65 歲以上)；而圖(3.2.a)表示各群中的三組分別爲：高中程度、大學程度與碩、博士程度；最後，圖(3.2.b)表示各組中的兩小組分別爲男性、女性。

MDS 透過歐氏空間點與點的距離來表達資料間的相似度，所以藉由以上的分析可以繪製此大資料的樹狀圖，圖(3.3)；換句話說，我們從這原始資料的六大類共 214 個欄位中，找到了一個清楚的分析架構。

在大資料分析中常談到「 Np hard 」的問題，因爲每筆資料都包含數個分類，在不知道各分類間的重要性或優先順序的情況下，分析大資料有如大海撈針；然而，透過 SC-MDS ，使我們可以抽絲撥繭地抓出各個分類的重要性，並加以排列優先順序，如此就像是建構一張大資料的地圖，讓分析者能按圖索驥從大資料中找到其隱藏的資訊。以下，我們將藉由這樣的勞工基本資料的樹狀圖，結合勞工特質的資料，想進一步分析不同群組間，勞工在特質 (characteristics) 上的差異。

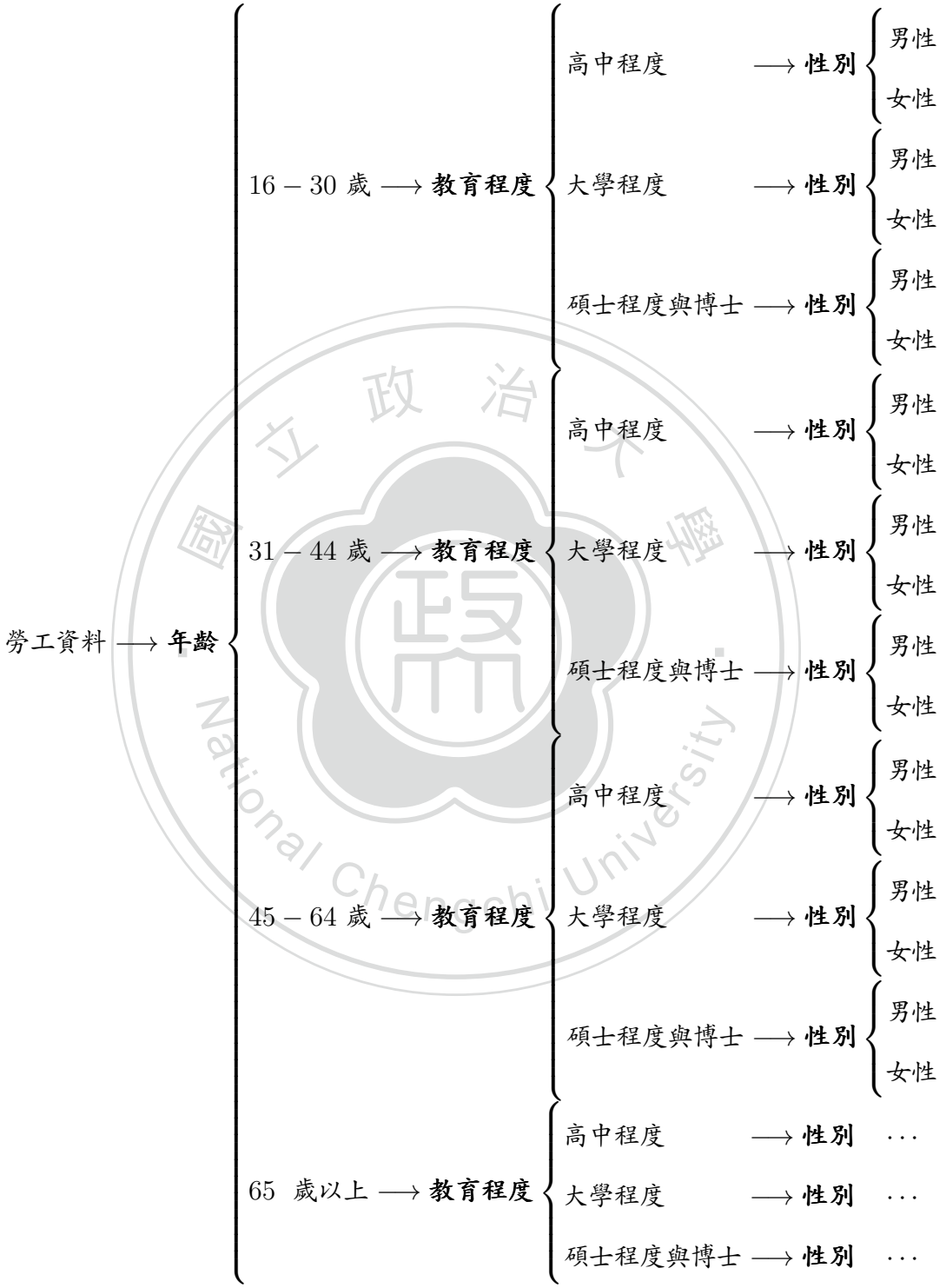


圖 3.3: 勞工資料的樹狀圖結構

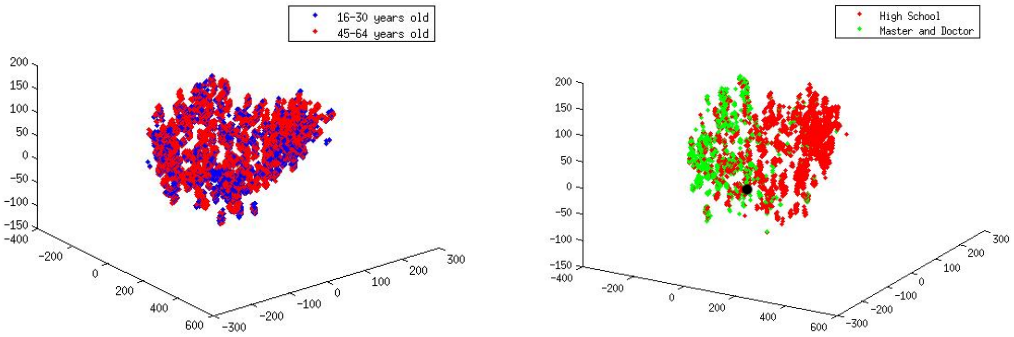
4 結合職位特質資料

4.1 資料的收集與整理

職位資訊聯合資料庫 (Occupational Information Network, O*NET) 是一個由美國勞工局就業與訓練部所設立的免費線上資料庫，O*NET 設立的目的是想透過刻劃上千個職位 (occupation) 特質去協助減少應徵者與雇主的媒合 (matching) 成本。O*NET 從能力 (abilities)、興趣 (interests)、知識 (knowledge)、技能 (skills)、工作活動 (work activities)、工作內容 (work context) 與工作價值 (work vaulues) 去評估一個職位該由具備什麼樣特質的勞工來擔任，各個項目的衡量又包含了重要性 (importance) 與等級 (level) 兩種。同樣的技能對於很多職位都相當重要，如能否清楚地向他人傳達資訊這類的說話能力；但對於同樣的技能，每個職位所要求的等級就有很大的不同。以說話技巧爲例，說話的技巧對於律師與律師助理而言都是相當重要的技能，然而律師常常需要在法庭上作辯論，但律師助理往往都是從事幕後的工作，所以律師的說話能力在等級上則是會高於律師助理。

我們選取有職位等級的項目：能力、知識、技能與工作活動，並透過職位與 CPS 資料結合，來觀察這 106,711 名勞工所具有的職位特質有何差別。此外，對於等級的遺失值，是以其它重要性相同的職位的等級取平均後替代。接著因爲 O*NET 將職位細分爲 1,086 個職位別，但 CPS 卻只有 509 個職位別，但職位別的差異只是在於 O*NET 分的較細，所以我們將對應到相同 CPS 職位別的 O*NET 等級 [6] 取平均後，作爲該 CPS 職位別的等級分數。舉例來說，在賭場工作的發牌員 (game supervisors) 與吃角子老虎的服務員 (slot supervisors) 在O*NET 分別爲不同的職位別，但 CPS 將這兩者都歸類於第一線賭場工作人員 (first-line supervisors/managers of gaming workers)，所以在我們合併資料時，就將發牌員與吃角子老虎的服務員的等級分數取平均，作爲 CPS 中第一線

[6]對應的職位分類，參考美國勞工局網站:
ftp://ftp.bls.gov/pub/special.requests/ep/classification.crosswalks



(a) 青年與中年勞工　　(b) 中年男性勞工具高中程度或碩、博士程度

圖 4.1: O*NET 資料藉 SC-MDS 得到的三維視圖

賭場工作人員的分數。

O*NET 資料的內容中，能力分爲認知能力 (cognitive abilities)、體能能力 (physical abilities)、心理動作能力 (psychomotor abilities) 與感官能力 (sensory abilities)，共 52 項；知識共 33 項；技能分爲基本技能 (basic skills)、解決複雜問題技能 (complex problem solving skills)、資源管理技能 (resource management skills)、社交技能 (social skills)、系統性分析與決策技能 (system skills)與技術技能 (technical skills)，共 35項；工作活動分爲資訊吸收 (Information Input)、與人互動 (interacting with others)、心理過程 (mental processes)與工作產出 (work output)，共 41 項 [7]。因此，在結合 CPS 與 O*NET 資料庫後，我們可以得到一個 $106,711 \times 161$ 的資料矩陣；同樣地，使用 SC-MDS 作降低維度的方法。

4.2 職位等級資料 SC-MDS 的三維視圖

首先，由圖(4.1a與b)中觀察在 SC-MDS 後的三維視圖中勞工特質資料分群效果並不明顯，所以我們列出勞工特質資料在 SVD 後的特徵值。由圖(4.2)中可

[7]詳細的項目列表，請參考附錄

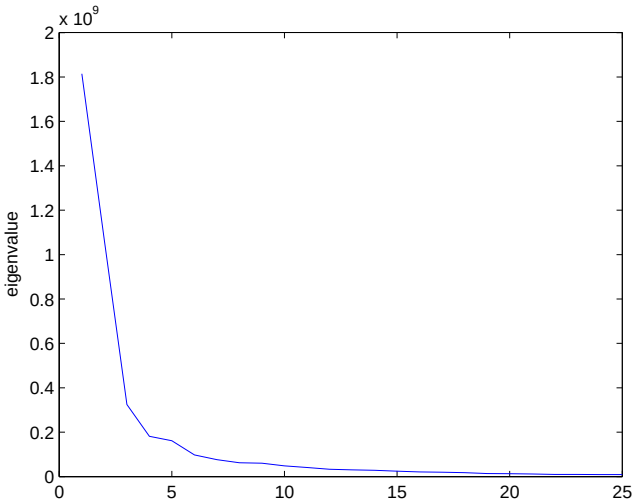


圖 4.2: 勞工特質資料在 SVD 後前 25 項的特徵值

以發現，在前三項爲最大的主項且快速地遞減；勞工特質資料在 SVD 後的前三項特徵值佔前 250 個特徵值的比重爲 73.54 %；相較於圖(3.1)的勞工基本資料在 SVD 後的前三項特徵值佔前 250 個特徵值的比重爲 36.36%，但圖(3.1)就已很清楚地分群，所以我們可以相信在三維度的歐氏空間中作 MDS 的空間配置是適當的。

圖(4.1a)爲青年與中年勞工的 O*NET 資料的 SC-MDS的結果，我們可以發現，在這兩個年紀階層中，在ONET資料中是非常靠近的，像是相互糾纏在一起；這樣的結果是很合直覺的，因爲在這兩個年齡層所擔任的職位可能會有所不同，但大部份的職位仍是有可能重複；譬如，擔任郵局的櫃檯服務員，可能是剛出社會的大學新鮮人，也有可能是已經快退休的資深員工。

圖(4.1b)爲中年的男性勞工具高中程度與碩、博士的 O*NET 資料的 SC-MDS的結果。可以發現一個有趣的現象，在最高與最低的這兩種教育程度中，在擔任職位的特質有明顯地分開；這結果雖然也符合直覺，因爲只受高中程度教育程度的勞工可能只能擔任較爲基層的工作，但擁有碩、博學位的勞工則較有機會擔任管理職位，因此這兩種職爲其所需具備的特質是有所差異的。此外，在

圖(4.1b)中，有部份勞工相互交錯，這也可能是因爲在職場的發展較不受教育程度限制。以鴻海企業的郭台銘董事長爲例，他只有高職學歷，但職掌兆元企業；相反地，日前出現碩、博士生應徵基層清潔人員的工作，如此教育程度與職位不相稱的社會現象便是圖(4.1b)所反應的。

4.3 以 SC-MDS 樹狀圖分析勞工特質差異

本節我們將採用上章所繪製的勞工分群樹狀圖，圖(3.3)，分析各個群組間的勞工，在職位特質上有何差異；爲了指出不同群組間職位特質的差異項目，我們採用獨立樣本的 T 檢定：T 檢定是用以檢定兩群體特性的期望值是否相等的統計方法。[8]

男性與女性在勞工特質上的差異

表(4.1)中爲在不同教育程度中性別對勞工特質的差異。我們可以發現，兩點有趣的現象：第一，受過高等教育的勞工，性別在勞工特質差異數目上比僅受高中程度少。以 $16-30$ 歲這個年齡層爲例，當勞工僅受過高中程度的教育時，男性與女性在技能上有高達 75% (27/35)以上的特質不同；然而，在擁有碩、博士學歷的勞工中，卻只有平均只有 20% (7/35)的項目有顯著差異。第二，雖然勞工受過高等教育，但隨年紀的增長，勞工特質的差異數又會明顯倍增。以擁有碩士程度學位的勞工爲例，在 45 歲以前，每當年紀增加一個層級，勞工特質差異數就會倍增。

這是個很有趣的現象，我們對以上兩個現象的解釋爲：第一，因爲在高中程度，都是接受普通或通才教育，但卻在勞工特質上有如此多的項目具顯著差異。對此，我們的解釋爲因爲僅受過高中程度教育的勞工，在投入就業市場後大多會從事更仰賴專業技能的工作，而這些工作大多與性別有關。舉例來說：搬運工人與保姆；搬運工人往往不需要較高的學歷，反而需要較強壯的生理特質，男性就

[8]以下的分析，我們設定 T 檢定的 p-value 爲 0.01。

較容易符合這職位的需求；相反地，保姆這項職位也不需要勞工擁有太高的教育水準，反倒是需要細心與耐心的心理特質，女性就較男性容易擁有這些特質。換句話說，在教育水準不高的情況下，勞工往往會選擇與自己的天生特質較符合的職位，並在日後培養出完全不同的職業特質。

年齡	能力 (52)		知識 (33)		技能 (35)		工作活動 (41)		項目總合 (161)	
16 – 30 歲	34	3	25	7	27	7	33	8	119	25
31 – 44 歲	34	13	36	23	30	12	25	14	125	62
45 – 64 歲	48	26	30	26	32	30	36	30	146	121
65 歲以上	27	28	51	28	32	25	34	28	144	115
	高中	碩博	高中	碩博	高中	碩博	高中	碩博	高中	碩博

表 4.1: 性別對勞工特質的差異

第二，對於受過高等教育的勞工身上，可以發現勞工特質差異數隨年齡倍增的情形，這可以解釋爲女性在家庭與事業上的選擇。一般來說，女性都較容易受家庭因素影響，如生育行爲。勞工本身所受的教育程度的高低會弱化就業後職位對生理特質的依賴性；舉例來說，一對同時考取到律師執照的夫妻分別都在法律事務所擔任律師的職位，這個職位並沒有男女之分，只賴於當初在法學院所接受的教育與訓練。但當妻子懷孕、生產時，妻子較容易離開就業市場。當這名妻子多年後再重回就業市場，其所能擔任職位便與一直都在職場上工作的丈夫在職位特質上有很大的差異。

不同教育程度間勞工特質上的差異

表(4.2)爲在壯年的勞工中，比較不同教育程度間勞工特質有顯著差異的項目。我們可以發現教育水準對於勞工在能力、知識、技能與工作活動上都有極大的不同，且又以高中程度與碩、博士間的差異最多；大學與碩、博士的差異最少。以技能爲例，表(4.2)中顯示，在男性壯年的勞工中，不同的教育程度，所擔任職位的特質完全不同。換句話說，不同的教育程度，會影響勞工在職場上擔任的職位。

教育程度	性別	能力 (52)	知識 (33)	技能 (35)	工作活動 (41)	項目總合 (161)
高中程度與大學程度	男性	49	26	35	36	146
	女性	47	26	32	37	142
高中程度與碩、博士程度	男性	50	30	35	41	156
	女性	50	26	32	40	148
大學程度與碩、博士程度	男性	45	24	34	40	143
	女性	39	23	25	33	120

表 4.2: 不同教育程度間壯年勞工在勞工特質上的差異

不同年齡層級間勞工特質上的差異

表(4.3)- (4.5)爲男性勞工在青年 $(16-30$ 歲$)$ 與壯年 $(31-44$ 歲$)$年齡層間(以下簡稱青壯年) 和壯年 $(31-44$ 歲$)$ 與中年 $(45-64$ 歲$)$ 年齡層間 (以下簡稱爲中壯年) 在特質中顯著差異的項目。由表(4.3)- (4.5)中可以發現兩個重點：第一，青壯年較中壯年在勞工特質上較爲相似；第二，具碩博士教育程度的勞工其勞工特質上較其它教育程度的勞工差異較少。對此，我們的解釋如下：第一，成長過程、教育環境與產業發展。在近 20 年的現代社會發展，無論是科技的進步、觀念的創新甚至學習過程中差異，當時中年人所處的社會背景已與壯年人

或青年人有著天壤之別。以現今一名 50 歲的中年人爲例，在 30 年前所處的教育環境中並沒有網路、社群甚至蘋果電腦；然而對現今一名 35 歲的勞工而言，在他 20 歲所處的社會中，科技的發展已經主導社會的前進，並且在他初入社會時，產業的變動更快，新的觀念如湧泉而出，這些社會背景影響著勞工所具備的特質，以就產生了中壯年勞工在特質上較青壯年的勞工差異較多。第二，在碩博士的教育過程中，都是在專精於各個領域的學習與訓練；因此，在擁有碩博士教育程度的勞工除了在特定的領域一定學有所長，在其具備的勞工特質上也會有所不同；譬如說，經濟博士與物理博士，他們都在社會科學與自然科學中擁有特定的專長，但在取得博士學位前他們都需擁有獨自研究的能力外，在取得學位過程中的種種考驗也會使這兩名不同領域的博士生擁有類似的特質。以下，我們分別藉由能力、技能與工作活動來逐步地討論各個年齡層間的勞工特質差異。

	能力									
	認知 (21)		體能 (9)		心理活動 (10)		感官 (12)		項目總合 (52)	
高中程度	21	13	9	9	10	10	12	8	52	40
大學程度	20	11	8	9	10	10	9	6	47	36
碩、博士程度	19	5	0	0	0	4	4	5	23	14
	31-44 vs 45-64	31-44 vs 16-30	31-44 vs 45-64	31-44 vs 16-30	31-44 vs 45-64	31-44 vs 16-30	31-44 vs 45-64	31-44 vs 16-30	31-44 vs 45-64	31-44 vs 16-30

表 4.3: 男性勞工在年齡層間的能力特質差異類表。

首先，在表(4.3)所呈現的是能力方面的比較。在認知與感官能力中，青壯年的差異數都較少。但是，較爲特別的是在擁有碩、博士學位的勞工中，心理活動則是中壯年較爲相似，反而是青壯年有顯著差異，其項目是：四肢協調能力 (multilimb coordination)、節奏感 (rate control)、反應時間 (reaction time)、反應方向 (response orientation) [9]， 這四項都神經與四肢的協調性有

[9]在心理活動 (mental process) 的衡量上，O*NET 還有手部的穩定性 (arm-hand steadi-

關，所以我們能推測雖然在擁有碩、博士學位的勞工，雖然擔任較高等的職位，但還是需要親手實作以累積管理經驗。舉例來說， 在物流業擔任儲備幹部的社會新鮮人，需要在倉儲中學習作業流程，甚至親手搬運貨物以體會第一線工作人員的工作量，進而快速累積經驗；相較於擔任各倉儲的經理來說，因爲已累積了足夠的經驗，所以他們的工作則是在辦公室去設計或改善作業流程。同樣的道理也反應在體能的表現上，擁有碩、博士學歷的勞工無論年齡在體能能力的表現是相同的；這可以解釋爲碩、博士在就業市場中往往擔任較高等的職位，而對於這些職位在體能的要求並不高，所以當隨年紀增長時，仍可符合該職業的要求。

	技能							
	基本技能 (10)		解決複雜問題技能 (1)		資源管理 (4)			
高中程度	9	3	1	0	3	2		
大學程度	10	2	1	0	4	2		
碩、博士程度	9	1	1	0	4	0		
	社交技能 (6)		系統性分析與決策 (3)		技術技能 (11)		項目總合 (35)	
高中程度	3	3	3	2	11	10	30	20
大學程度	6	3	3	2	11	9	35	18
碩、博士程度	2	0	3	0	4	6	23	7
	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31

表 4.4: 男性勞工在年齡層間的技能特質差異類表。

接著，表(4.4)所呈現的是技能方面的比較：青壯年的男性勞工除了擁有碩、

ness)、肢體控制的精準度 (control precision)、手指的靈活度 (finger dexterity)、手指間的協調度 (manual dexterity)、四肢的移動速度 (speed of limb movement) 與 手腕與手指的移動速度 (wrist-finger speed)。

博士學位的勞工在技術技能外，其它技能特質的差異數都明顯小於中壯年的男性勞工；特別是基本技能與解決複雜問題技能這兩項。有趣的是，青壯年且擁有碩、博士學位的男性勞工在職位的基本技能上只有數學這項不同，而其它基本技能卻沒有顯著差異[10]，甚至在解決複雜問題技能、資源管理、社交能力與系統性分析與決策上也一樣沒有顯著差異；但相較於中年與壯年且受相同教育程度的男性勞工卻不同。這反應了以上的技能在受碩、博士教育的男性勞工在中年與壯年間有一個很明顯的落差，這樣落差的形成我們解釋爲職場經驗的累積。以富邦金融控股公司爲例，在該企業 2010 年「企業社會責任報告書」中顯示，其全體董事平均年齡爲 60 歲，且 12 位董事中僅有兩名不具有碩士、專業學院或博士學位；此外，在 502 名男性管理人才中，30 歲以下僅有一名，31 – 49 歲有 253 名，而 50 歲以上有 248 名[11]。這例子顯示受較高等教育的男性勞工若要擔任企業中核心的職位，是需要資歷上的累積；而這些核心職位，對於各項技能的要求就更高了，藉此產生了中年與壯年間技能上的落差。

最後，表(4.5)所呈現的是工作活動的比較：其中較爲特別的是工作產出這個部份[12]。在 9 項衡量工作產出的部份，青壯年的勞工在與電腦互動和操作機械裝置這兩項有顯著差異，但在中壯年的勞工卻只有與電腦互動上有顯著差異。對此的解釋如同能力的比較，因爲青年的勞工能藉親手操作而快速累積經驗，所以

[10]在基本技能的衡量上，O*NET 分爲主動學習 (active learning)、聆聽他人說話 (active listening)、關鍵思考 (critical thinking)、學習策略 (learning strategies)、數學 (mathematics)、監控 (monitoring)、閱讀 (reading comprehension)、科學 (science)、說話 (speaking)、寫作 (writing)，共 10 項。

[11]資料來源：「富邦金控 2010 年企業責任報告書」。

[12]在工作產出的衡量上，O*NET 分爲機器與流程控管 (controlling machines and processes)、資訊記錄(documenting, recording information)、設計與具體需使用的設備儀器 (drafting laying out and specifying technical devices parts and equipment)、物體的搬移 (handling and moving object)、與電腦互動 (interacting with computers)、操作機械裝置 (operating vehicles mechanized devices or equipment)、生理活動表現 (performing general physical activities)、保養與維修電子設備 (repairing and maintaining electronic equipment)、保養與維修機械設備 (repairing and maintaining mechanical equipment)，共 9 項。

在操作機械裝置上，與壯年的勞工有顯著差異。此外，在與人互動和心理過程這兩項工作活動上，我們也可發現無論教育背景，中年與壯年的勞工有較多顯著差異的項目。對此我們解釋爲較年長者在待人處事與分析事務上累積了較多的經驗，進而產生這樣的落差。

	工作活動									
	資訊吸收 (5)		與人互動 (17)		心理過程 (10)		工作產出 (9)		項目總合 (41)	
高中程度	5	5	11	10	10	6	8	8	34	29
大學程度	5	5	16	7	10	4	9	5	40	21
碩、博士程度	4	4	9	4	9	5	1	2	18	15
	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31	31-44 vs 45-64	31-44 vs 16-31

表 4.5: 男性勞工在年齡層間的工作活動特質差異類表。

5 結論

隨科技的進步，無論是生物資訊、消費行爲或是勞工資料都能被詳細地記錄；然而，資料的價值來自於其中所包含的資訊，但資訊該如何被挖掘出來，便成了目前各個領域中爭相討論的議題。在本文的勞工資料中，我們遇到兩個傳統方法無法解決的問題：第一個是大資料的分析，傳統上的方法在電腦科學上無法處理如此大量的資料；第二是統計的方法論，一般序述統計的方式無法清楚地找出此資料的特性，所以需要使用分解與合併 MDS 爲我們找出全部資料的分析架構。MDS 爲我們建立一個從分析勞工資料的分層架構，分別是年齡、教育程度、性別。並根據這個架構分析勞工在職位上的特質上的差異，並發現了教育程度會影響性別間在勞工職位特質上的差異，且這些差異的數量又會隨年齡的增加而增加；教育程度在各個年齡層都對勞工職位特質產生很大的差異；最後，青年與壯年的勞工在職位特質上相較於壯年與中年勞工相似。

相似性高的勞工對經濟社會的發展是否是好的?從前述的討論中我們可以發現美國高等教育中所培養的人才都具有很高的相似性，而相似性的高等技術勞工過多，往往會造成人才供給過剩，進而使該勞動就業市場薪資降低或排擠低技術勞工，甚至當本國勞動市場無法消化人才時會產生高技術勞工外移，產生高成低就的社會現象；而當低技術勞工受到排擠且不願再將其薪資向下調整時，此時顧主往往會藉由引進外籍勞工解決低技術勞工市場失衡的問題。高等教育無法適量地培養就業市場所需的勞工，進而產生就業市場的波動，甚至引發人口的遷移，所以就現今經濟社會快速變化的 21 世紀中，政府當局不應再將人才專精化、制式化，而是培養更多元、更具差異與彈性的勞工。很慶幸台灣這幾年的高等教育所提倡的多元化、多面化的施教方針，如國立交通大學設立客家文化學院、國立政治大學設立應用物理系與眾多大專學院開設跨領域課程等，這些將有助於增加未來台灣勞工在就業市場的競爭力。

此外， 我們也建議勞工局能參考美國職位資訊聯合資料庫去進行一個台灣勞工的職位特質分析，讓教育政策制定者能就當下或是未來的教育方針上能有個更具體的衡量指標。

參考文獻

Bellman, Richard Ernest (1957), *Dynamic Programming*, Princeton : Princeton University Press.

Chalmers, M. (1996), "A linear iteration time layout algorithm for visualising high – dimensional data", *IEEE Visualization*, 127–132.

Cox, Trevor F. and Cox, Michael A. A. (2001), *Multidimensional scaling*, London : Chapman & Hall, 2 edition.

Dasgupta, S. and Gupta, A. (1999), "An elementary proof of the johnson-lindenstrauss lemma", Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA.

Dempster, Arthur, Laird, Nan, and Rubin, Donald (1977), "Maximum likelihood from incomplete data via the em algorithm", *Journal of the Royal Statistical Society, Series B*, 39(1), 1–38.

Dwyer, Tim and Gallagher, David R. (2004), "Visualising changes in fund manager holdings in two and a half dimensions.", *Information Visualization*, 3, 227–258.

Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1992), "Knowledge discovery in databases: An overview", *AI Magazine*, 213–228.

Groenen, Patrick J.F. and Franses, Philip Hans (2000), "Visualizing time-varying correlations across stock markets.", *Journal of Empirical Finance*, 7, 155–172.

Johnson, W.B. and Lindenstrauss, J. (1984), "Extensions of lipshitz mapping into hilbert. space", volume 26, 189–206, In Conference in modern analysis

and probability, volume 26 of Contemporary Mathematics, Amer. Math. Soc.

Knuth, Donald E. (1973), *The art of computer programming*, Boston, Mass. : Addison-Wesley.

Kruskal, J.B (1964), "Nonmetric multidimensional scaling: a numerical method.", *Psychometrika*, 29, 115–129.

Lloyd, S. P. (1957), "Least square quantization in pcm", *Bell Telephone Laboratories Paper*.

Lloyd, S.P. (1982), "Least squares quantization in pcm", *IEEE Transactions on Information Theory*, 28(2), 129–137.

Morrison, Alistair, Ross, Greg, and Chalmers, Matthew (2003), "Discussion of a set of points in terms of their mutual distances.", *Information Visualization*, 2, 68–77.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space.", *Philosophical Magazine*, 2(6), 559–572.

Shepard, R.N. (1962), "The analysis of proximities: Multidimensional scaling with an unknown distance function.", *Psychometrika*, 27(2), 125–140.

Torgerson, Warren (1952), "Multidimensional scaling: I. theory and method", *Psychometrika*, 17, 401–419.

Tzeng, Jengnan, Lu, Henry Horng-Shing, and Wen-Hsiung, Li (2008), "Multidimensional scaling for large genomic data sets.", *BMC Bioinformatics*, 9, 1 – 17.

White, Tom. (2009), *Hadoop: The Definitive Guide*, O'Reilly Media, 1 edition.

Young, G.W and Householder, A.S (2003), "Fast multidimensional scaling through sampling, springs and interpolation.", *Information Visualization*, 2, 68–77.

附錄

Abilities	Cognitive Abilities (21)	Category Flexibility
		Deductive Reasoning
		Flexibility of Closure
		Fluency of Ideas
		Inductive Reasoning
		Information Ordering
		Mathematical Reasoning
		Memorization
		Number Facility
		Oral Comprehension
		Oral Expression
		Originality
		Perceptual Speed
		Problem Sensitivity
		Selective Attention
		Spatial Orientation
		Speed of Closure
		Time Sharing
		Visualization
		Written Comprehension
		Written Expression
	Physical Abilities	Dynamic Flexibility
		Dynamic Strength

Abilities	(9) Physical Abilities (9)	Explosive Strength Extent Flexibility Gross Body Coordination Gross Body Equilibrium Stamina Static Strength Trunk Strength
	Psychomotor Abilities (10)	Arm-Hand Steadiness Control Precision Finger Dexterity Manual Dexterity Multilimb Coordination Rate Control Reaction Time Response Orientation Speed of Limb Movement Wrist-Finger Speed
	Sensory Abilities (12)	Auditory Attention Depth Perception Far Vision Glare Sensitivity Hearing Sensitivity Near Vision Night Vision

Abilities	Sensory Abilities (12)	Peripheral Vision Sound Localization Speech Clarity Speech Recognition Visual Color Discrimination
Knowledge (33)	Administration and Management Biology Building and Construction Chemistry Clerical Communications and Media Computers and Electronics Customer and Personal Service Design Economics and Accounting Education and Training Engineering and Technology English Language Fine Arts Food Production Foreign Language Geography History and Archeology Law and Government	

Knowledge (33)	Mathematics Mechanical Medicine and Dentistry Personnel and Human Resources Philosophy and Theology Physics Production and Processing Psychology Public Safety and Security Sales and Marketing Sociology and Anthropology Telecommunications Therapy and Counseling Transportation	
skills	Basic Skills (10)	Active Learning Active Listening Critical Thinking Learning Strategies Mathematics Monitoring Reading Comprehension Science Speaking Writing

skills	Complex Problem Solving Skills		
	Resource Management (4)	Management of Financial Resources Management of Material Resources Management of Personnel Resources Time Management	
	Social Skills (6)	Coordination Instructing Negotiation Persuasion Service Orientation Social Perceptiveness	
	Systems Skills (3)	Judgment and Decision Making Systems Analysis Systems Evaluation	
	Technical Skills (11)	Equipment Maintenance Equipment Selection Installation Operation and Control Operation Monitoring Operations Analysis Programming Quality Control Analysis Repairing Technology Design	

		Troubleshooting
Work Activities	Information Input (5)	Estimating the Quantifiable Characteristics of Products Events or Information Getting Information Identifying Objects Actions and Events Inspecting Equipment Structures or Material Monitor Processes Materials or Surroundings
	Interacting With Others (17)	Assisting and Caring for Others Coaching and Developing Others Communicating with Persons Outside Organization Communicating with Supervisors Peers or Subordinates Coordinating the Work and Activities of Others Developing and Building Teams Establishing and Maintaining Interpersonal Relationships Guiding Directing and Motivating Subordinates Interpreting the Meaning of Information for Others

		Monitoring and Controlling Resources
		Performing Administrative Activities
		Performing for or Working Directly with the Public
		Provide Consultation and Advice to Others
		Resolving Conflicts and Negotiating with Others
		Selling or Influencing Others
		Staffing Organizational Units
		Training and Teaching Others
Work Activities	Mental Processes (10)	Analyzing Data or Information
		Developing Objectives and Strategies
		Evaluating Information to Determine Compliance with Standards
		Judging the Qualities of Things Services or People
		Making Decisions and Solving Problems
		Organizing Planning and Prioritizing Work
		Processing Information
		Scheduling Work and Activities
		Thinking Creatively
		Updating and Using Relevant Knowledge

	Work Output (9)	Controlling Machines and Processes Documenting Recording Information Drafting Laying Out and Specifying Technical Devices Parts and Equipment Handling and Moving Objects Interacting With Computers Operating Vehicles Mechanized Devices or Equipment Performing General Physical Activities Repairing and Maintaining Electronic Equipment Repairing and Maintaining Mechanical Equipment