論述 》專論 · 評述

# 主計資料大數據分析－建立主計主題模型之研究

本研究人事費歲計估算模型法協助人事費預算編製，藉以突破歲出法律義務支出僵固性缺口特性，進而增加業務費或設備及投資等費用編列時的靈活性。主計主題輿情模型可以深入了解民眾對議題的看法。去識別化研究解決「使用他人提供的去識別化工具進行資料去識別化」使用情境，建立一系統可提供間接去識別化資料整合之方法。

**謝邦昌、蕭育仁、李紹綸、丁台怡、張葦憶**

（台北醫學大學管理學院院長兼大數據研究中心主任、台北醫學大學管理學院副院長兼 EMBA 教授、亞東技術學院管理暨健康學群資訊管理系副教授兼系主任、台北醫學大學管理學院大數據研究中心顧問、台北醫學大學管理學院大數據研究中心秘書）

## 壹、前言

大數據分析可協助了解整體趨勢與預測未來，有助於政府進行前瞻施政規劃，優化政府施政。行政院主計總處（以下簡稱主計總處）分別於 105 及 106 年度委託謝邦昌教授研究團隊進行大數據分析相關研究，並於 105 年度蒐整主計領域大數據分析發展趨勢及需求，運用大數據分析技術，就「村里常住人口推估」、「歲計估算」、「科技跨域整合資料」、「客製化 Text Mining」等四項研究議題建構資料模型，示範主計資料大數據分析之可行作法。106 年度延續與精進已建立之應用雛型，強化人事費歲計進行推估模型之研究、建立主計主題輿情模型與去識別化資料整合模型，以助於主計資料大數據分析之應用發展。

## 貳、模型研究

### 一、人事費歲計估算模型

人事費歲計估算模型主要目的是在協助主計總處人事歲計支出編列的準確性。由於人事費具有不得自其他用途別科目流入，以及有賸餘亦不得流

出的法規限制。因而容易造成業務費或設備及投資等費用編列時的排擠性和資金運用效能的侷限性。本模型可以協助預算編製與審議決策參考，藉以突破歲出法律義務支出僵固性缺口特性，藉由 GBA、薪資及人事銓敘等系統資料的蒐集發現，來探究主計總處人力運用與人事費的支用的關係，就歲出用途別而言，研究的範圍將人事類別、實際員額、考績晉級狀態、薪資、會計、預決算等資料進行歸戶並建立關係聯結，萃取出人事類別、實際員額變動、考績晉級及預決算、會計金額關係樣態，據以建立估測預算趨勢模型。

本模型主要的資料來源區分爲兩大系統，其一爲 GBA 系統內的 102 年 1 月至 105 年 12 月主計總處預決算資料，我們使用與人事費（第一級科目 01）相關的第二級用途別科目代碼區分，人事費支付類別上可以區分爲九大類，分別爲政務人員待遇、法定編制人員待遇、約聘僱人員待遇、技工及工友待遇、獎金、其他給與、加班值班費、退休離職儲金與保險。另一個資料來源爲人事處的資料庫，包括 103 年 1 月至 105 年 12 月主計總處薪資表、102 年至 105 年主計總處考績借支表、102 年至 105 年主計總處年終獎金與 102 年至 105 年主計總處不休假加班費及超過 14 天補助費清冊。本計畫即利用 102 年～ 104 年的數據資料作爲樣本內估計期（in the sample），將 105 年的數據作爲樣本外資料（out of the sample）進行估計模型驗證。透過 102 年 1 月至 105 年 12 月主計總處預決算、人事及薪資等歷史資料分析，依穩定性及不穩定性科目，藉由其他相關數據的輔助和大數據分析模組的進行，建立人事費歲計估算模型。

在 GBA 數據資料庫分析模組，針對第二級科目－固定頻率支付（以法定編製人員待遇之職員待遇爲例）分析說明，我們同時使用三項移動平均、四項移動平均與指數平滑法進行估計，並使用 102 年 1 月至 104 年 12 月當作歷史資料，105 年 1 月至 105 年 12 月做爲檢測區。針對第二級科目－分固定頻率支付（以獎金（0111）爲例）分析說明，受限資料型態的限制，我們則使用過去的平均數據進行估算，人事費歲計估算總表詳見下頁表 1。

## 二、主計主題輿情模型

主計主題輿情模型將分爲主計輿情分析及主計文件之文字探勘兩部分。

（一）主計輿情分析

主計業務職掌預決算、統計普調查（如物價、GDP、三大普查等）等重要數據發布，與民眾習習相關，若能透過科技建立主計主題輿情預測模型，快速收集民眾意見，除能配合「開放政府」政策推動外，將可協助主計人員於發布主題前後觀測民意輿情動向、社群特性及議題燃燒程度，回饋至主計主題上，使得主計業務決策更加周延平衡。經過需求收集及系統建置評估後，主計輿情分析平台模組可以自動蒐集新聞如中國時報、聯

# 論述 》專論 · 評述

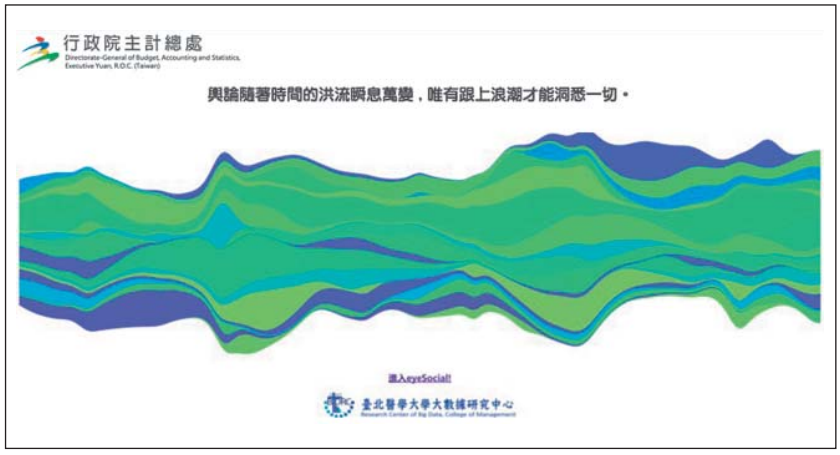
## 表 1　人事費歲計估算總表

單位：元

| 項目 | 類別 | 實際值 | 三項移動平均預測值 | 四項移動平均預測值 | 指數平滑法預測值 |
|---|---|---|---|---|---|
| 待遇 | | 530, 320, 029 | 531, 663, 310 | 531, 711, 845 | 530, 805, 303 |
| 預測誤差 | | | 1, 343, 281 | 1, 391, 816 | 485, 274 |
| 比率 | | | 0. 2533% | 0. 2624% | 0. 0915% |
| 退休離職儲金 | | 39, 573, 139 | 39, 662, 476 | 39, 664, 584 | 39, 619, 637 |
| 預測誤差 | | | 89, 338 | 91, 446 | 46, 498 |
| 比率 | | | 0. 2258% | 0. 2311% | 0. 1175% |
| 健保保險補助 | | 30, 659, 949 | 30, 944, 239 | 30, 946, 833 | 30, 803, 572 |
| 預測誤差 | | | 284, 290 | 286, 884 | 143, 623 |
| 比率 | | | 0. 9272% | 0. 9357% | 0. 4684% |
| 公保保險補助 | | 12, 093, 586 | 11, 966, 787 | 11, 967, 956 | 12, 025, 214 |
| 預測誤差 | | | -126, 799 | -125, 630 | -68, 372 |
| 比率 | | | 1. 0485% | 1. 0388% | 0. 5654% |
| 勞保保險補助 | | 9, 935, 156 | 9, 989, 958 | 9, 987, 737 | 9, 965, 222 |
| 預測誤差 | | | 54, 802 | 52, 581 | 30, 066 |
| 比率 | | | 0. 5516% | 0. 5292% | 0. 3026% |
| 獎金 | 考績獎金 | 41, 845, 194 | 40, 015, 450 | 40, 015, 450 | 40, 015, 450 |
| | 特殊功勳獎金 | 233, 800 | 221, 233 | 221, 233 | 221, 233 |
| | 年終工作獎金 | 67, 405, 926 | 66, 956, 674 | 66, 956, 674 | 66, 956, 674 |
| 其他給予 | 婚喪及生育補助 | 2, 771, 489 | 4, 211, 660 | 4, 211, 660 | 4, 211, 660 |
| | 子女教育補助 | 1, 904, 800 | 1, 906, 267 | 1, 906, 267 | 1, 906, 267 |
| | 休假補助 | 12, 016, 995 | 12, 038, 936 | 12, 038, 936 | 12, 038, 936 |
| 加班費 | 超時加班費 | 15, 086, 208 | 15, 436, 357 | 15, 436, 357 | 15, 436, 357 |
| | 不休假加班費 | 13, 168, 320 | 13, 474, 670 | 13, 474, 670 | 13, 474, 670 |
| 小計 | | 154, 432, 732 | 154, 261, 248 | 154, 261, 248 | 154, 261, 248 |
| 預測誤差 | | | -171, 484 | -171, 484 | -171, 484 |
| 比率 | | | 0. 1110% | 0. 1110% | 0. 1110% |
| 總計 | | 777, 014, 591 | 778, 488, 019 | 778, 540, 203 | 777, 480, 196 |
| 預測誤差 | | | 1, 473, 428 | 1, 525, 613 | 465, 605 |
| 比率 | | | 0. 1896% | 0. 1963% | 0. 0599% |

資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

## 圖 1　主計輿情分析平台




資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

合財經網、自由時報、蘋果日報、東森新聞網等，以及社群媒體如 FACEBOOK 粉絲團、PTT 及 Mobile01 等民眾討論內容。

從分析平台首頁（圖 1）進入後，選擇熱門文章（下頁圖 2）可以快速瀏覽討論人氣排名的主題、正面情緒排名、負面情緒排名及時間排名之文章；從熱門頻譜（下頁圖 3）可以分析最熱門的字詞；針對民眾對分析議題的情緒可以分爲正面及負面情緒感受，同時可以觀察各媒體民眾對該議題的感受程度，以及分析一段時間內議題討論正評及負評的聲量程度趨勢變化。

（二）文字探勘分析

文字探勘是以各種資料探勘方式來進行文件的文字資料分析，透過其分析來取得文字間的關聯性。與資料探勘不同之處，在於文字探勘是針對文字進行分析，且文字多屬半結構化或非結構資料，因此要先對文字進行前處理，並透過某些統計

方法與演算法，對文字進行分析與運用，進而取得必要的資訊，作爲決策的參考依據。以審查報告態樣議題分析爲例，針對 105 年度中央政府總決算審核報告進行文字態樣的文字探勘，其分析流程是先建立 excel 資料表，規劃欲分析欄位包括：A. 年度、B. 主管機關、C. 行動、D. 計畫、E. 機關、F. 原因及 G. 原因類別。接著，定義 excel 資料表萃取規則如下頁表 2。

最後，以 excel 資料表中分別進行文字探勘分析，以整份 excel 資料表進行文字探勘分析，篩選出共 152 個字詞形成詞庫。文字探勘結果包括：詞雲、長條圖、關鍵詞頻率次數表、LDA 關聯分析[1]（第 37 頁圖 4）。

## 圖 2　熱門文章示意圖


資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

## 圖 3　熱詞頻譜示意圖


資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

## 三、去識別化資料整合

本次去識別化資料整合可以將公務登記資料或調查資料導入去識別化資料整合模型後，經驗證產出去識別化後之最終整合資料庫，系統提供採用特定格式之資料匯入功能，並將去識別化後資料，匯入系統進行資料整合。

實驗內容採用 SHA－512 加密演算法，在 .NET Framework4.6 Platform，採用 Visual Studio 2017 C# 語言撰寫，結合 SQL Server 2016 進行資料加密處理與儲存，針對機敏性資料欄位（單鍵）提供去識別化，例如：A 機關 ID 欄位加密爲 ID－A，B 機關 ID 欄位加密爲 ID－B，再根據 ID－A 與 ID－B 整合聯結，爲考量 A 機關 ID－A 欄位和 B 機關 ID－B 欄位之資料筆數可能不盡相同，可能發生 1 對多、多對 1、多對多情

# 論述 》專論 · 評述

表 2　萃取規則表

| | | |
|---|---|---|
| A | 年度 | 105 年度 |
| B | 主管機關 | 遇“主管”關鍵詞，即萃取列入機關主管欄位，並建立詞庫。 |
| C | 行動 | 遇“研謀”、“改善”、“檢討”、“加強”、“精進”、“允宜”、“輔導”、“落實”、“督促”、“協處”、“拓展”、“強化”、“積極辦理”、“研商”、“議”、“強化”等關鍵詞，即萃取出列入行動欄位，並建立至詞庫。 |
| D | 計畫 | 遇“計畫”等詞關鍵詞<br>1. 若同時遇到“辦理”“執行”，則萃取“辦理”“執行”等詞之後至“計畫”前的文字，列入計畫欄位，並將相關字詞建立至詞庫。<br>2. 若未遇到“辦理”“執行”，則萃取該句，以及“計畫”之前文字。 |
| E | 機關 | 遇機關名稱，即萃取建立詞庫後，進行篩選列入機關欄位。 |
| F | 原因 | 1. 遇“惟”關鍵詞，即萃取“惟”之後的文字，列入原因欄位；<br>2. 接著遇“允”關鍵詞，即萃取“允”之後的文字，列入原因欄位；<br>3. 最後遇“未依規定”、“未盡理想”、“未能”、“不足”、“欠缺”、“尚須”、“預算執行率”、“執行率”、“使用比率”、“疑義”、“未如預期”、“欠乏”、“未落實”、“未盡周妥”、“尚欠周妥”、“未臻”、“執行缺失”、“進度落後”、“有待強化”、“調整”、“短絀”、“尚乏”、“失衡”、“有限”、“比率仍低”、“尚未研訂”、“尚無定見”、“亦待清理”、“欠周延”、“情事”、“閒置”、“不利”、“尚未完成”、“未建置完成”、“尚未完備”、“未結清”關鍵詞，即萃取該關鍵詞之後的文字，列入原因欄位。 |
| G | 原因類別 | 將原因欄位的關鍵詞篩選出，列入原因類別欄位，並建立至詞庫。 |

資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

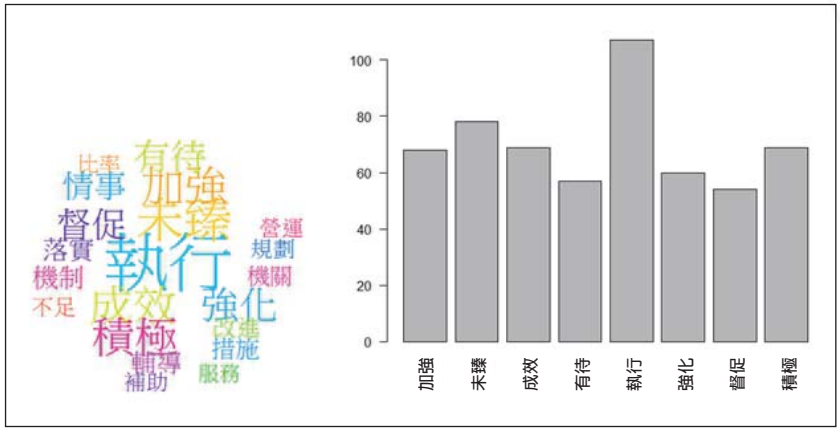
形，本研究採完全外部連結（FULL OUTER JOIN）方式進行 ID － A 與 ID － B 整合聯結。本研究亦可針對複合鍵提供去識別化，例如：A 機關 ID ＋出生年月日等欄位加密爲複合鍵 ID － A1，B 機關 ID ＋出生年月日等欄位加密爲複合鍵 ID － B1，再根據複合鍵 ID － A1 與複合鍵 ID － B1 整合聯結，爲考量 A 機關 ID － A1 欄位和 B 機關 ID － B1 欄位之資料筆數可能不盡相同，可能發生 1 對多、多對 1、多對多情形，本研究建議採完全外部連結（FULL OUTER JOIN）方式進行 ID － A1 與 ID － B1 整合聯結。去識別化研究提供採用特定格式之資料匯入功能，解決「使用他人提供的去識別化工具進行資料去識別化」使用情境，針對機敏性資料欄位（單鍵及複合鍵）提供去識別化，並透過建立一系統可提供去識別化資料整合之方法，此系統可進行資料去識別化以及資料整合，資料整合模型情境詳見下頁圖 5。

## 參、結論與發現

### 一、人事費歲計估算模型

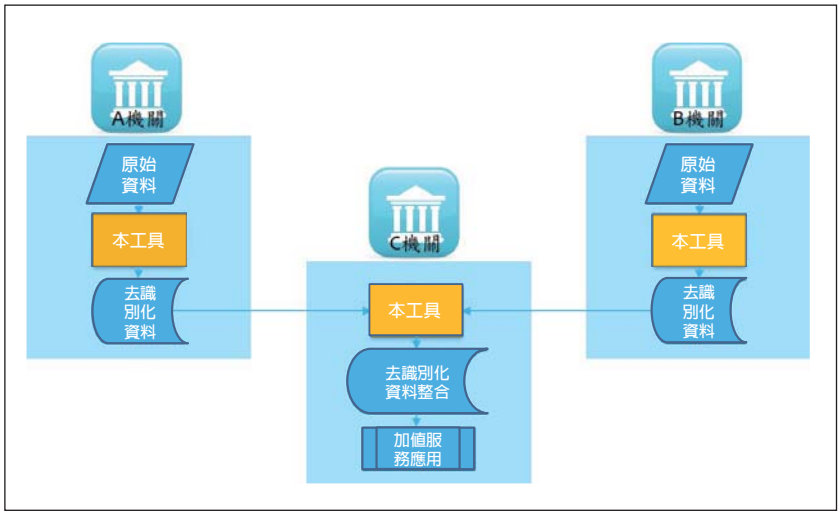
總結上述人事費歲計估算模型結果，主計總處每年將近7.5億元的人事費支出，經由精算模型下指數平滑法的人事費估算誤差只有465,605元，估算誤差達到0.0599%，透過此法協助人事費預算編製，藉以突破歲出法律義務支出僵固性缺口特性，進而增加業務費或設備及投資等費用編列時的靈活性，達成主計單位把錢花在刀口上的使命。在歲計估算模型研究結果發現，GBA系統資料在人事費部分屬彙總性資料，分析資料筆數較少，無法顯示人事實際員額變動、考績晉級狀態，對於解釋問題能力較爲薄弱，須再結合更細緻資料予以分析。

圖4　105年度中央政府總決算審核報告詞雲及詞頻分析




資料來源：主計資料大數據分析－建立主計主題模型研究計畫。

圖5　資料整合模型情境




資料來源：作者自行繪製。

### 二、建立主計主題輿情模型

本研究提供簡易快速收集新聞、社群網站等輿情，或是預決算書、統計調查報表等非結構化資料，進行深度挖掘分析，可使決策者迅速有效掌握民眾輿情或文字資料意涵。在輿情分析部分，由於輿情內容來源因版權使用的問題，本研究係以免費的社群網路資源進行探索與雛型實作，並於公有雲方式進行雛型建置；未來在導入輿情分析工具或資訊服務時，可結合有付費新聞頻道、

增加新聞探勘頻率、推播方式來擴大其應用效益。在主計文件文字探勘部分，未來可導入人工智慧分析方式如 deep learning，來豐富主計詞庫，增加文字探勘斷詞及詞意精確度，並運用詞庫來進行主計文件的分析。在已開發的客製化 Text Mining 平台雛型模組，可以提供簡易快速收集新聞、社群網站等輿情，或是預決算書、統計調查報表等非結構化資料，進行深度挖掘分析，未來主計總處可延續平台雛型模組擴充功能，創造更多資料應用價值效益。

## 三、去識別化資料整合研究

本次去識別化研究提供採用特定格式之資料匯入功能，解決「使用他人提供的去識別化工具進行資料去識別化」使用情境，並將去識別化後資料，匯入系統進行資料整合。模組採用 SHA － 512 加密演算法，在 .NET Framework4.6 Platform，採用 Visual Studio 2017 C# 語言撰寫，結合 SQL Server 2016 及其 In － Memory 特性，加速資料加密處理與儲存，針對機敏性資料欄位（單鍵及複合鍵）提供去識別化，並透過建立一系統可提供去識別化資料整合之方法，此系統可進行資料去識別化以及資料整合。同時，加入去間接識別化研究，提供給未來有資料整合需求之單位研究參考。

## 註釋

1. LDA 關聯分析（Latent Dirichlet Allocation, LDA）是非監督機器學習技術，可以用來識別大規模文檔集（documentcollection）或語庫（corpus）中潛藏的主題資訊。

## 參考文獻


1. 陳敦源、蕭乃沂、廖洲棚（2015），「邁向循證政府決策的關鍵變革：公部門巨量資料分析的理論與實務」，國土及公共治理季刊，第 3 卷，第 3 期，第 33 － 44 頁。
2. 蕭乃沂、陳敦源、廖洲棚、楊立偉、呂俊宏（2014），政府應用巨量資料精進公共服務與政策分析之可行性研究（編號：NDC － MIS － 103 － 003）。臺北市：國家發展委員會。
3. 呂建億（2015），民眾對政府輿情分析方法之信任研究－民意調查與網路輿情分析的比較。國立政治大學公共行政研究所，未出版，臺北市。
4. 劉宗熹（2016），公務機關巨量資料分析應用推動簡介。政府機關資訊通報（341），頁 1 － 9。
5. 朱斌妤、黃東益、洪永泰、曾憲立、李仲彬（2015），數位國家治理（2）：國情追蹤與方法整合（編號：NDC － MIS － 103 － 001）。臺北市：國家發展委員會。
6. 行政院國家資訊通信發展推動小組（2015），政府機關運用巨量資料分析推動說明。❖