國立政治大學應用數學系

碩士學位論文

台灣 1950 年至 2010 年出版社演變情況之資料視覺化方法

Data visualization method of Taiwan publishing house evolution from 1950 to 2010

國立政治大學
National Chengchi University

指導教授：曾正男　博士

研究生：吳昱辰　撰

中 華 民 國 110 年 7 月

DOI:10.6814/NCCU202101092

致謝

首先，很高興能在三年前考進政治大學應用數學系，這三年期間同學與老師都給了我很大的幫助，在第二年的時候曾正男老師確定成為我的指導教授，從開始跟著老師後在老師身上學到很多東西，不管是學術上的還是生活上的，老師都會給予指導及建議，也非常感謝老師在我撰寫論文期間給予的指導，撰寫的過程中很多學術上的問題都是老師撥冗與我慢慢討論後，接著進行下一步，才能如期在碩三這年結束撰寫論文的過程。

在修課上面印象最深的就是實變函數論，非常感謝陳天進老師的教學，每堂課都能收穫滿滿的筆記，也是這堂課讓我對於數學的能力更加進步，私下老師也很照顧學生，常常會找大家吃飯，聊一些學習上或生活上的事情，也是經由這些過程才與老師慢慢熟悉，到現在要畢業了回想起來都是珍貴的回憶。

最後，不免俗地還是要感謝家人，畢竟家人一直是我堅強的後盾，沒有家裡支持我沒辦法在這三年取得現在的收穫，而在學校中，感謝系辦的助教在這三年的照顧，感謝林澤佑學長在看見我編譯程式的時候主動過來教我一些技巧，當時教的技巧很多都很實用，還有羅令宇同學，從大學就一直在一個班，感謝他在這段時間對我的幫助，感謝同屆的同學們，大家對我的照顧，謝謝這三年中對我照顧的所有人。

DOI:10.6814/NCCU202101092

中文摘要

多重對應分析早期應用於應用於語言學研究，主要是利用該方法將獲得資料進行降維分析，讓資料可以呈現在二維或三維空間，透過視覺就能對資料進行解讀。本文探討的是多重對應分析是否真的能夠將資料原始型態在降維後呈現，實驗中的檢定方式會利用到特徵值去觀察，我們所使用的資料是台灣 1950 至 2010 年出版社資料，將討論幾種視覺化分析方法所分析出來的內容差異性，經實驗觀察多重對應分析在此資料的研究中出現解釋率不足的問題，因此再比較多元尺度分析及潛在語義索引兩種方法，另外兩種方法在解釋率上都有明顯的提升，而潛在語義索引是三種方法中結果表現最為突出的。

關鍵字: 多重對應分析、多維尺度分析、潛在語義索引、1950~2010 出版社演變

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

Abstract

The early application of multiple correspondence analysis in linguistic research was mainly to use this method to perform dimensionality reduction analysis on the obtained data, so that the data can be presented in two-dimensional or three-dimensional space, and the data can be interpreted through vision. This article is discussing whether multiple correspondence analysis can really present the original form of the data after dimensionality reduction. The verification method in the experiment will use eigenvalues to observe. The data we use are Taiwanese publishing houses from 1950 to 2010. The content differences analyzed by several visual analysis methods will be discussed. After experimental observation, multiple correspondence analysis has the problem of insufficient interpretation rate in the research of this data. Therefore, we will compare the two methods of multi-scale analysis and latent semantic indexing. Both methods have a significant improvement in interpretation rate, and latent semantic indexing is the most prominent result among the three methods.

Keywords: multiple correspondence analysis, multidimensional scale analysis, latent semantic indexing, 1950～2010 publishing house evolution

DOI:10.6814/NCCU202101092

目錄

致謝 ii
中文摘要 iii
Abstract iv
目錄 v
表目錄 vii
圖目錄 viii
第一章 緒論 1
第一節 研究背景 1
第二節 研究動機 2
第二章 文獻回顧 3
第一節 MCA 多重對應分析 3
一、對應分析介紹 3
二、多重對應分析介紹 3
三、多重對應分析理論 4
第二節 MDS 多維度分析 4
一、多維度分析背景介紹 4
二、多維度分析理論 5
第三節 潛在語義索引 10
一、潛在語義索引介紹 10
二、潛在語義索引方法 11
第三章 資料描述 12
第一節 資料來源 12
第四章 問題描述 15
第一節 MCA 的侷限性 15

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

第五章 研究方法 18
第一節 MDS 距離定義過程 18
第六章 實驗結果 23
第一節 MDS 資料分析圖 23
第二節 LSI 資料分析圖 36
第七章 結論 38

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202101092

表目錄

表 2.2.1　城市相對距離表格 (單位: 公里) 6
表 2.2.2　原始出版社資料數量 10
表 2.3.3　原始出版社資料數量 11
表 3.1.1　原始出版社資料數量 13
表 4.1.1　MCA 解釋率一覽表 17
表 5.1.1　出版社出版文類統計表 19
表 5.1.2　出版社相對距離表格 20
表 5.1.3　距離算法式 (5.1)，$w_1 = 1$、$w_2 = 1$ 21
表 5.1.4　距離算法式 (5.2)，$w_1 = 1$、$w_2 = 1$ 21
表 5.1.5　距離算法式 (5.1)，$w_1 = 0.9$、$w_2 = 0.7$ 22
表 5.1.6　距離算法式 (5.2)，$w_1 = 0.9$、$w_2 = 0.7$ 22

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

圖目錄

圖 2.2.1　多維度分析流程圖 9

圖 3.1.1　出版數量 1 ～ 6 本出版社分配圖 14

圖 4.1.1　利用 Python 分析 1950 ～ 1970 出版社分群情況 15
圖 4.1.2　利用 Python 分析 1971 ～ 1990 出版社分群情況 16
圖 4.1.3　利用 Python 分析 1991 ～ 20100 出版社分群情況 16

圖 6.1.1　1950～1970 年拿掉出書少於 3 本出版社不同分法分析圖 24
圖 6.1.2　1971～1990 年拿掉出書少於 3 本出版社不同分法分析圖 25
圖 6.1.3　1991～2010 年拿掉出書少於 3 本出版社不同分法分析圖 26
圖 6.1.4　1950～1970 年拿掉出書少於 4 本出版社不同分法分析圖 27
圖 6.1.5　1971～1990 年拿掉出書少於 4 本出版社不同分法分析圖 28
圖 6.1.6　1991～2010 年拿掉出書少於 4 本出版社不同分法分析圖 29
圖 6.1.7　1950～1970 年拿掉出書少於 5 本出版社不同分法分析圖 30
圖 6.1.8　1971～1990 年拿掉出書少於 5 本出版社不同分法分析圖 31
圖 6.1.9　1991～2010 年拿掉出書少於 5 本出版社不同分法分析圖 32
圖 6.1.10　1950～1970 年拿掉出書少於 6 本出版社不同分法分析圖 33
圖 6.1.11　1971～1990 年拿掉出書少於 6 本出版社不同分法分析圖 34
圖 6.1.12　1991～2010 年拿掉出書少於 6 本出版社不同分法分析圖 35
圖 6.2.13　LSI 第一期剃除資料少於三本出版社分析圖 36
圖 6.2.14　LSI 第一期剃除資料少於三本出版社分析圖 36
圖 6.2.15　LSI 第一期剃除資料少於三本出版社分析圖 37
圖 6.2.16　1950 ～ 2010 出版文類數量統計圖 37

DOI:10.6814/NCCU202101092

第一章　緒論

第一節　研究背景

自從光復初期發展之後，此時期的台灣在各方面皆以驚人的速度成長，本土寫作量大增，反共文學，鄉土文學興起，各種社會潮、文化教育書籍等引入，出版業也是蓬勃發展，直到民國 70 年，根據行政院新聞局的統計，全台就有 2123 家出版社，其中還不包括在民國 40 年到 70 年間，如雨後春筍般的出現，又默默凋零的出版社，或者偷偷印行禁書的地方出版社。此時期各大出版社紛紛推出翻譯書、經典古籍、文庫本等讀物，或是重新編印之前自家出版社出過的書籍。

在民國 70 年六月，遠流、爾雅、戶外生活、遠景等 7 位來自各大出版社的發行人、總編齊聚一堂、針對出版社市場的行銷通路做了一次 [知識之道] 的座談。其中便大膽預測 [大型書店將在兩年後出現]、[開一家類似百貨公司一樣大的書店]，並主張出版業應當引進科學化，企業化的管理模式，其產銷流程應分工，專業化，讓出版書專司出版，由書報業、圖書經銷業者負責到各行銷通路鋪貨，而有 [書店] 負責將書籍推銷給社會大眾，如此一來出版品質才能提高。另外，書店應有 [都市化] 和 [社區化] 兩種走向，以滿足不同讀者群的需求。座談中還有提及國內應鼓勵本土寫作，大量翻譯外文書籍充斥圖書市場，使台灣淪為外國文化的殖民地。

以今觀之，當時的出版人還蠻滿洞見。在出版產銷方面，隨著台灣文化產業日趨成熟，出版和圖書行銷通路分工已呈現今主流。出版業仍保留其書店經營者，乃有其歷史因素。在提倡本土創作方面，現今台灣雖有各類文學獎鼓勵新人創作，整體圖書出版和消費環境仍舊不利於發展，翻譯書仍市場主流，也常常市書店內的主打。這樣看來，台灣的出版書似乎還是沒什麼改變。而關於大型書店的崛起、書店社區化多多少少都構成當今書店風貌。

DOI:10.6814/NCCU202101092

隨著台灣經濟起飛，出版業也更加傾向企業化運作自由度和科技含量相輔相成，高度結合，釀就了台灣出版業的空前繁榮 (2010 年達 6000 家出版社)。由此我們知道，出版社依但回歸自由寬鬆的狀態，就會真正成為社會和讀者之間的虹橋。

第二節　研究動機

近幾年大數據時代的崛起，對於資料分析的需求更是提高許多，其中 MCA 更是應用在許多地方進行研究，SPSS 統計軟體甚至有該方法的套件能夠直接應用，社會科學家也會利用該軟體進行研究，本文這次想觀察的是對於經過多重對應分析呈現出來的結果圖，是否存在那些疑慮，所以本文想利用這次的實驗資料觀察 MCA 是否能夠呈現該資料的分布型態，以及對於另外兩種資料分析方法 MDS 與 LSI，觀察這三種方法對於實驗資料的分析結果。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202101092

第二章　文獻回顧

第一節　MCA 多重對應分析

一、 對應分析介紹

1935 年，Hirschfeld 發表了一篇數學論文 [15]，解釋了與列聯表的列和行續在的關係相關的代數表達式。他被認為是對應分析的創始人，但重要的先驅是 Fisher 和 Gattman。Hirschfeld 和 Fisher 的工作側重於二維數據，而 Guttman 的工作側重於多維數據集。Guttman 陸續在 1944 年9、1954 年 [8] 和 1968 年 [10] 發表了該方法的一連串研究成果。此後他的結果得到了重新開發。然而，這只是研究的一部分，因為研究是由許多國家在同一時間進行，例如法國、日本、荷蘭和其他英語國家，但在知識交流上面受到語言不同的阻礙. 這些平行發展催生了各種流派和函數分析，但 Gifti 和 Tenenhaus and Young 等研究人員對該方法進行了不同的修改。整合成兩個一致的目標。

事實上，對應分析在法國已經成為資料分析方法主流，它也是數據分析的主要方法。主要感謝 Jean-Paul Benseci (1973) 和他的同事們的努力。首先，Benzekri 在 1977 年研究了列為行為輔音和元音的中文字母。他發現列和行的組合可以成為一種理解如何使用漢字的語言。從這個意義上說，法語中的“corresponence”是指列和行之間的“關聯性系統 (system of associations)”。英文直譯是“對應”，但實際上本義是轉換，即將復雜無差別的事物轉化為簡單易懂的形式。如今，對應分析在美國社會科學界非常流行。這確實歸功於法國社會學家的努力（Clausen，1998：Greenacre，1984）。

二、 多重對應分析介紹

多重對應分析（MCA）[5] [14] 在二戰前出現在歐洲，但其潛力尚未在社會科學中得到研究。2000 年左右方法流傳美國後，被應用於語言學研究，成為該領域的重要研究方法（Glynn, et al., 2014; Glynn,& Robinson, 2014）。

DOI:10.6814/NCCU202101092

在過去的五年裡，隨著 R 語言和套件的發展，這種方法已經由法國學者通過出版專書和世界各地數據分析師的眾多套件而發展起來。

然而，多重對應分析是通過指定觀察值（個體）和類別值來量化名義（類別）數據，這樣，同一類別的個體在圖形上將彼此靠近，而不同類別的個體在圖形上將被分離。每個對象將盡可能靠近包含該對象的類別的類別點。通過這種方式，將個體按類別劃分為同質性的群組。將同一類別的個體歸為同一子群組的變量則稱同質性變量。

三、 多重對應分析理論

由於 MCA 適用於分類變量如多項選擇題）的統計推斷，因此首先要做的就是將定量數據（如年齡、身高、體重、一天中的時間等）分類。當數據集完全表示為分層變量時，可以創建相應的完整下拉表。我們展示這張表 X。如果有 I 人回答一個問卷 J 個多重選擇題，每題有 4 個答案，X 會有 I 行和 $4J$ 列。理論上的，表 X 是完全藉由 I 個觀察值所選出的 K 分類變量。。假定有 $k-th$ 變量每個變量有 J_k 個級別使得 $J=\sum_{k=1}^{K} J_k$ 而表 X 是一個裡面係數只有 0 跟 1 的矩陣，矩陣大小為 $I \times J$ ，另外這邊定義一個 N 是 X 矩陣裡面所有係數總和並且定義 $Z=\frac{X}{N}$ 在 MCA 中還有兩個特殊向量；第一個是 $r_{I\times 1}$ 其中 r 裡面的元素是矩陣 Z 的每一列的總和，和向量 $c_{J\times 1}$，其中的 c 裡面的元素是矩陣 Z 的每一行的總和。$D_r = diag(r)$ 和 $D_c = diag(c)$，對角矩陣包含 r 和 c 分別作為對角線。根據這些符號，去計算 M 矩陣:

$$M = D_r^{-\frac{1}{2}}(Z - rc^t)D_c^{-\frac{1}{2}} \tag{2.7}$$

再去分解式 (2.6) 的 M 矩陣使得 $M = P\Delta Q^t$，其中 P、Q 都是正交矩陣，Δ 是廣義奇異值對角矩陣。算出 P、Δ、Q^t 這三個矩陣之後就可以開始定位座標，觀測值的座標為 $F = D_r^{-\frac{1}{2}}P\Delta$；變量座標為 $G = D_c^{-\frac{1}{2}}Q\Delta$

第二節　MDS 多維度分析

一、 多維度分析背景介紹

多維度分析 [17] [18] [15] 的出現和發展與其在心理學的各個分支中的應用密不可分。1940 年代是其開始和準備階段，1950 年代是計量多維度分析的發展時期。Torgerson 提出了第一個 J-measure 多維度分析的數學模型，分別於 1952 及 1958 年提出兩篇論文 [17] [18]，為將來的發展奠定了基礎。1960 年代是非計量多維度分析的發展

DOI:10.6814/NCCU202101092

時期。1970 年代後，以上提出的各種方法逐漸成熟並以算法的形式出現，而多維度分析主要是使用距離矩陣執行一些操作並轉換質心以通過 SVD 解析找到我們想要的數據。

MDS 是由 Torgerson(1952) 所正式提出，簡單來說，MDS 是一種 N 個主體根據 P 的準則貧孤 M 個客體的統計方法。假如我們以購買鞋子為例，今天拿到的資料有 N 的消費者購買鞋子，我們想知道購買鞋子的 N 個消費者消費關係，那我們可以分類為來購買的族群年齡 (第一類群)，選擇的品牌 (第二類群)，購買的球鞋總類 (第三類群)，根據這三個類群的準則來給定分數。在這樣的情況下，MDS 就是一個很好的分析方法讓我們去觀察在這幾個類群投影至可是圖的情況下，彼此間的相似及差異性。

MDS 的精隨是利用每筆資料間的相關性質轉換成一個距離矩陣，他可以協助在歐式空間 (Euclidean space) 下觀察出原始資料隱含的訊息；也就是說，多元尺度法是利用主體以圖上的點來代表，進而在圖上觀察主體中每個個體之間的距離關係。

二、 多維度分析理論

MDS 依據拿到的資料型態可以分為兩種情況，第一種是計量多元尺度法 (Metric Multidimen¬sional Scaling) 和非計量多元尺度法 (Non-Metric Multidimen¬sional Scaling)，計量多元尺度法前面有提到是 Torgerson(1952) 所提出，而資料中輸入的值必須是連續性的尺度: 另一方面，非計量多元尺度法是在 1962 年由 Shepard 以及 1964 年 Kruskal 所提出的理論，其中輸入的資料是間斷性尺度。

MDS 研究有幾個步驟：

1. 解決問題-需要比較哪些變量？需要比較多少個變量？研究目的是什麼？

2. 獲取輸入數據-這邊輸入的矩陣需要如距離矩陣的形式，及行向量列向量的跟指標是相同的，舉例如表 2.1.1，這樣就可以利用這個矩陣裡面的值去做後面把它座標化的運算。

3. 進行 MDS 統計程序-許多統計軟件中都提供了用於運行該過程的軟件。只要把資料輸入到這些軟件中就可以直接套用，並且運算出結果。

4. 確定維度數-研究人員需要先決定他們最後成果要呈現的維度數。MDS 解決方案

DOI:10.6814/NCCU202101092

表 2.2.1 城市相對距離表格 (單位: 公里)

	ATL	ORD	DEN	HOU	LAX
ATL	0	587	1212	701	1936
ORD	587	0	920	940	1745
DEN	1212	920	0	879	931
HOU	701	940	879	0	1374
LAX	1936	1745	831	1374	0

的可解釋性通常很重要，而可解釋性通常與維度成正比，維度越高解釋性越高；反之，維度越低解釋性越低，而低維解決方案通常更易於解釋和可視化。但是，維度選擇是在於平衡欠擬合和過度擬合的問題。低維度的解決方案可能會由於遺漏相異性數據的重要維度而不合適。高維解決方案可能會過分適合差異性測量中的雜訊。

5. 映射結果並定義維度-統計程序（或相關模塊）將映射出數據分析出來的結論。地圖將繪製每個個體所代表的點（通常在二維空間中）。產品彼此之間的接近程度表明它們的相似程度，相似性越高在距離上會越近；反之，差異性越高，距離會越遠

計量多元尺度法 (mds for large dataset jengnan tzeng) 這裡我們將詳細解釋步驟 3 的統計方法。假設有 n 個數據類型的主題，每個主題都有 p 個標準，數據類型可以作為矩陣 $X_{p\times n}$，矩陣中的每條街道列就是主題的原則，$x_i = (x_{1i}, x_{2i}, ..., x_{pi})$，它被認為是維度為 p 的歐式空間中的一個點。為了確保 X 在 MDS 之後是唯一的，並且每個主題的相對位置不影響數據的分析，首先將每個數據移動到質心。然後將 D 定義為 X 距離的方陣。$D = [d_{ij}]_{n\times n}$，從距離的方陣可以得到如下關係。

DOI:10.6814/NCCU202101092

$$\frac{1}{n}\sum_{i=1}^{n} d_{ij} = \frac{1}{n}(\sum_{i=1}^{n} x_i^T x_i - x_j^T \sum_{i=1}^{n} x_i - x_j^T \sum_{i=1}^{n} x_i + n x_j^T x_j)$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i + x_j^T x_j \tag{2.1}$$
$$\frac{1}{n}\sum_{j=1}^{n} d_{ij} = \frac{1}{n}(\sum_{j=1}^{n} x_j^T x_j - x_i^T \sum_{j=1}^{n} x_j - x_i^T \sum_{j=1}^{n} x_j + n x_j^T x_j)$$
$$= \frac{1}{n}\sum_{j=1}^{n} x_j^T x_j + x_j^T x_j \tag{2.2}$$
$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij} = \frac{1}{n^2}\sum_{j=1}^{n}(\sum_{i=1}^{n} x_i^T x_i + n x_j^T + n x_j^T x_j)$$
$$= \frac{1}{n}(\sum_{i=1}^{n} x_i^T x_i + \sum_{j=1}^{n} x_i^T x_j) \tag{2.3}$$

要獲得每個 x_i 的相對關係，請將 B 定義為 X 的內積矩陣，$B = X^T X$；在 Young and Householder (2003) 裡面有提到，在歐幾里德空間中定義的實點之間的相互距離相等 $(d_{ij} = d_{ji})$ 之充份並且要求為 $B = X^T X$ 為半正定矩陣 (semi-positive definition Mat rix)；這個集合中的每個點都在歐幾里德變換 (Euclidean transformation) 後唯一。等式（2.1）和（2.3）的組合如下:

$$d_{ij} = x_i^T x_j = -\frac{1}{2}(d_{ij} - \frac{1}{n}\sum_{i=1}^{n} d_{ij} - \frac{1}{n}\sum_{j=1}^{n} d_{ij} + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij})$$
$$= -\frac{1}{2}(d_{ij} - d_{i+} - d_{+j} - d_{++}) \tag{2.4}$$

由 2.4，矩陣 B 可以轉換成以下形式:

$$B = X^T X = -\frac{1}{2}(D - \bar{D}_r - \bar{D}_c + \bar{D}_g) \tag{2.5}$$

式 2.5 當中:

$\bar{D}_r = diag[d_{i+}/n]$, d_{i+} 為第 i 列的數值總和，而 $\bar{D}_r$ 除了對角線有 d_{i+}/n 這些值以外，其他位置為 0 ; $\bar{D}_c = diag[d_{+j}/n]$, d_{+j} 為第 j 行的數值總和，而 $\bar{D}_c$ 除了對角線有 d_{+j}/n 這

DOI:10.6814/NCCU202101092

些值以外，其他位置為 0;$\bar{D}_g = [d_{++}/n]$，d_{++} 為矩陣 D 的所有的值總和。式 (2.5) 又稱為 double centering。

找到內積矩陣 B 之後就可以對其做 SVD 分解，因為 B 矩陣具備對稱矩陣的性質所以可以將其 SVD 分解表示為:

$$B = U\Sigma U^T \tag{2.6}$$

根據最後得到的 B 的矩陣，去做 SVD 分解之後，去觀察 Σ 裡面特徵值 (eigenvalue) 的比重關係，因為 Σ 是對角矩陣且裡面的值 $\sigma_{11} > \sigma_{22} > ... > \sigma_{nn}$，藉由式 (2.5) 以及式 (2.6) 可得 $X = \Sigma^{\frac{1}{2}} Z^T$。因為在降低維度的同時不能失去原始資料的相關性，在使用 MDS 方法去分析的過程過，會採用與主成份分析 (PCA) 類似手法去捨棄較小的特徵值，也就是說只保留較大的特徵值，假如今天要把資料投影到二維圖上那座標點的表示法就是 $(z_1\sigma_{11}^{\frac{1}{2}}, z_2\sigma_{22}^{\frac{1}{2}})$，$z_1$、$z_2$ 分別為 Z 矩陣第一個行向量跟第二個行向量。。

然而 MDS 這個方法的使用情況還是會出現侷限性，而問題出現在我們利用 MDS 這個方法通常是要投影至二維空間或三維空間，然而如果是投影到這兩個空間的話，所選擇的特徵值就會是前兩個或前三個，但不一定每筆資料的前三個的特徵值比重能夠足夠表達出這筆資料的特性，因此，在使用 MDS 方法時需要確認一下過程中的特徵值的情況，以免保出來的結果解釋性過低。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

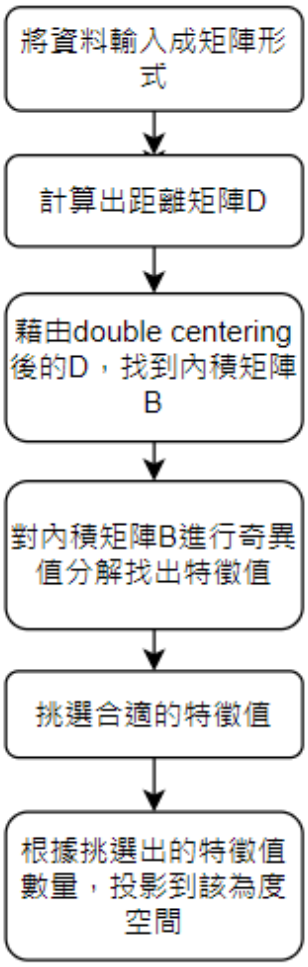


圖 2.2.1 多維度分析流程圖

非計量多元尺度法

上面介紹的多元尺度測量方法可以對連續尺度的數據進行分析，但是我們生活中得到的數據往往是不是連續性尺度，無法直接從這些數據中計算出距離矩陣。Shepard（1962）和 Kruskal（1964）) [13] 提出了一種解決不連續尺度的方法，稱為非計量 MDS。

非計量 MDS 與計量 MDS 差異其實不大，非計量 MDS 只是比計量 MDS 多了一個前置作業，裡面運算的精隨還是一樣的，差別只在於非計量因為是得到間斷尺度資料，所以只要把間斷尺度資料轉換為距離關係，而這個距離關係必須具有單調關係；也就是說，假如今天三個主體 i、ii、iii 為三個不同的外送業者，而有份針對不同外送業者的滿意度調查如表 (2.1.1)

此外，滿足主體 i、ii 之間且 ii、iii 之間的距離最遠的條件可能有很多方式可以表達；也就是說，有 N 個主體的話，就有 $C_2^N = \frac{N(N-1)}{2}$ 個次序關係。假設今天希望把這

DOI:10.6814/NCCU202101092

表 2.2.2 原始出版社資料數量

主體	服務態度	距離關係
i,ii	優良、普通	0.7
ii,iii	優良、待加強	1
i,iii	普通、待加強	0.7

N 個主體投影至 r 維空間上面，則需要 $N \times r$ 個座標點才能表達，但是，如果次序的關係相對大於坐標值的個數，則序數關係限制了每個坐標的相對位置，因此可以找到唯一的解決方案。MDS 的計算過程如下示。

1. 計算每個主題對之間的相似度順序。
2. 求出每個點的歐式空間距離。

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

3. 確定一個 r 維度的投影空間 $\hat{d}_{ij}$ 表示為 i,j 兩點在 r 維空間上的單調函數。如果有兩組或更多組相對距離匹配序數關係，則相對距離必須匹配原始數據的序數關係。
4. $\hat{d}_{ij} = d_{ij}$ 一旦有不一致的情形出現，則 $\hat{d}_{ij}$ 為其平均值。

第三節　潛在語義索引

一、 潛在語義索引介紹

使用潛在語義結構的信息檢索技術於 1988 年由 Scott Deerwester、Susan Dumais、George Furnas、Richard Harshman、Thomas Landauer、Karen Lochbaum 和 Lynn Streeter 獲得專利 [7]。將這項技術應用於信息查詢當中，也可稱之為潛在語義索引 (LSI)。

潛在索引既是一種搜索方法，也是一種索引。奇異值分解用於識別具有連接的非結構化文本集合中的模式。在相同的上下文中使用的單詞通常被認為具有相同的含義，而 LSI 是基於此規則的搜索方法。LSI 的一個重要特性是它可以通過在出現在同一上下文中的單詞之間創建鏈接來從文本中提取特定內容，這與僅檢索特定關鍵字的傳統搜

DOI:10.6814/NCCU202101092

索方法不同。它被稱為潛在語義索引的原因是您可以連接一組文檔中具有隱式語義連接的單詞。

二、 潛在語義索引方法

進行 LSI 需要找出資料矩陣，矩陣列向量指標為第一類變量，行向量指標為第二類變量，LSI 在於利用 SVD 分解資料矩陣去定位兩類變量在可視圖上關係，簡單介紹 LSI 在本篇文章中使用方法為先建立如下表 2.2.3

表 2.3.3 原始出版社資料數量

	小說	散文	詩	論述	傳記
一葦	3	5	5	71	6
三久	5	30	9	94	45
三友圖書	12	20	0	8	31
三民書局	7	9	79	0	13
上海人民	36	15	31	17	40

根據裡面的值可以定義矩陣 L

$$L = \begin{pmatrix} 3 & 5 & 5 & 71 & 6 \\ 5 & 30 & 9 & 94 & 45 \\ 12 & 20 & 0 & 8 & 31 \\ 7 & 9 & 79 & 0 & 13 \\ 36 & 15 & 31 & 17 & 40 \end{pmatrix}$$

接著對 L 矩陣進行奇異值分解 $L = U\Sigma V^T$ 進行 LSI 需要找出資料矩陣，矩陣列向量指標為第一類變量，行向量指標為第二類變量，LSI 在於利用 SVD 分解資料矩陣去定位兩類變量在可視圖上關係，假如今天資料矩陣為 L，$L = U\Sigma V^T$，則列向量變量座標: $U\sqrt{\Sigma_k}$，行向量變量: $\sqrt{\Sigma_k}V^T$ 其中 k 代表維度。

DOI:10.6814/NCCU202101092

第三章　資料描述

第一節　資料來源

本研究中所適用到的資料來源為，”2007 台灣作家作品目錄系統” 該系統一共收錄了約兩千六百多位作家小傳及作家的十萬多部作品，資料收錄至 2007 年 6 月，台灣也有相關研究是利用該筆資料 [3] [4]，這筆資料不僅僅具有統整文學史料的意義與價值，更是具體的收錄一個世紀以來，台灣的作家創過過程與出版之成果。

”2007 台灣作家作品目錄系統” 該目錄旨在蒐集臺灣現代文學作家及文學作品的資料，並藉以反映臺灣現代文學的整體成就。所收作家凡 2600 餘位作家，計十萬種作品。收錄標準為：

- 以長期生活在臺灣且從事現代文學創作之作家為主，包括在臺灣出生、成長。
- 曾在此受教育工作與生活，有作品曾在臺灣出版的現居國外之作家。
- 其入選基本條件是需有正式出版一本以上的個人作品集，且其作品已獲文學社群之肯定者。
- 文學社群肯定之條件:

    1. 獲校園以外的文學獎。
    2. 年度小說、散文及詩被選入者。
    3. 九歌文學大系、聯副三十年大系等大系被選入者。
    4. 文學評論部分要有一定學術地位及研究成果始能認定。

書本一直是文化傳承的重要媒介，而出版社在是書本輸出的重要媒介，在此研究之前也有相關研究在討論 [1] [2]，本次研究將這次拿到的資料分為三個區間分別為 1950～1970、1971～1990、1991～2010，這三個時間區間去做比較，去分析 3 個年份區

DOI:10.6814/NCCU202101092

間台灣出版社出書的差異性，藉此分析出台灣不同年份時期的閱讀習慣差異。

這次的資料分析，主要在瞭解台灣文學生產中，作家與出版社之間的關係。關於這主題之前文學研究界有許多印象，如文學出版社、或作家出版社的說法，或是來自於其他國家所的文學社會學理論的說法，如文學性出版與商業性出版社形成兩極光譜的理論。但是，關於這次研究問題是以資料來分析，看看台灣作家與出版社之間關係是不是可以找到其分群的模式，並加以解釋。

此次研究拿到的資料量為如表 3.1.1，其中為了避免分析過程中出現太多不必要偏差有事先整理剔除出版社出版書數量少於三本、四本、五本、六本的情況，因為有些出版社出版數量太少可能是因為某些特殊原因，而我們是希望分析普遍出版社的分群情況，所以先把資料量較少的出版社拿掉，再去分析出版社分群情況。

表 3.1.1 原始出版社資料數量

	1950 ～ 1970	1970 ～ 1990	1990 ～ 2010	total
原始資料量	3668	10719	15255	29662
拿掉小於 3 本出版社	3288	10154	14388	27830
拿掉小於 4 本出版社	3196	9934	14112	27242
拿掉小於 5 本出版社	3092	9777	13865	26734
拿掉小於 6 本出版社	2996	9622	13616	26274

DOI:10.6814/NCCU202101092

圖 3.1.1 為出版社出書一本到六本的出版社數量，可以明顯看到出書數量 1 ～ 3 本的出版社數量明顯突出，為了避免研究過程被出書數量過低的出版社影響，所以研究資料從出書數量少於 3 本開始剔除。

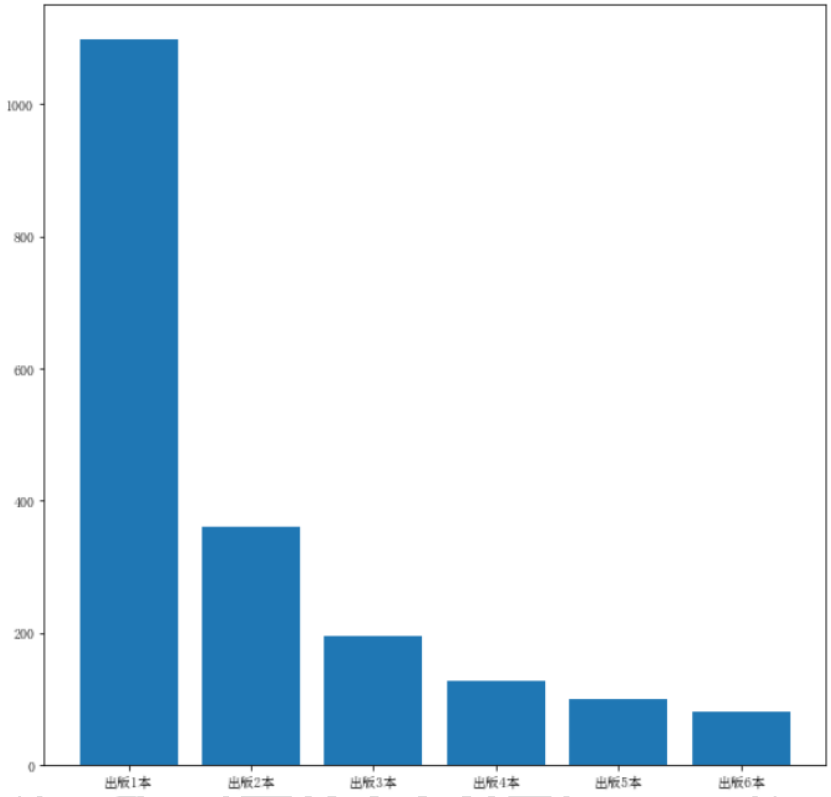


圖 3.1.1　出版數量 1 ～ 6 本出版社分配圖

National Chengchi University

DOI:10.6814/NCCU202101092

第四章　問題描述

第一節　MCA 的侷限性

以下圖表為 MCA 針對三種時期進行的分析結果。

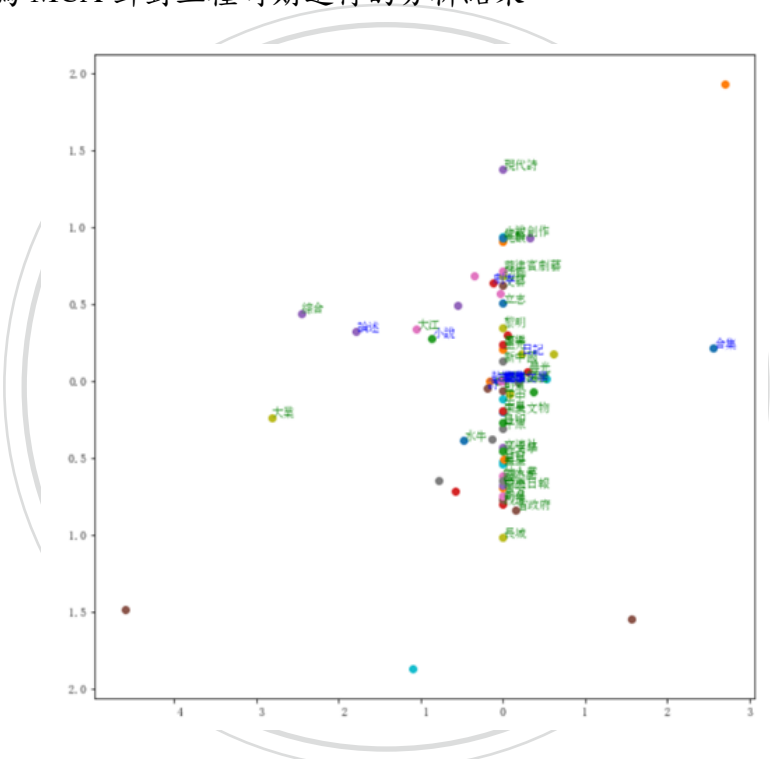


圖 4.1.1　利用 Python 分析 1950 ～ 1970 出版社分群情況

DOI:10.6814/NCCU202101092

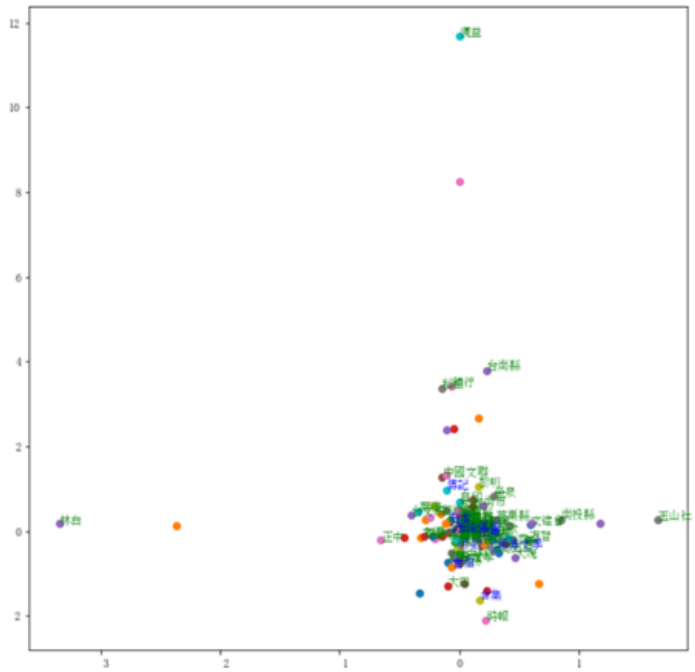


圖 4.1.2　利用 Python 分析 1971 ～ 1990 出版社分群情況

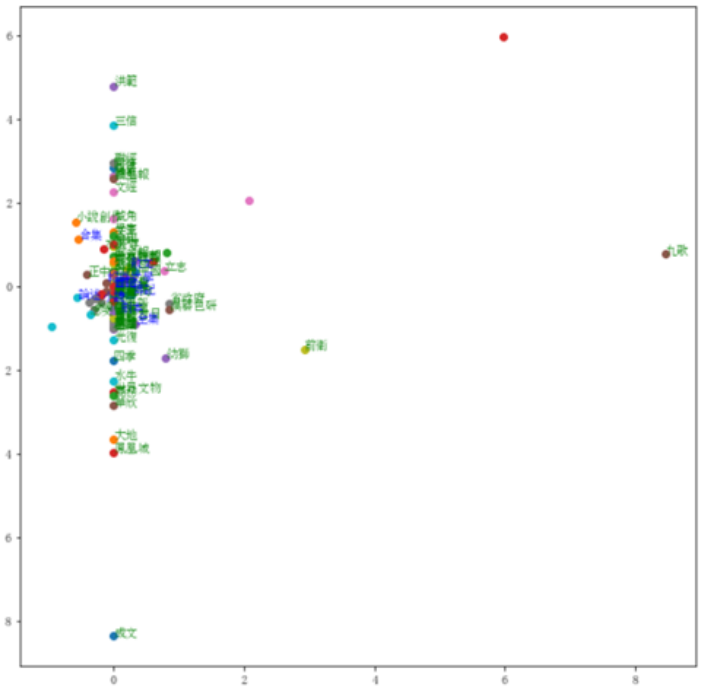


圖 4.1.3　利用 Python 分析 1991 ～ 20100 出版社分群情況

從一開始著手這個問題的時候，先是想到了利用多重對應分析 (MCA) 去分析資料，而這次的實驗使用的程式語言是 Pyhton 於是利用 Python 開始打 MCA 的演算法，然而在編碼的過程中發現資料的解釋量不夠，甚至可以說是相當的低，於是利用 SPSS

DOI:10.6814/NCCU202101092

軟體去做驗證，SPSS 軟體也是目前在統計學上經常會使用到的軟體，裡面很多套件都是事先安裝好的，裡面也有 MCA 的套裝軟件，之後發現 SPSS 的分析結果跟 python 的結果大同小異，但是這邊發現了一個問題，雖然 SPSS 套裝軟體可以得到想要的結果，但就結果的解釋量來說是不夠的。

為了可以更直觀的去觀察數據的解釋量問題可以根據解釋量的定義 $e = \frac{\sum_{i=1}^{r} \sigma_{ii}}{\sum_{i=1}^{n} \sigma_{ii}}$，去觀測資料投影至 k 維可視圖的解釋量，解釋量的比對圖如表 5.1.1，在圖表 5.1.1 中可以發現算出來的解釋量基本落在 0% ～ 1% 之間，所以 MCA 對於這組資料的實驗結果可以說是相當不理想的。在圖表 5.1.1 之後也有放上 Python 跟 SPSS 比對圖，從圖上去分析的分群效果其實也不明顯。

表 4.1.1 MCA 解釋率一覽表

	拿掉少於 3 本	拿掉少於 4 本	拿掉少於 5 本	拿掉少於 6 本
1950 ～ 1970 年	0.8%	0.9%	1.1%	1.2%
1971 ～ 1990 年	0.4%	0.5%	0.5%	0.6%
1991 ～ 2010 年	0.3%	0.4%	0.5%	0.5%

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

第五章　研究方法

第一節　MDS 距離定義過程

從最一開始的利用 MCA 方法去做資料分析，可能因為資料結構的關係，造成解釋量的問題，所以接著我們去嘗試利用 MDS 的分析方法去做下一步分析。

在本文中第二章的部分有提到 MDS 的理論分析，其中現在 MDS 主要分為兩種情況，第一種是計量多維度分析，第二種是非計量多維度分析，根據這次拿到的資料為名義變量資料，所以我們是要利用非計量多維度分析的方法去處理這次拿到的出版社資料。

根據資料內容有出版書的書名、出版文類、出版年、作者，這幾個重要欄位，與我們的問題出版社分群情況演變，相關的欄位主要是出版年份、文類和作者，與出版年份相關是因為我們要去探討不同時期的出版社分群情況，選擇作者跟出版文類是希望可以透過這兩個欄位去分出我們希望的群聚效果。

首先因為拿到的資料裡的欄位都是名義變項，所以需要先將這些欄位進行數值化去造出一個 P 矩陣，去做之後的計量 MDS 分析，那根據這次需要分群的情況，分別把出版社名稱跟出版文類當作行跟列的指標。簡單舉例先把資料整理成像是下方表格(5.1.1):

DOI:10.6814/NCCU202101092

表 5.1.1 出版社出版文類統計表

	小說	散文	詩	論述	傳記
一葦	3	5	5	71	6
三久	5	30	9	94	45
三友圖書	12	20	0	8	31
三民書局	7	9	79	0	13
上海人民	36	15	31	17	40

根據表 5.1.1 我們可以提取裡面的數值，定義一個 Q 矩陣使得 Q 可以表示成:

$$Q = \begin{pmatrix} 3 & 5 & 5 & 71 & 6 \\ 5 & 30 & 9 & 94 & 45 \\ 12 & 20 & 0 & 8 & 31 \\ 7 & 9 & 79 & 0 & 13 \\ 36 & 15 & 31 & 17 & 40 \end{pmatrix}$$

所以 Q 裡面的每一行代表一家出版社的出書數量，而第一步距離定義方式是計算每行之間各個元素的差值的差值，例如定義符號 d_{ij} 為第 i 間出版社以及跟第 j 間出版社的距離，再定義 MDS 距離的過程中通常會利用到兩種算法:

$$d_{ij} = \|d_i^T - d_j^T\| \tag{5.1}$$

$$d_{ij} = 1 - < \frac{d_i^T}{\|d_i^T\|}, \frac{d_j^T}{\|d_j^T\|} > \tag{5.2}$$

其中，d_i 為矩陣 D 第 i 行的行向量，代表第 i 間出版社的各種文類出書數量;d_j 為矩陣 D 第 j 行的行向量，代表第 j 間出版社的各種文類出書數量，式 (5.1) 第一種計算 MDS 距離矩陣的方法，其算法是直接算出兩個向量的距離長度，當第 i 間出版社與第 j 間出版社之間出書數量差距越多，則 d_{ij} 會越大; 反之，當第 i 間出版社與第 j 間出版社

DOI:10.6814/NCCU202101092

之間出書數量差距越小，則 d_{ij} 會越小。

式 (5.2)，利用到計算相似度的概念，相似度是不相似度的對立面。在實際應用中，可以獲得樣本之間的相似度。通常內積可以被視為計算相似度的一種方法，如果兩個向量都是單位向量的话，那內積之後得到的結果是在 $(0,1)$ 之內的，此時，為了得到不相似度，直接將相似度減 1 即可求得，即可將不相似度視為一種距離關係。

以下表 5.1.2 為利用式 (5.1) 去計算出來的距離矩陣

表 5.1.2 出版社相對距離表格

	一葦	三九	三友圖書	三民書局	上海人民
一葦	0	62.2	69.2	103.3	77.3
三久	62.2	0	88.4	123.3	76.2
三友圖書	69.2	88.4	0	82.3	31
三民書局	103.3	123.3	82.3	0	63.9
上海人民	77.3	76.2	31	63.9	0

根據上方定義出來的表 5.1.2，定義出距離矩陣 D:

$$D = \begin{pmatrix} 0 & 62.2 & 69.2 & 103.3 & 77.3 \\ 62.2 & 0 & 88.4 & 123.3 & 76.2 \\ 69.2 & 88.4 & 0 & 82.3 & 31 \\ 103.3 & 123.3 & 82.3 & 0 & 63.9 \\ 77.3 & 76.3 & 31 & 63.9 & 0 \end{pmatrix}$$

依據算出來的 D 矩陣還可以進行微調，定義兩個權重參數:

w_1: 假如同一作者在第 i 家跟第 j 家出版社都出版過書。

w_2: 假如同一作者在第 i 家跟第 j 家出版社都出版過書，且出書的文類相同。

根據 w_1、w_2 後微調的矩陣定義為 $\widehat{D}$，$\widehat{D}$ 裡面的元素定義為 $\widehat{d_{ij}}$，且 $\widehat{d_{ij}} = w_1^s w_2^t d_{ij}$，s 代表有在不同兩家出版書籍且書籍文類不同的作者數量，t 代表有在不同兩家出版書籍且

DOI:10.6814/NCCU202101092

書籍文類相同的作者數量。根據調整過的距離矩陣定義一個解釋率符號 e 代表根據選擇的 k 維度所乘呈現的資料解釋量算法為 $e = \frac{\sum_{i=1}^{k}\sigma_{ii}}{\sum_{i=1}^{k}\sigma_{ii}}$，然而根據實驗結果 w_1、w_2 因為定義方式是拉近相異出版社間的距離權重，所以在矩陣上經過這兩個參數微調後，矩陣的相對距離會拉近，造成解釋率下降。

在實驗過程中有針對式 (5.1) 這與式 (5.2) 兩種算法的解釋率去做比較，比較結果顯示出式 (5.1) 解釋率比式 (5.2) 的解釋率高出一些，以下圖表有做簡單的比較，圖表內的值以 (a, b) 形式表示，其中 a 表示為三維空間解釋率，b 表示為二維空間解釋率

表 5.1.3 距離算法式 (5.1)，$w_1 = 1$、$w_2 = 1$

	$1950 \sim 1970$	$1971 \sim 1990$	$1991 \sim 2010$
拿掉少於 3 本出版	(43.4%,38.4%)	(41.9%,37.1%)	(44.8%,38.5%)
拿掉少於 4 本出版	(43.6%,38.5%)	(42.5%,37.5%)	(41.5%,39.2%)
拿掉少於 5 本出版	(44.0%,38.7%)	(42.8%,37.8%)	(46.5%,40.0%)
拿掉少於 6 本出版	(44.3%,39.0%)	(43.3%,38.2%)	(47.4%,40.9%)

表 5.1.4 距離算法式 (5.2)，$w_1 = 1$、$w_2 = 1$

	$1950 \sim 1970$	$1971 \sim 1990$	$1991 \sim 2010$
拿掉少於 3 本出版	(38.4%,29.8%)	(31.9%,23.8%)	(28.4%,21.6%)
拿掉少於 4 本出版	(37.5%,29.0%)	(31.1%,22.9%)	(27.0%,20.5%)
拿掉少於 5 本出版	(37.9%,29.6%)	(28.9%,21.4%)	(26.1%,19.5%)
拿掉少於 6 本出版	(37.6%,29.2%)	(29.4%,21.6%)	(25.1%,18.6%)

DOI:10.6814/NCCU202101092

表 5.1.5 距離算法式 (5.1)，$w_1 = 0.9$、$w_2 = 0.7$

	1950 ∼ 1970	1971 ∼ 1990	1991 ∼ 2010
拿掉少於 3 本出版	(29.6%,26.1%)	(27.9%,23.9%)	(24.8%,21.0%)
拿掉少於 4 本出版	(30.3%,26.7%)	(28.0%,24.0%)	(25.1%,21.3%)
拿掉少於 5 本出版	(30.9%,27.2%)	(28.2%,24.2%)	(25.4%,21.5%)
拿掉少於 6 本出版	(31.7%,28.0%)	(28.5%,24.5%)	(25.9%,21.9%)

國立政治大學 National Chengchi University

表 5.1.6 距離算法式 (5.2)，$w_1 = 0.9$、$w_2 = 0.7$

	1950 ∼ 1970	1971 ∼ 1990	1991 ∼ 2010
拿掉少於 3 本出版	(37.6%,29.2%)	(31.2%,23.3%)	(27.6%,21.1%)
拿掉少於 4 本出版	(36.8%,28.4%)	(30.5%,22.5%)	26.3%,20.0%)
拿掉少於 5 本出版	(37.3%,29.1%)	(28.4%,21.0%)	(25.5%,19.1%)
拿掉少於 6 本出版	(37.0%,28.7%)	(28.9%,21.2%)	(24.6%,18.2%)

DOI:10.6814/NCCU202101092

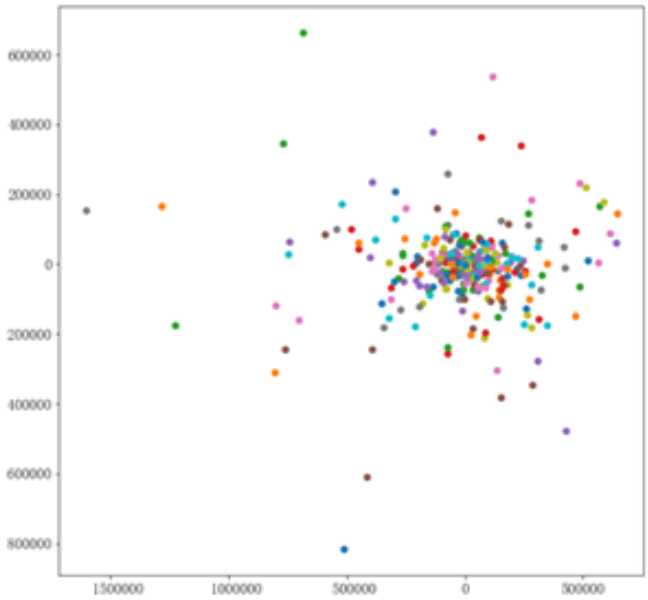
第六章　實驗結果

第一節　MDS 資料分析圖

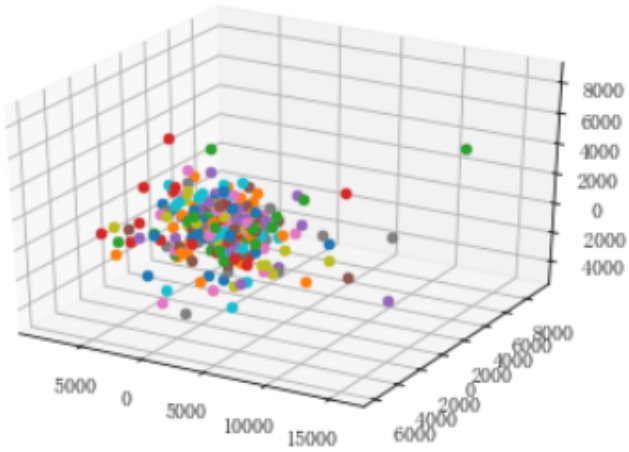
經過一開始的 MCA 方法因為解釋量過低所以改成 MDS 方法去分析之後，解釋量有明顯提升許多大約提升至 40% ～ 50%，所以利用剔除少於 3-6 本不等的出版社資料去分析資料結構，但就資料分析出來的圖形及分佈的趨勢來看差異性不大。其中第一期資料圖形因為數量最少故較稀疏，圖形看起來比較呈水平分佈。MDS 的第一期四個圖都有相同的分佈，MDS 的少於 6 本的特別明顯。第二期圖形高度一致，都很集中在中間地帶，剔除少於 5-6 本就比較清楚可見，可是還是非常集中，漸往外圍均勻分佈。整個看起來像一個球。這作家與出版社之間大量集中在短暫和淺層的關係，即 10 本以內的特別多，可能 10-30 本之間的也很平均。不管如何，這個時期作家與出版社健全發展，大家都有機會，但整體看來比較沒有特色。第三期就圖形來看沿著水平走趨但上下兩邊也都有很平均的分佈。尤其拿掉五、六本後較明顯。不過整體而言還是呈現球形樣態，只不過比較往水平集中的球體而已。

下方圖表 6.1.1～6.1.12 為 MCA 與 MDS2 為投影與 3 為投影的比較圖，比較的資料依序為拿掉少於三本出版資料的三個時期，依序到拿掉少於六本出版資料的出版社的三個時期，根據第五章對於 MDS 解釋量最高的參數選擇，選擇的是 $w_1 = 0.9, w_2 = 0.8$ 這一組參數，可以發現在解釋量最高的 MDS 可視圖下出來的結果都是呈現發散情況。

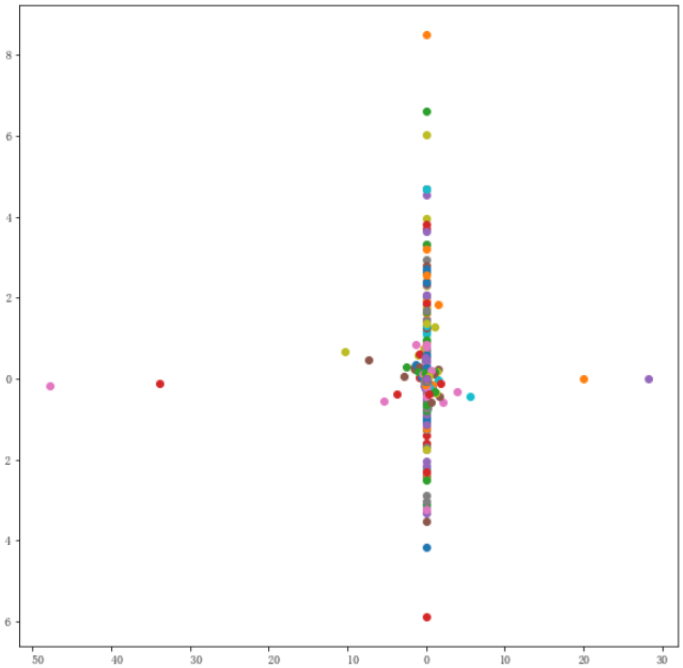
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 34\%$



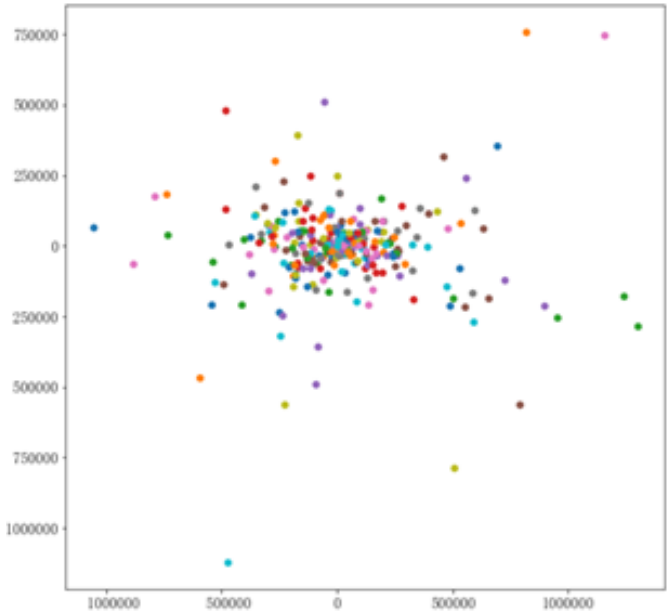
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 40.4\%$



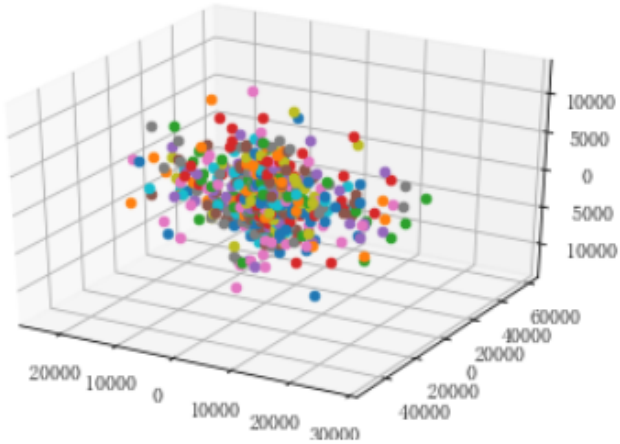
MCA2 維可視圖,$e = 0.8\%$

圖 6.1.1 1950~1970 年拿掉出書少於 3 本出版社不同分法分析圖

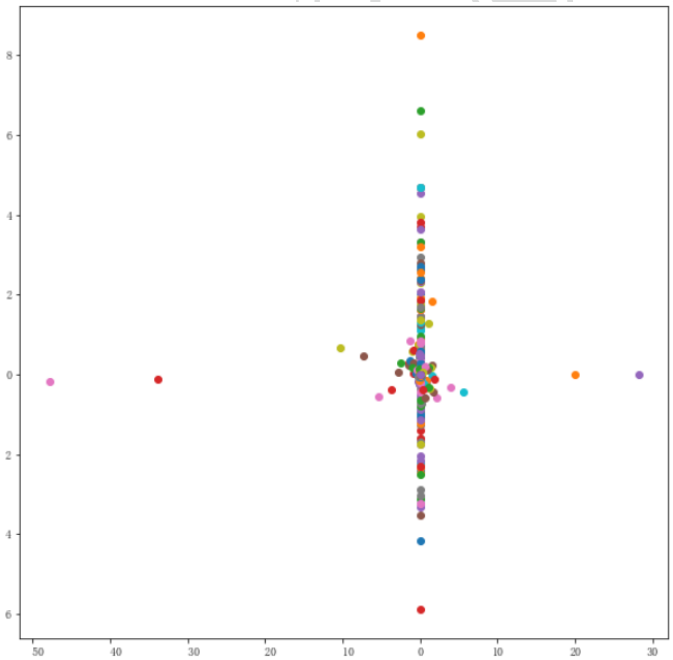
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 40.3\%$



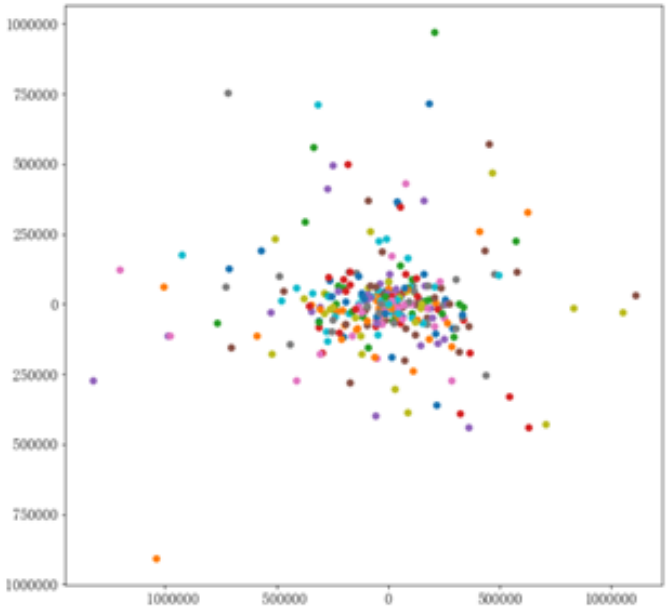
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 47.2\%$



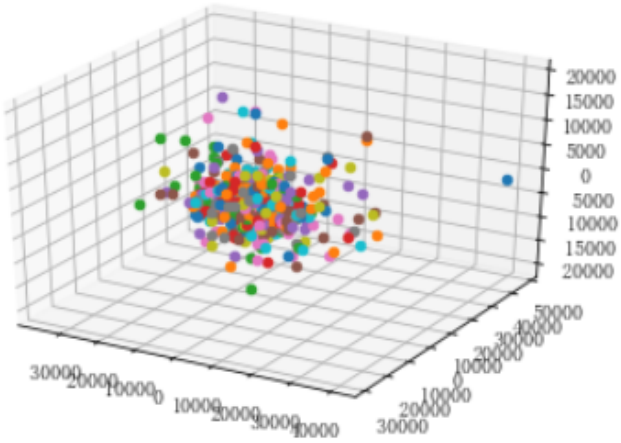
MCA2 維可視圖,$e = 0.4\%$

圖 6.1.2 1971～1990 年拿掉出書少於 3 本出版社不同分法分析圖

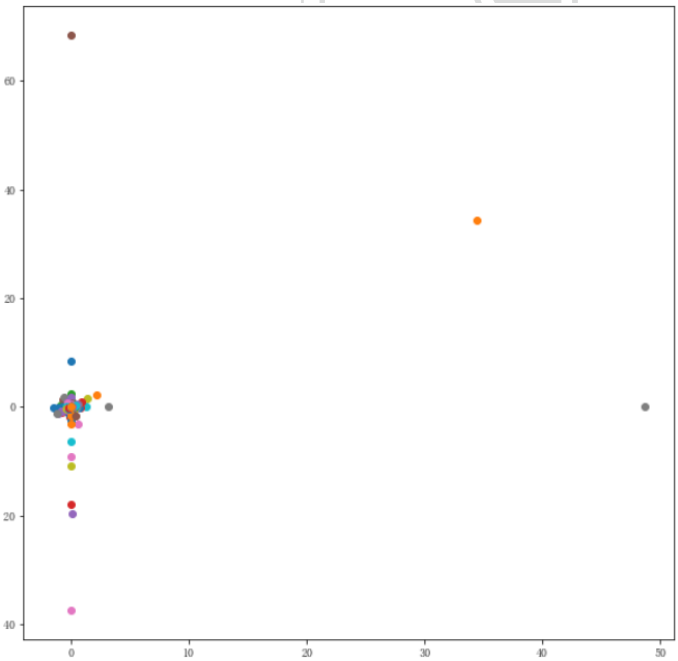
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 34\%$



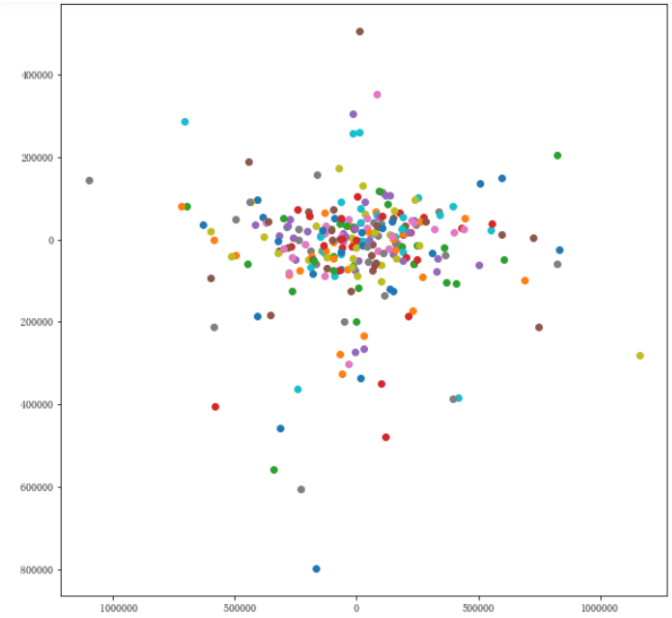
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 40.9\%$



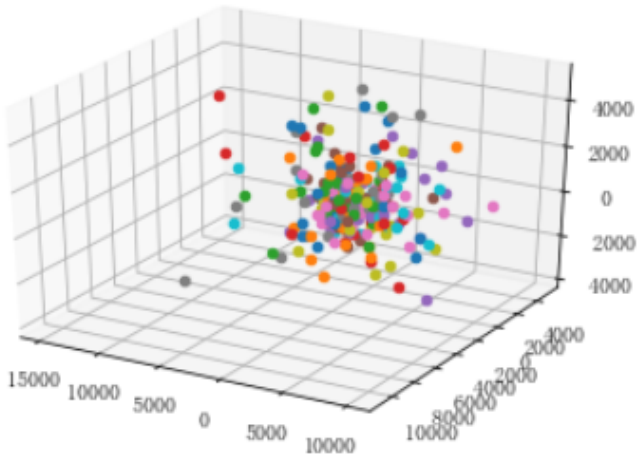
MCA2 維可視圖,$e = 0.3\%$

圖 6.1.3 1991~2010 年拿掉出書少於 3 本出版社不同分法分析圖

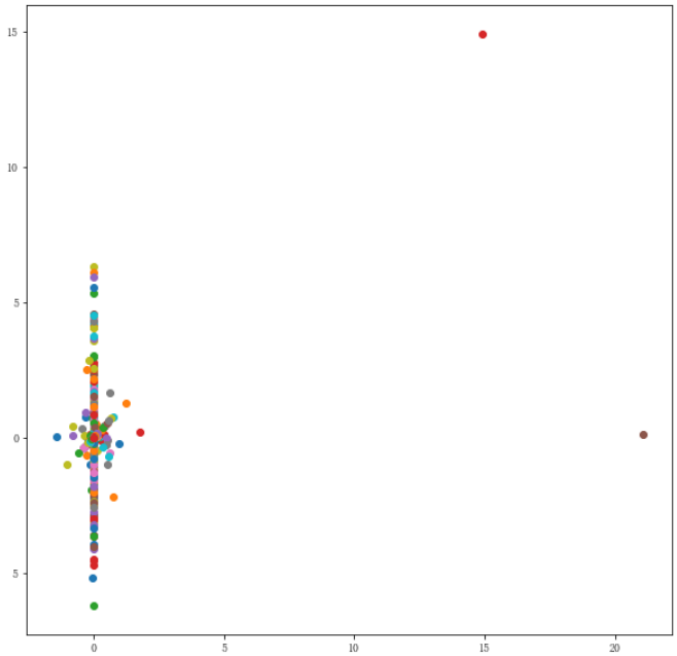
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 36.9\%$



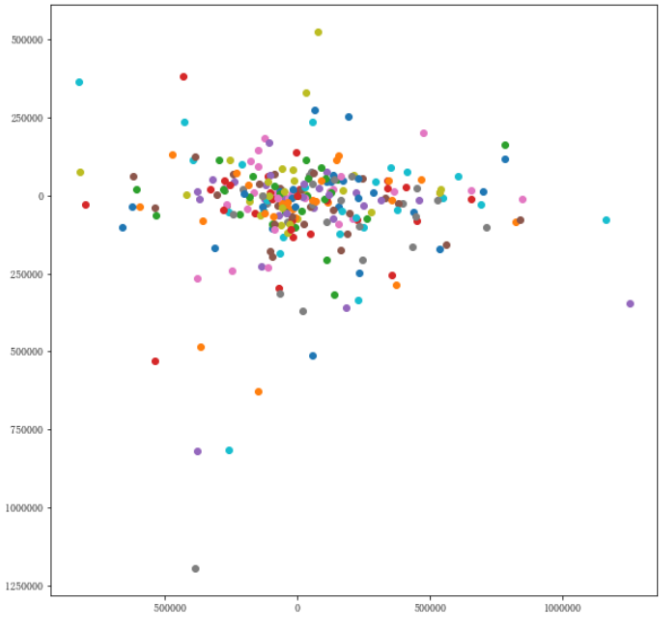
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 43.3\%$



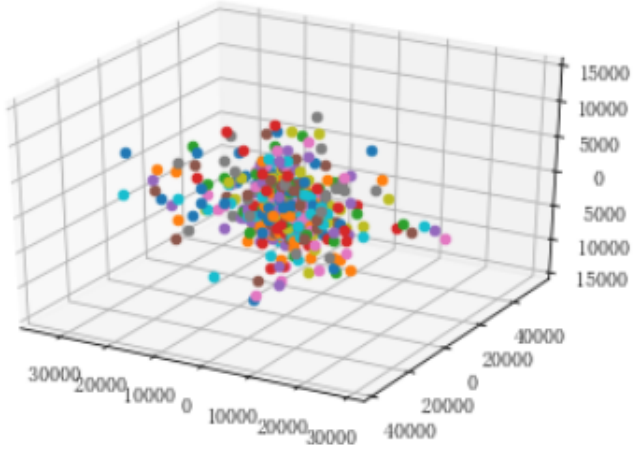
MCA2 維可視圖,$e = 0.9\%$

圖 6.1.4 1950~1970 年拿掉出書少於 4 本出版社不同分法分析圖

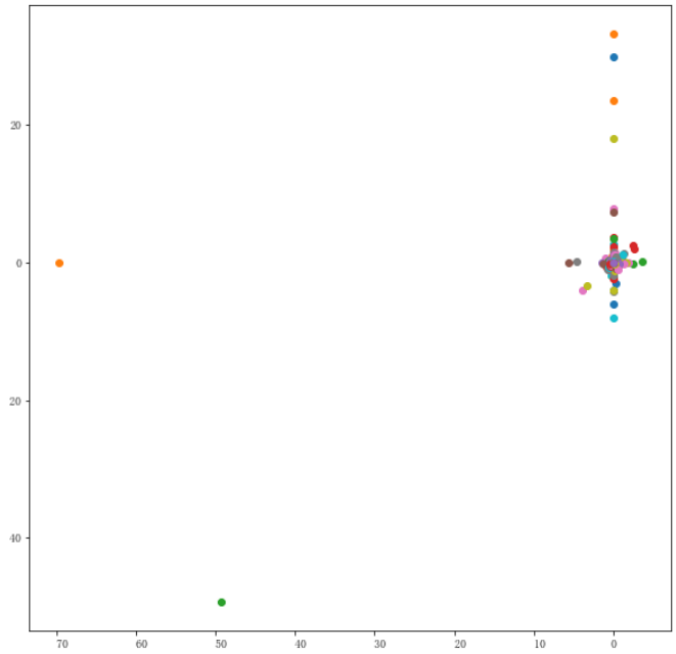
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 42\%$



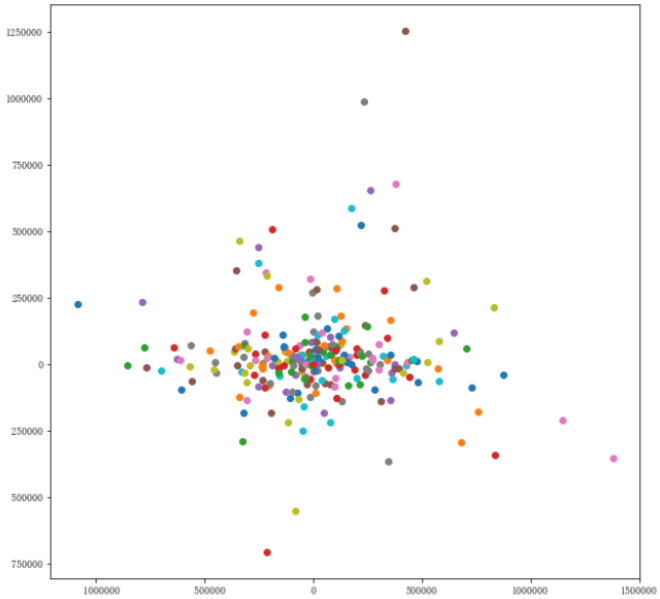
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 49\%$



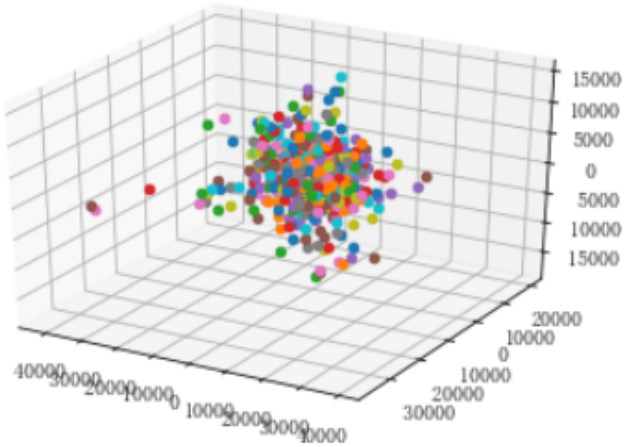
MCA2 維可視圖,$e = 0.5\%$

圖 6.1.5 1971～1990 年拿掉出書少於 4 本出版社不同分法分析圖

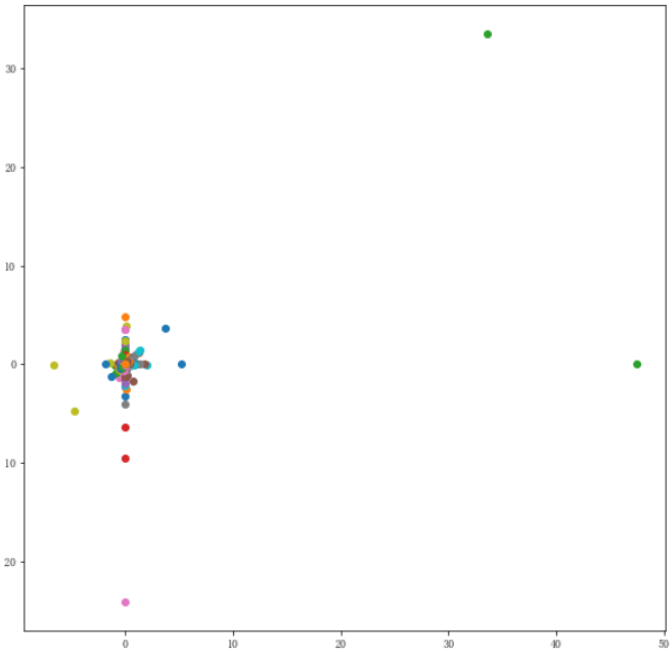
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 36.6\%$



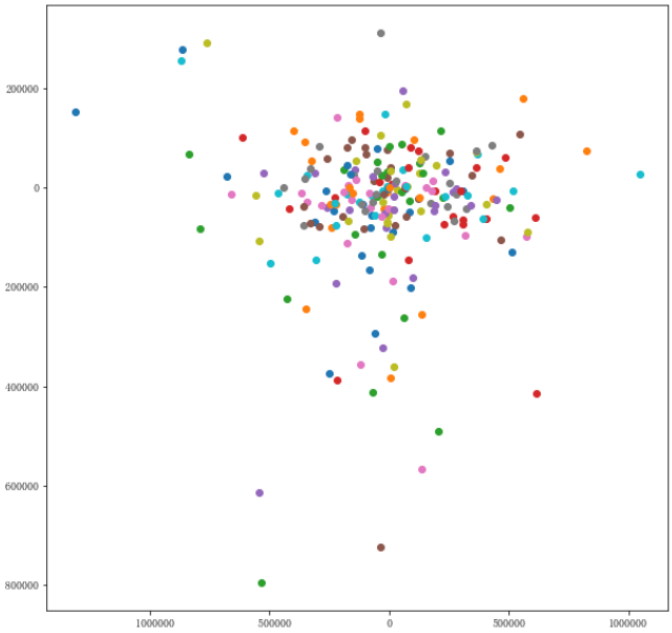
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 43.4\%$



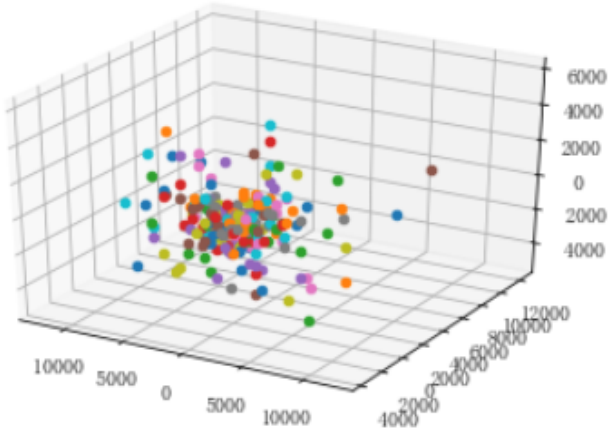
MCA2 維可視圖,$e = 0.4\%$

圖 6.1.6 1991~2010 年拿掉出書少於 4 本出版社不同分法分析圖

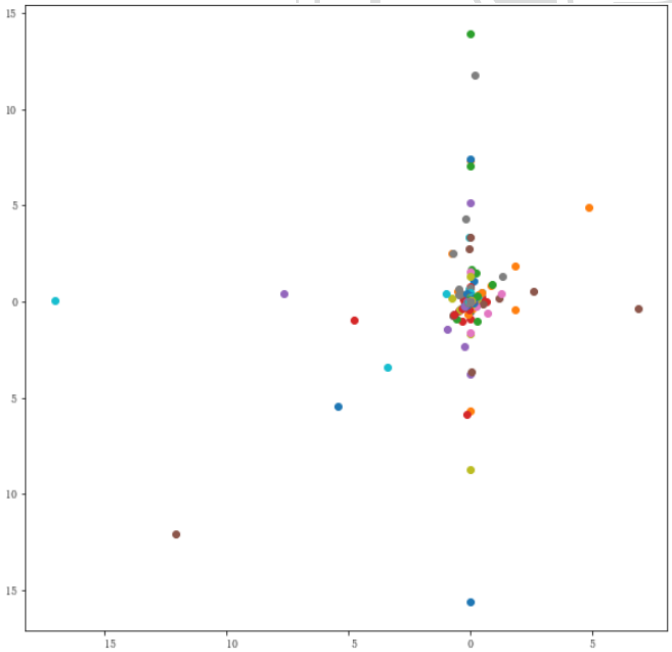
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 37.2\%$



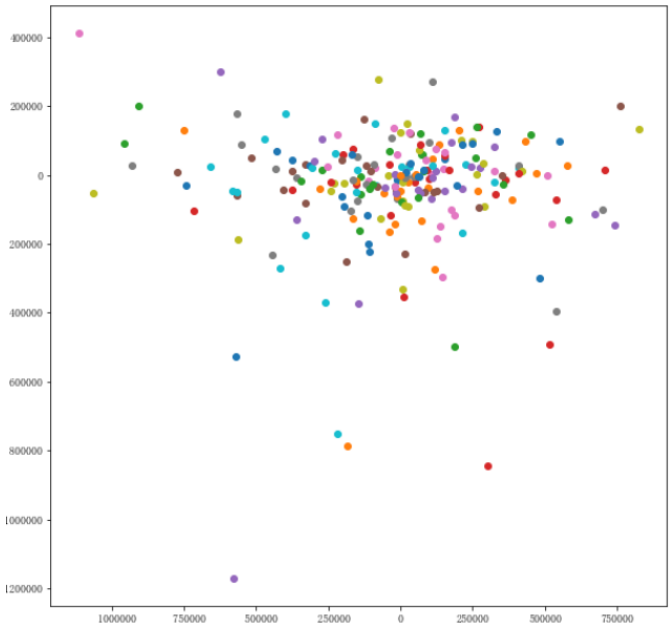
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 43.7\%$



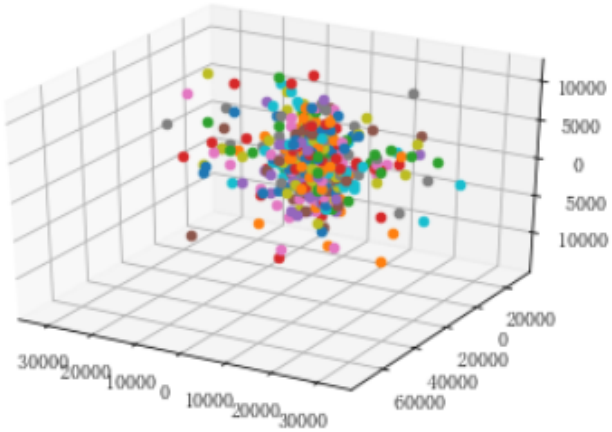
MCA2 維可視圖,$e = 1.1\%$

圖 6.1.7 1950~1970 年拿掉出書少於 5 本出版社不同分法分析圖

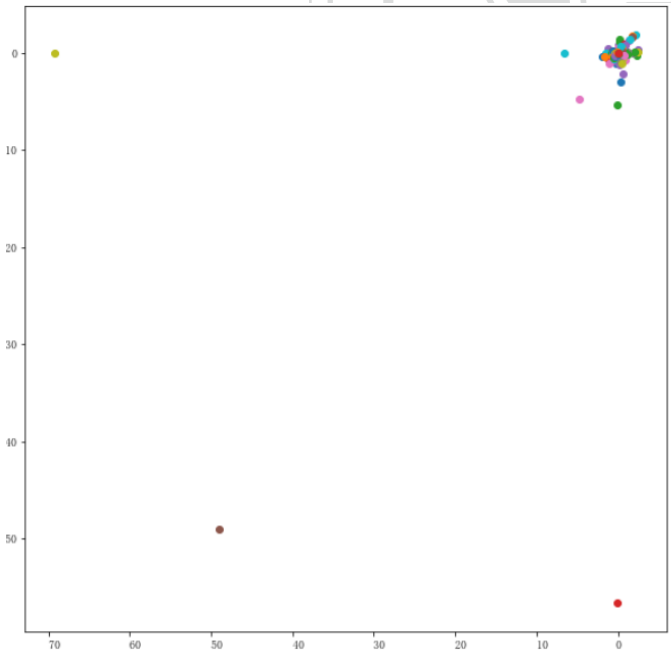
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 42.5\%$



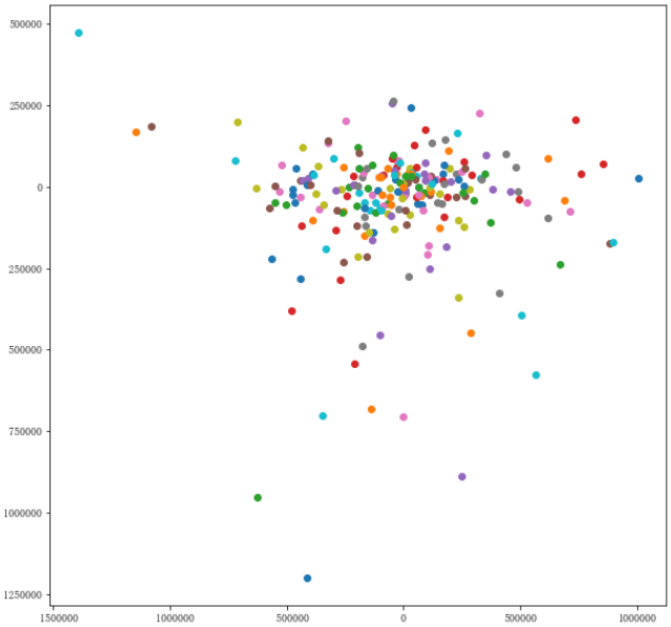
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 49.1\%$



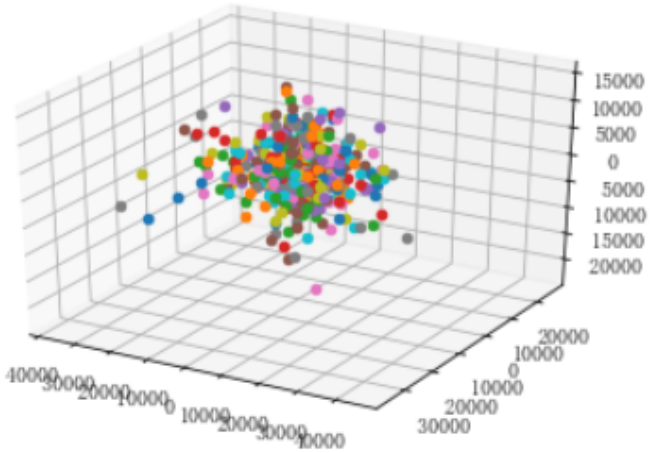
MCA2 維可視圖,$e = 0.5\%$

圖 6.1.8 1971～1990 年拿掉出書少於 5 本出版社不同分法分析圖

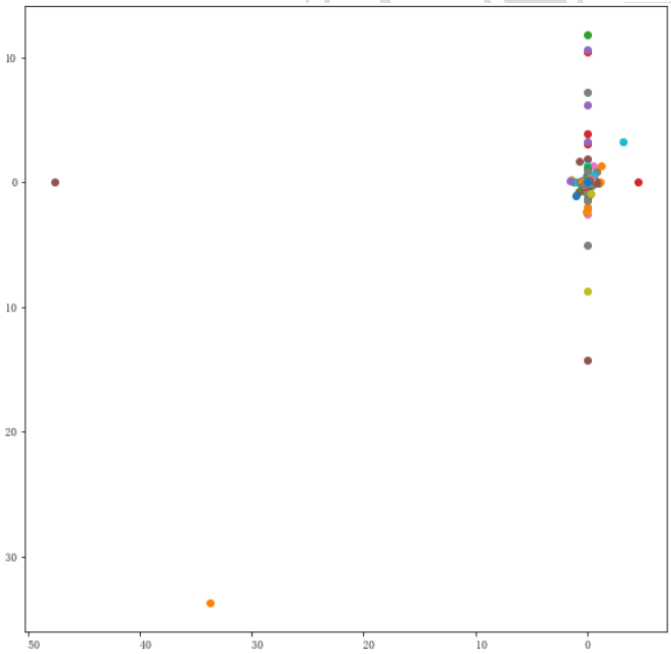
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 36.8\%$



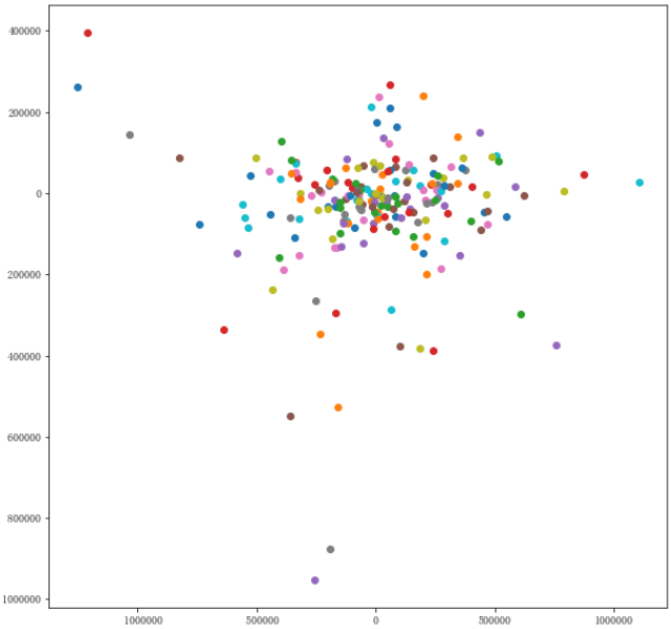
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 43.5\%$



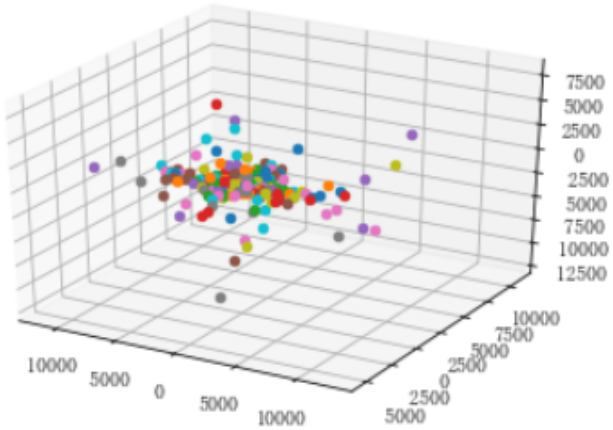
MCA2 維可視圖,$e = 0.5\%$

圖 6.1.9 1991~2010 年拿掉出書少於 5 本出版社不同分法分析圖

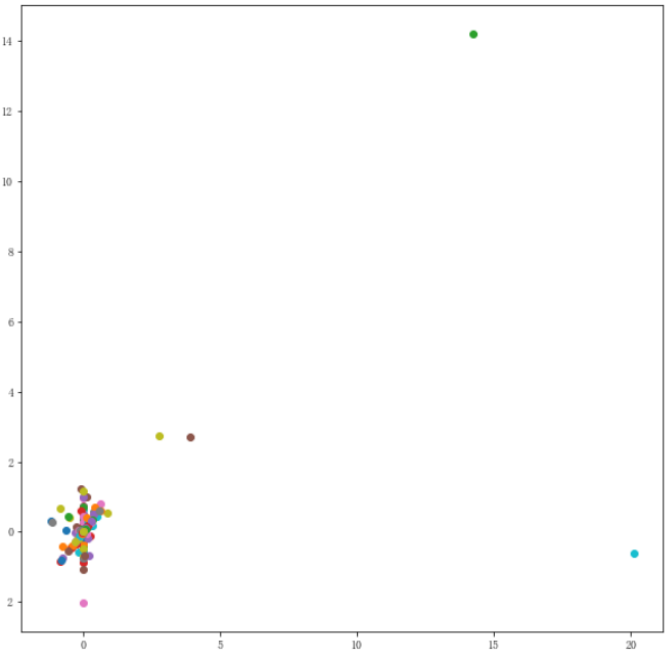
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 37.6\%$



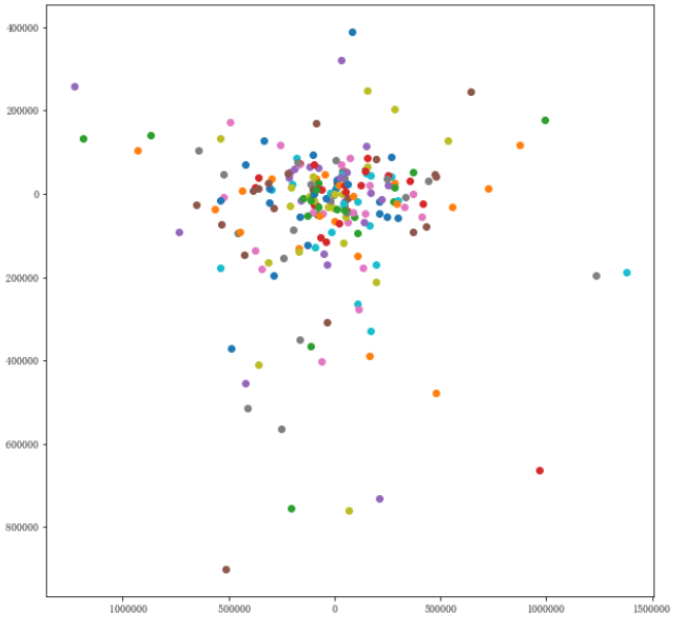
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 44.2\%$



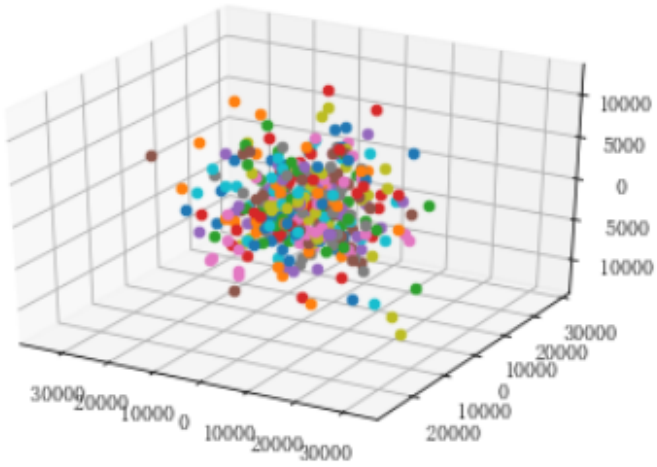
MCA2 維可視圖,$e = 1.2\%$

圖 6.1.10 1950~1970 年拿掉出書少於 6 本出版社不同分法分析圖

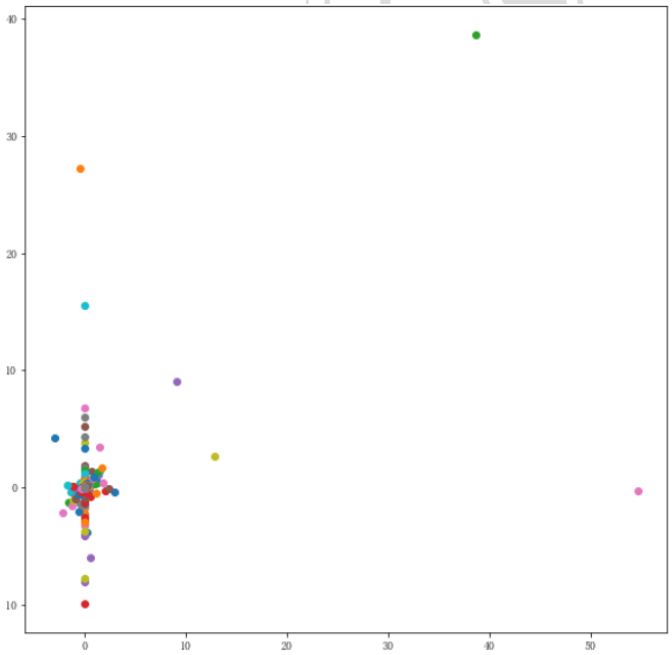
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 42.8\%$



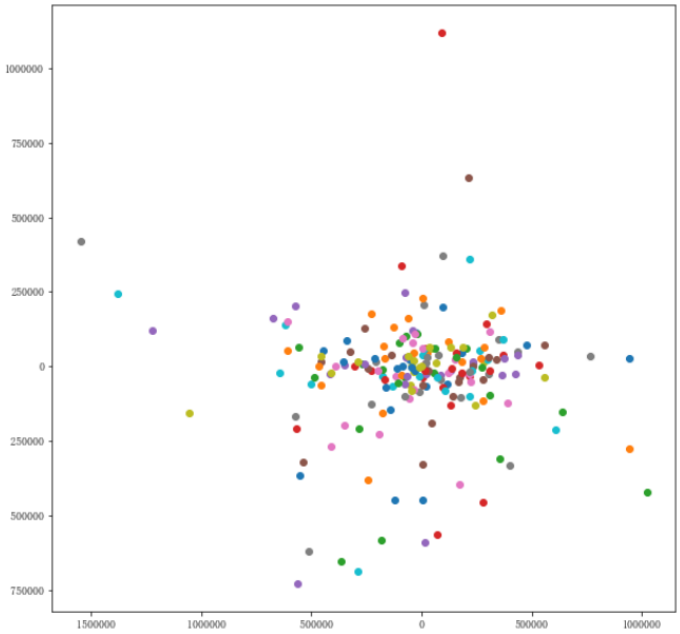
MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 49.4\%$



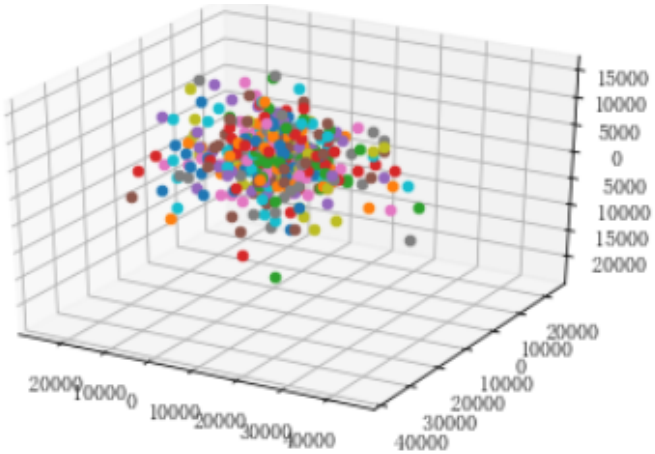
MCA2 維可視圖,$e = 0.6\%$

圖 6.1.11 1971～1990 年拿掉出書少於 6 本出版社不同分法分析圖

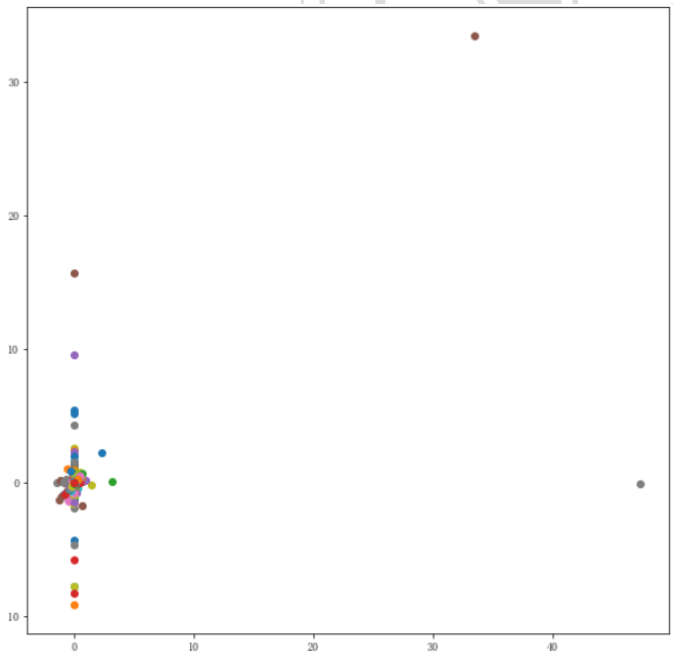
DOI:10.6814/NCCU202101092



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 37.2\%$



MDS:$w_1 = 0.9$,$w_2 = 0.8$,$e = 43.9\%$



MCA2 維可視圖,$e = 0.5\%$

圖 6.1.12 1991~2010 年拿掉出書少於 6 本出版社不同分法分析圖

DOI:10.6814/NCCU202101092

第二節　LSI 資料分析圖

圖 6.2.13 為 1950 ～ 1970 資料剃除少於三本出版社進行 LSI 分析出來的結果，左右兩張圖皆是同筆資料，左圖為右圖中心放大版本，從圖中可以看出在 1950 ～ 1970 年之間兒童文學與小說為出版社分佈重點指標，並且該圖解釋率為 56.3%。

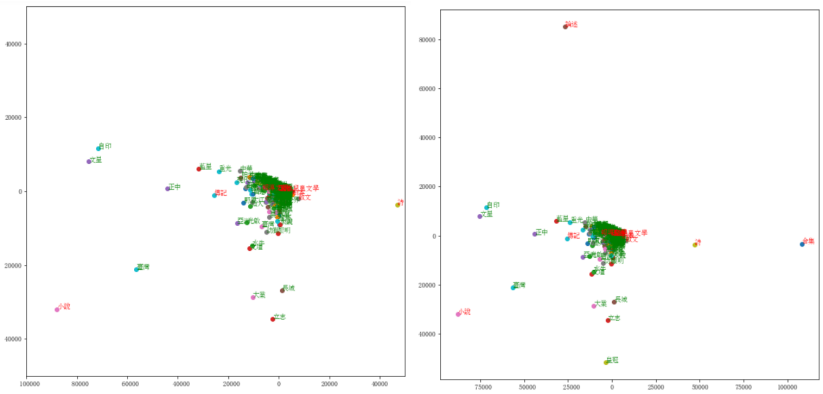


圖 6.2.13　LSI 第一期剃除資料少於三本出版社分析圖

圖 6.2.14 為 1971 ～ 1990 資料剃除少於三本出版社進行 LSI 分析出來的結果，左右兩張圖皆是同筆資料，左圖為右圖中心放大版本，從圖中可以看出在 1971 ～ 1990 年兒童文學與小說為出版社分佈重點指標，並且該圖解釋率為 53.1%。

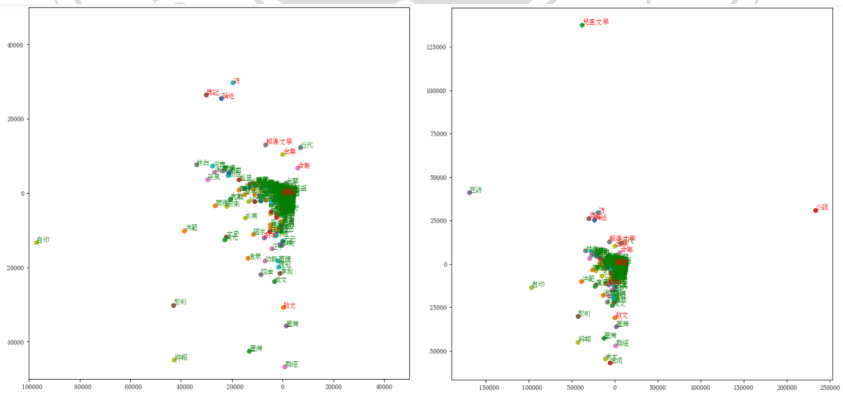


圖 6.2.14　LSI 第一期剃除資料少於三本出版社分析圖

圖 6.2.15 為 1991 ～ 2010 資料剃除少於三本出版社進行 LSI 分析出來的結果，左右兩張圖皆是同筆資料，左圖為右圖中心放大版本，從圖中可以看出在 1991 ～ 2010 年之間散文與小說為出版社分佈重點指標，並且該圖解釋率為 55.1%。

DOI:10.6814/NCCU202101092

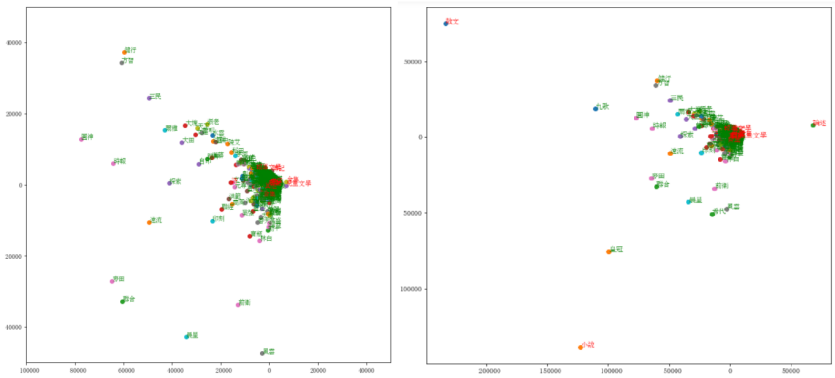


圖 6.2.15　LSI 第一期剃除資料少於三本出版社分析圖

圖 6.2.16 為 1950 ～ 2006 出版社出版文類數量，包括詩、小說、論述及散文數量折線圖，圖中可以發現小說出版數量一直穩定第一或第二，如利用 LSI 分析出來的效果相同，再者從圖中可以發現散文的發展相當迅速，從 1975 年開始出版數量就領先其他文類，在 LSI 分析當中第三期的圖表也顯示出散文對於出版社的分佈行為有較大的影響力。

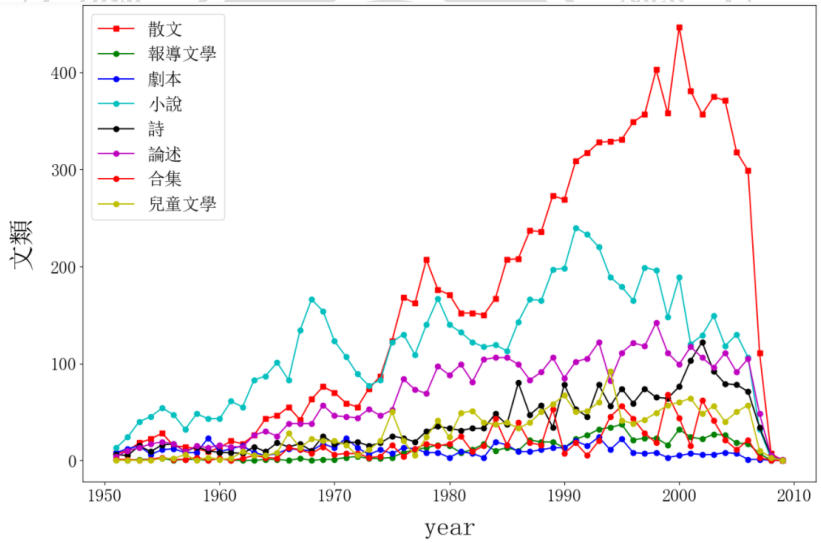


圖 6.2.16　1950 ～ 2010 出版文類數量統計圖

DOI:10.6814/NCCU202101092

第七章　結論

本研究中對於獲取資料進行多重對應分析，利用 Python 進行實驗，觀察出解釋率過低問題以及可視化結果圖分佈效果不佳，利用統計套件出來的結果圖亦是如此，後來利用多元尺度分析及潛在語義索引下去試做，明顯看出可視化結果圖的分佈情況較為清晰，其中多元尺度分析出來的結果，呈現圓形發散，肉眼無法看出分群效果，代表出版社資料沒有特殊結構化特性。而潛在語義索引可以看出小說在三個時期對於出版社的出書比例有著不容小覷的地位，意味者小說在三個時期中一直在台灣人民閱讀習慣中扮演重要腳色，再者，從第一期的論述、第二期的兒童文學以及第三集的散文，都是在該期表現最為突出的方向，也代表者台灣人閱讀或購買書籍文類的改變。

研究中使用到三種資料分析方法相同的是都利用到奇異值分解去定位變數座標，不同的是根據資料創建出來的矩陣手法。應用多元尺度分析所定義出來的距離矩陣，建立出來的矩陣在對原始資料的解釋率上也有大幅提升，二維圖大約提升到 30% ～ 40%，三維圖大約提升到 40% ～ 50%，而潛在語義索引解釋率更是高於多元尺度分析，在二維圖上解釋率就達到 40% ～ 50% 與多重對應分析相比，後兩者分析解釋率都有明顯的提升。

DOI:10.6814/NCCU202101092

Bibliography

[1] 孙云. ” 台湾政治转型后政党体制的演变及发展趋势.” (2004).

[2] 管宁. 当代台湾出版业现状与发展趋势. Diss. 2008.

[3] 寂寞的群像 ── 台灣新生代小說家的書寫與思維. 2008.

[4] 台灣作家群聚現象與網絡關係. 2017. PhD Thesis.

[5] Abdi, Hervé, and Dominique Valentin. ”Multiple correspondence

[6] Cox, Michael AA, and Trevor F. Cox. ”Multidimensional scaling.” Handbook of data visualization. Springer, Berlin, Heidelberg, 2008. 315-347.

[7] Deerwester, Scott, et al. ”Indexing by latent semantic analysis.” Journal of the American society for information science 41.6 (1990): 391-407.

[8] Guttman, Louis. ”Some necessary conditions for common-factor analysis.” Psychometrika 19.2 (1954): 149-161.

[9] Guttman, Louis. ”A basis for scaling qualitative data.” American sociological review 9.2 (1944): 139-150.

[10] Guttman, Louis. ”A general nonmetric technique for finding the smallest coordinate space for a configuration of points.” Psychometrika

[11] Hirschfelder, Joseph O., et al. Molecular theory of gases and liquids. Vol. 165. New York: Wiley, 1964.

[12] Kruskal, Joseph B. Multidimensional scaling. No. 11. Sage, 1978. analysis.” Encyclopedia of measurement and statistics 2.4 (2007): 651-657.

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202101092

[13] Kruskal, Joseph B., and Roger N. Shepard. ”A nonmetric variety of linear factor analysis.” Psychometrika 39.2 (1974): 123-157.

[14] Le Roux, Brigitte, and Henry Rouanet. Multiple correspondence analysis. Vol. 163. Sage, 2010.

[15] Sjöström, Lars, et al. ”Effects of bariatric surgery on mortality in Swedish obese subjects.” New England journal of medicine 357.8 (2007): 741-752.

[16] Tzeng, Jengnan, Henry Horng-Shing Lu, and Wen-Hsiung Li. ”Multidimensional scaling for large genomic data sets.” BMC bioinformatics 9.1 (2008): 1-17.

[17] Torgerson, Warren S. ”Multidimensional scaling: I. Theory and method.” Psychometrika 17.4 (1952): 401-419.

[18] Torgerson, Warren S. ”Theory and methods of scaling.” (1958).
33.4 (1968): 469-506.

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202101092